

# A UNIFIED THEORY OF PROBABILISTIC MODELING AND EPISTEMIC CONFLICT

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Oliver E. Richardson

August 2024

© 2024 Oliver E. Richardson  
ALL RIGHTS RESERVED

# A UNIFIED THEORY OF PROBABILISTIC MODELING AND EPISTEMIC CONFLICT

Oliver E. Richardson, Ph.D.

Cornell University 2024

What should you do with conflicting information? To be *rational*, you must immediately resolve the inconsistency, so as to maintain a consistent (probabilistic) picture of the world. But how? And is it really critical to do so immediately? Inconsistency is clearly undesirable, but we stand to gain a lot by representing it.

This thesis develops a broad theory of how to approach probabilistic modeling with possibly-inconsistent information, unifying and reframing much of the literature in the process. The key ingredient is a novel kind of graphical model, called a *Probabilistic Dependency Graph* (PDG), which allows for arbitrary (even conflicting) pieces of probabilistic information. In [Part I](#), we establish PDGs as a sweeping generalization of other models of mental state, including traditional graphical models such as Bayesian Networks and Factor Graphs, as well as causal models, and even generalizations of probability distributions, such as Dempster-Shafer Belief functions. In [Part II](#), we show that PDGs also capture modern neural representations. Surprisingly, standard loss functions can be viewed as the inconsistency of a PDG that models the situation appropriately. Furthermore, many important algorithms in AI are instances of a simple approach to resolving inconsistencies. In [Part III](#), we provide algorithms for PDG inference, and uncover a deep algorithmic equivalence between the problems of inference and calculating a PDG's numerical degree of inconsistency. We also develop powerful yet inuitive principles for reasoning with (and about) PDGs.

## **BIOGRAPHICAL SKETCH**

TODO

TODO

## **ACKNOWLEDGEMENTS**

TODO

## TABLE OF CONTENTS

Biographical Sketch . . . . .	iii
Dedication . . . . .	iv
Acknowledgements . . . . .	v
Table of Contents . . . . .	vi
List of Figures . . . . .	xii
<b>1 Introduction</b>	<b>1</b>
1.1 In Defense of Inconsistency . . . . .	1
1.2 Overview of Results . . . . .	4
1.3 Themes and Motifs . . . . .	8
1.3.1 Mathematical Precision for Informal Reasoning . . . . .	8
1.3.2 Qualitative and Quantitative Information . . . . .	9
1.3.3 Equivalent Representations . . . . .	10
<b>2 Background and Mathematical Preliminaries</b>	<b>12</b>
2.1 Basic Concepts and Notation . . . . .	12
2.2 Variables . . . . .	17
2.3 Probability . . . . .	20
2.3.1 Probability in the Finite Case . . . . .	21
2.3.2 Probability, in General . . . . .	26
2.4 The Basic Theory of Graphs and Hypergraphs . . . . .	30
2.5 Probabilistic Graphical Models . . . . .	36
2.5.1 Bayesian Networks and Variants . . . . .	37
2.5.2 Markov Random Fields, Factor Graphs, and Variants . . . . .	39
2.5.3 Other Graphical Models . . . . .	40
2.6 Information Theory . . . . .	41
2.6.1 Shannon Entropy and the Information Profile . . . . .	42
2.6.2 Relative Entropy . . . . .	49
<b>I A Universal Modeling Language</b>	<b>52</b>
<b>3 Probabilistic Dependency Graphs (PDGs)</b>	<b>53</b>
3.1 Introduction and Examples . . . . .	53
3.2 Syntax . . . . .	61
3.2.1 Alternate Equivalent Definitions of PDGs . . . . .	63
3.2.2 Combining PDGs . . . . .	65
3.3 The Semantics of PDGs . . . . .	68
3.3.1 PDGs As Sets Of Distributions . . . . .	68
3.3.2 PDGs As Scoring Functions over Joint Distributions . . . . .	69
3.3.3 PDGs As Unique Distributions . . . . .	78
3.3.4 Further Semantics: PDGs as Degrees of Inconsistency and as Transformations on Distributions . . . . .	79

3.4	Relationships to Other Graphical Models . . . . .	80
3.4.1	Bayesian Networks . . . . .	80
3.4.2	Factor Graphs . . . . .	83
3.4.3	Factored Exponential Families . . . . .	86
3.5	Discussion . . . . .	90
<hr/>		
Chapter 3 Appendices		
3.A	Proofs . . . . .	92
3.A.1	Properties of the Scoring Semantics . . . . .	92
3.A.2	Bayesian Networks as PDGs . . . . .	98
3.A.3	Factor Graph Proofs . . . . .	101
<b>4</b>	<b>Representing Things with PDGs</b>	<b>105</b>
4.1	Probabilities and Random Variables . . . . .	106
4.2	Widgets . . . . .	108
4.2.1	Relations and Constraints . . . . .	108
4.2.2	“Soft” Constraints and Barriers . . . . .	111
4.2.3	Couplings and Wasserstein Metrics . . . . .	112
4.2.4	Incomplete CPDs and Individual (Conditional) Probabilities	114
4.2.5	Probability Ranges . . . . .	120
4.3	Other Representations of Knowledge and Uncertainty . . . . .	121
4.3.1	Convex Sets of Probabilities . . . . .	122
4.3.2	Belief and Plausibility Functions . . . . .	124
4.3.3	Pseudomarginals on Cluster Graphs . . . . .	129
4.3.4	Causal Models . . . . .	130
4.3.5	Implicit Neural Representations . . . . .	131
<hr/>		
Chapter 4 Appendices		
4.A	Proofs . . . . .	132
<b>5</b>	<b>Qualitative Mecahnism Independence</b>	<b>134</b>
5.1	Introduction . . . . .	135
5.2	Qualitative Independent-Mechanism (QIM) Compatibility . . . . .	138
5.3	QIM-Compatibility and Causality . . . . .	143
5.3.1	The Equivalence Between QIM-Compatibility and Arising from a Randomized PSEM . . . . .	146
5.3.2	Interventions, and the Correspondence Between Witnesses and Causal Models . . . . .	148
5.4	QIM-Compatibility and Information Theory . . . . .	152
5.4.1	A Necessary Condition for QIM-Compatibility . . . . .	154
5.4.2	A Scoring Function for QIM-Compatibility . . . . .	158
5.5	Discussion . . . . .	159
<hr/>		
Chapter 5 Appendices		
5.A	Proofs . . . . .	160
5.A.1	From CPDs to Distributions over Functions . . . . .	160

5.A.2	Results on (In)dependence . . . . .	162
5.A.3	Causality Results of Section 5.3 . . . . .	170
5.A.4	Information Theoretic Results of Section 5.4 . . . . .	177
5.B	Constructions and Counterexamples . . . . .	182
5.B.1	Parallel Arcs without Functional Dependency . . . . .	182
5.B.2	Counter-Examples to the Converse of Theorem 5.7 . . . . .	184
5.C	From Causal Models to Witnesses . . . . .	185
<b>II</b>	<b>A Universal Objective</b>	<b>187</b>
<b>6</b>	<b>Loss as the Inconsistency of a PDG: Choose your Model, not your Loss</b>	<b>188</b>
6.1	Introduction . . . . .	188
6.2	Standard Metrics as Inconsistencies . . . . .	192
6.2.1	Three Dimensions of Log-Likelihood . . . . .	192
6.2.2	Accuracy and Square Loss . . . . .	195
6.3	Regularizers and Priors as Inconsistencies . . . . .	197
6.4	Statistical Distances as Inconsistencies . . . . .	198
6.5	Variational Objectives and Bounds . . . . .	202
6.5.1	PDGs and Variational Approximations . . . . .	202
6.5.2	Variational Auto-Encoders and PDGs . . . . .	204
6.5.3	The $\beta$ -VAE Objective . . . . .	205
6.6	Free Energy as Factor Graph Inconsistency . . . . .	206
6.7	Beyond Standard Losses: A Concrete Example . . . . .	207
6.8	Reverse-Engineering a Loss Function? . . . . .	209
6.9	Conclusions . . . . .	210
<hr/>		
Chapter 6 Appendices		
6.A	Notes . . . . .	212
6.B	Further Results and Generalizations . . . . .	214
6.B.1	Full Characterization of Gaussian Predictors . . . . .	214
6.B.2	Full-Dataset ELBO and Bounds . . . . .	217
6.B.3	More Variants of Log Likelihood Results . . . . .	218
6.C	Proofs . . . . .	219
6.C.1	Additional Proofs for Inline Claims . . . . .	236
<b>7</b>	<b>The Local Inconsistency Resolution (LIR) Algorithm</b>	<b>246</b>
7.1	Introduction . . . . .	246
7.2	Parametric PDGs . . . . .	247
7.3	Local Inconsistency Resolution (LIR) . . . . .	249
7.4	LIR in the Classification Setting . . . . .	251
7.5	The EM Algorithm as LIR . . . . .	253
7.6	Generative Adversarial Training as LIR . . . . .	254
7.7	Message Passing Algorithms as LIR . . . . .	255
7.8	Discussion and Future Work . . . . .	259

Chapter 7 Appendices	
7.A Proofs . . . . .	260
<b>III Algorithms, Logic, and Complexity</b>	<b>263</b>
<b>8 Inference for PDGs, via Exponential Conic Programming</b>	<b>264</b>
8.1 Introduction . . . . .	264
8.2 Preliminaries & Related Work . . . . .	267
8.3 Inference as a Convex Program . . . . .	272
8.3.1 Minimizing Incompatibility ( $\gamma = 0$ ) . . . . .	273
8.3.2 $\gamma$ -Inference for small $\gamma > 0$ . . . . .	275
8.3.3 Calculating the $0^+$ -semantics ( $\gamma \rightarrow 0$ ) . . . . .	277
8.4 Polynomial-Time Inference Under Bounded Treewidth . . . . .	278
8.5 Experiments . . . . .	285
8.6 Discussion and Conclusion . . . . .	288
Chapter 8 Appendices	
8.A Proofs . . . . .	290
8.A.1 Properties of PDG Semantics Needed for Inference . . . . .	290
8.A.2 Correctness and Complexity Analysis for PDG Inference via Exponential Conic Programming . . . . .	299
8.B The Convex-Concave Procedure, and Implementation Details . .	337
8.C Details on the Empirical Evaluation . . . . .	340
8.C.1 Synthetic Experiment: Comparison with Black-Box Optimizers, on Joint Distributions. . . . .	340
8.C.2 Synthetic Experiment: Comparing with Black-Box Optimizers, on Tree Marginals . . . . .	345
8.C.3 Comparing to Belief Propagation, on Tree Marginals. . . . .	350
<b>9 Lower Bounds, and the Deep Connection between Inconsistency and Inference</b>	<b>352</b>
9.1 A Semantic Connection . . . . .	352
9.2 The Computational Complexity of (Approximate) Inference and Inconsistency Calculation . . . . .	354
9.3 A Deeper, Computational Connection . . . . .	359
9.4 The Reductions . . . . .	361
9.4.1 Hardness . . . . .	361
9.4.2 Inference via Inconsistency Minimization. . . . .	367
<b>10 Reasoning with PDGs</b>	<b>396</b>
10.1 Observational (Quantitative) Monotonicity and Equivalence . .	396
10.2 Structural (Qualitative) Monotonicity and Equivalence . . . . .	396
10.2.1 Qualitative Monotonicity . . . . .	396
10.2.2 QIM Equivalence . . . . .	400

10.3 A Logic Based on PDGs . . . . .	402
10.3.1 A Natural Preorder on PDGs . . . . .	402
10.3.2 “Propositional” PDG Logic . . . . .	407
10.3.3 Epistemic Logic . . . . .	409
<hr/>	
Chapter 10 Appendices	
10.A Proofs . . . . .	411
10.B Negative Results and Anti-Conjectures . . . . .	416
<b>IV Foundations</b>	<b>418</b>
<b>11 Learner’s Confidence</b>	<b>419</b>
11.1 Introduction To Learner’s Confidence . . . . .	419
11.2 A Formal Model of Confidence, Learning, and Belief . . . . .	428
11.2.1 Abstract Confidence Domains . . . . .	429
11.2.2 Belief States and Commitment Functions . . . . .	430
11.2.3 Modeling Observations: Degree of Belief, and Structural Symmetry . . . . .	435
11.3 Commitment on Confidence Continua . . . . .	437
11.3.1 The Vector Fields of Commitment Functions, and Orderless Combination . . . . .	438
11.3.2 Optimizing Learners . . . . .	443
11.3.3 Optimizing Commitment Functions for Probabilistic Beliefs	444
11.4 Further Examples, in Depth . . . . .	447
11.4.1 Update Rules for Discrete Probabilities . . . . .	447
11.4.2 Update Rules for Parametric Families . . . . .	447
11.4.3 Kalman Filters . . . . .	448
11.5 Discussion . . . . .	448
<hr/>	
Chapter 11 Appendices	
11.A Further Discussion: Full Confidence, Incremental Confidence, and Independence . . . . .	449
11.A.1 Updating with Full Confidence . . . . .	449
11.A.2 Discussion on Incremental Confidence and Independence Assumptions 1 . . . . .	451
11.A.3 Sequential Observations and Input Independence . . . . .	456
11.B More Examples . . . . .	458
11.C Extra Properties of Update Rules . . . . .	465
11.D Proofs . . . . .	467
<b>12 Relative Entropy Soup</b>	<b>470</b>
<b>13 The Category Theory of PDGs</b>	<b>471</b>
13.1 A Primer on Category Theory . . . . .	471
13.2 A Categorical Picture of PDGs . . . . .	475

13.2.1	Limits . . . . .	478
13.2.2	Natural Transformations . . . . .	480
<b>V</b>		<b>482</b>
<b>14</b>	<b>Conclusions</b>	<b>483</b>
14.1	Summary . . . . .	483
14.2	Future Work and Open Questions . . . . .	485
14.3	Impact: Implementing and Applying the Theory . . . . .	489

## LIST OF FIGURES

2.1	Examples of directed hypergraphs and their duals . . . . .	35
2.2	Illustration of the information profile $I_\mu$ . . . . .	44
3.1	A BN without arcs, its corresponding PDG, and how the latter can be augmented (losslessly) with additional cpds. . . . .	54
3.2	The classic “smoking” BN, its corresponding PDG, and a restriction of the latter, which can be augmented with other information about cancer. . . . .	57
3.3	Grok’s prior (left) and combined (right) knowledge. . . . .	60
3.4	Illustrations of the information deficiency ( $SDef_{\mathcal{A}}$ ) for various hypergraphs $\mathcal{A}$ . . . . .	75
3.5	Illustration of how a factor graph is converted to a PDG . . . . .	85
3.6	Converting a PDG to a <i>strict</i> PDG . . . . .	85
4.1	A widget PDG for capturing a single conditional probability: a statement of the form $\Pr(Y=y \mid X=x) = p$ , for $p \in [0, 1]$ . . . . .	115
4.2	A widget implementing a CPD whose presence is conditioned on a guard variable, i.e., $p(B A, G=1)$ . . . . .	119
4.3	A widget PDG for capturing a conditional probability range, a statement of the form $\Pr(Y=y \mid X=x) \in [a, b]$ . . . . .	120
5.1	An illustration of the interaction information identity . . . . .	153
6.1	Variants of log-probability based losses, across three orthogonal dimensions: conditional vs unconditional, multi-sample vs single-sample, and latent-variable vs full-information . . . . .	195
6.1	Statistical distances as inconsistencies: a map of the inconsistency of $p(X)$ and $q(X)$ , as their respective confidences vary. . . . .	198
6.2	A visual proof of the data-processing inequality for all PDG divergences, with monotonicity . . . . .	202
7.1	Two Illustrations for adversarial training: with a PPDG and a PDG	253
8.1	Empirical results: accuracy and resource costs for the inference algorithm and baselines . . . . .	285
8.2	Comparison of convex solver and black-box optimization baselines. Memory footprints, and accuracy/time costs for the cluster setting. . . . .	286
8.C.1	More details: resource costs for joint-distribution optimization setting . . . . .	341
8.C.2	differences in performance between the Gibbs and simplex parameterizations of probabilities. . . . .	344
8.C.3	A disaggregated version of Figure 8.1 . . . . .	345

8.C.4 Plot of the accuracy gap for inference methods and baselines, for various values of $\gamma$ .	346
8.C.5 Objective gap vs time in the cluster setting; shows more separation	347
8.C.6 Resource costs for the cluster setting.	347
8.C.7 Gap vs inference time for the small PDGs in the <code>bnlearn</code> repository	349
8.C.8 A variant of Figure 8.C.1, with with gap (accuracy) information on the left, and slightly different parameter settings.	351
11.1 Different representations of update functions, and the relationships they have with one another.	425

# CHAPTER 1

## INTRODUCTION

### 1.1 In Defense of Inconsistency

*“A man with one watch always knows the time;  
a man with two watches is never sure.”  
(San Diego Union 1930)*

A central principle of scientific inquiry is this: if you want to eventually understand things clearly, then you must take seriously the possibility that your understanding today might be wrong. And not only might you be wrong in the margins—wrong about the details of a calculation or under-informed—you could have a fundamentally misguided understanding of how the world works. You could be using meaningless concepts, confusing cause and effect, and misplacing your trust in those that seek to manipulate you. For those of us who prize logic and rationality (including, famously, [Descartes \(1637\)](#)), perhaps most terrifying of all is the possibility that your own internal thoughts might not even be self-consistent. If there is one thing on which we can all agree, it is a contempt for inconsistency. The foundational principles of microeconomics entitle us to our own tastes (and beliefs, to some extent), but only if they admit a consistent internal logic. For without one, we risk being swindled ([Vineberg 2022](#)) or misled ([Priest et al. 1996](#)), if not humiliated ([Finocchiaro 1981](#); [Russell 1902](#)).

Computer scientists like us, who build cognition from scratch, have capitalized on an opportunity to eliminate contradictions in artificial agents altogether, by using epistemic representations that are consistent by design. Theorists are careful to prove things about clearly defined models, while applied computer

scientists have long built artificial agents using representations that, by construction, cannot even encode inconsistency. We would rather have one definition of  $X$  and one technique for computing  $X$  than several different approaches and implementations that could, in principle, disagree. We would rather use one library for telling time than two. And no matter how many parameters we add, the goal is to parameterize a single coherent probability distribution. Why? Because computing things in just one way grants an invulnerability to our one common enemy: inconsistency.

Yet this invulnerability comes with shackles. Would *you* trade *your* freedom to see things from two angles at once—to puzzle over paradoxes, to experience deep surprise and be challenged by internal conflict—for a guarantee of self-consistency? In order to always present a consistent front, it is necessary to fully process new information before moving on. Mulling is prohibited, as is waiting for clarification. Half-baked ideas are a nonstarter. This regimented consistency-first lifestyle works well if you are only mistaken in the margins, and need only to occasionally update a few parameters. But what if you’re wrong in a deeper way? Can you imagine how hard it would be to learn to see things completely differently, without ever venturing into a confused state? Picture having to substantially refactor a large codebase with an interface that only allows you to see code that compiles. So if (through no fault of your own) you do happen to be wrong about how the world works—that is, really, fundamentally wrong—then taking on inconsistency may be worthwhile.

Engineers and computer scientists generally agree that *redundancy*—by which we mean multiple independent approaches, tests, and safeguards that all point in the same direction—is highly desirable. But, in a sense, redundancy and inconsistency are opposite sides of the same coin. If we use only representations that

avoid inconsistency by preempting the possibility of two meaningfully distinct answers to a question (using only one watch, so to speak), then we also cannot have meaningful agreement, and hence no redundancy. Without opportunity for internal conflict, there can be no opportunity for internal agreement; it is the possibility of disagreement that makes agreement valuable.

Overlapping and redundant representations are incredibly useful in practice for guarding against catastrophic failure. For this reason, even modern AI systems have recently been slipping away from the probabilistic consistency to which they aspire. For instance, it is expensive and even undesirable to exactly model your training data; instead we try to do the opposite, by adding priors (also known as regularizers; see [Section 6.3](#)) to prevent this. The priors are typically inconsistent with the data (which is arguably the point), although neither is ever modified. Still, the result is a far more robust learning process, both theoretically ([McMahan 2014](#); [Livini 2017](#)) and practically ([Girosi et al. 1995](#)). Even clearer examples of inconsistency arise in variational inference (e.g., mean field approximation methods, variational autoencoders ([Kingma and Welling 2014](#))), and more generally when combining multiple networks in nontrivial ways. In some cases, there is even a term representing “consistency” in the training objective ([Zhu et al. 2017](#)), or the name of the approach ([Zhou et al. 2003](#); [Dwibedi et al. 2019](#)). In each case, the goal is to be (mostly) consistent, but seldom is it possible to actually get there, and tolerating inconsistency is critical to the learning process. These techniques are billed as pragmatic ways of approximating some presumably consistent we must have intended. But why perpetuate these unrealistic rationality standards? I argue that we should just call them what they are: (possibly) inconsistent belief states.

Doing so intelligently unlocks a beautiful theory of probabilistic modeling,

that unifies a broad range of concepts, representations, and algorithms across artificial intelligence, and the study of rational belief and decision making. The framework developed in this dissertation gives us the ability to model, measure, and mitigate inconsistency. This affords an agent an enormous amount of epistemic flexibility, making it far easier to make deep structural changes to its epistemic representations. So, in an ironic twist, the possibility of internal inconsistency might ultimately be what saves you from being deeply wrong. Make no mistake: inconsistency is still bad. Indeed, our measure of it is the only conception of “bad” needed to do much of machine learning ([Chapters 6](#) and [7](#)). There is no question that chronically exhibiting a high degree of inconsistency is problematic. Yet, as with all internal conflict, it is more productive to engage with it and recognize it as a catalyst for positive change than it is to avoid it entirely.

Is really worse to be gullible than it is to be stuck and very wrong? It is said that a man with two watches never knows the time. But is a man with only one watch really better off? And how are you supposed to use multiple watches, anyway? Let me show you.

## 1.2 Overview of Results

This thesis endeavors to provide a broad unified picture of many representations, concepts, and algorithms used in modern AI systems. At a high level, the idea is simple: represent everything in probabilistic terms, and then identify and resolve inconsistencies between them. The details, of course, are more involved. [Chapter 2](#) reviews the relevant background material, focusing particularly on those elements of the theory of probabilistic graphical models and information theory;

it also develops a (novel) account of *variables* that unifies standard notation across different communities. Both aspects serve as the building blocks for the theory that follows. The rest of the material is based on three previously published conference papers ([Richardson and Halpern 2021](#); [Richardson 2022](#); [Richardson, Halpern, and De Sa 2023](#)), two workshop papers ([Richardson 2023](#); [Richardson and Bao 2024](#)), additional papers in earlier stages of the peer review process ([Richardson, Peters, and Halpern 2024](#)), and a trove of original unpublished notes.

**A Universal Model.** [Part I](#) is about subjective representations of knowledge and uncertainty. [Chapter 3](#) introduces the key mathematical object at the heart of this dissertation, called a *probabilistic dependency graph* (PDG). It is based on our AAAI paper of the same name ([Richardson and Halpern 2021](#)), but augmented with significant elements from the other papers. In that chapter, we motivate by example the need for a representation with the kind of flexibility and modularity offered by PDGs, develop their formal syntax and semantics, and relate them to traditional graphical models—which are special cases. [Chapter 4](#) takes the PDG representation further, showing how PDGs not only capture graphical models, but also other notions of uncertainty, such as credal sets ([Walley 1991](#)) and Dempster-Shafer belief functions ([Shafer 1976](#)). In the process, we will develop a number of smaller tools, which we call *widgets*, which capture small fragments of epistemic states as PDGs—enabling us to seamlessly convert things like constraints on the values of variables, or on their probability ranges, to PDGs. [Chapter 5](#), also based on an eponymous paper ([Richardson, Peters, and Halpern 2024](#)), develops a concept that we call *qualitative mechanism independence*. The concept is a significant generalization of independencies in other graphical models that can also describe functional dependencies and give meaning to

cyclic structures. As we shall see, qualitative mechanism independence is closely related to (and can be captured by) the qualitative information in a PDG. To summarize, Part I establishes PDGs as a sweeping generalization of classical knowledge and uncertainty representations, especially those that have formed the backbone of AI systems in the last 25 years.

**A Universal Objective.** In Part II, we will see that PDGs also turn out to capture more modern learned representations, and even learning algorithms and objectives behind them. Up until this point, we will have seen a few benefits of being able to tolerate inconsistency, but here we will begin to see the benefits of measuring inconsistency precisely the way we have. In Chapter 6, based on *Loss as the Inconsistency of a Probabilistic Dependency Graph: Choose Your Model, not Your Loss Function* (Richardson 2022), we will see that PDGs capture not only the part of a modern machine learning (ML) system typically thought of as the representation (i.e., the networks and their architectures), but also the loss function used to train that system. Previously the two aspects of an ML system were considered separate design choices, and the loss was a matter of pragmatics, not truth. But our findings suggest that there may be a “universal” way of getting at the “correct” loss function simply by laying out your (possibly inconsistent) beliefs. Indeed, a wide variety of standard losses can be viewed as measuring the inconsistency of a PDG that models the situation appropriately. From these results, it follows that much of machine learning can be viewed as resolving inconsistencies. In Chapter 7, we operationalize this process, by giving a generic recipe for *how* one might resolve inconsistencies: focus your attention to some (small) view of the picture, and then make (small) changes to reduce the inconsistency of that picture. This idea turns out to be a generalization of belief propagation that applies to arbitrary PDGs, and that captures techniques such as

adversarial training and variational inference in the process.

**Algorithms, Logic, and Complexity.** [Part III](#) develops algorithms and reasoning techniques for PDGs. [Part I](#) positions PDG semantics as a potentially quite useful generalization of traditional graphical models, but to actualize any of that potential, we need to be able to do inference on them; meanwhile, [Part II](#) establishes calculating (and minimizing) a PDG’s degree of inconsistency as an important problem of interest. In [Part III](#), we solve both problems ([Chapter 8](#)), characterize their computational complexity, and investigate the deep connection between them ([Chapter 9](#)). Both [Chapters 8](#) and [9](#) are based on the full version of the PDG inference paper ([Richardson, Halpern, and De Sa 2023](#)). In [Chapter 10](#), we then flesh out a reasoning principle that appears in both [Parts I](#) and [II](#) ([Chapters 5](#) and [6](#)), called *monotonicity of inconsistency*: believing more things cannot make you any less inconsistent. This turns out to not only be the basis of an intuitive visual calculus for deriving important inequalities in the literature, but also forms the basis of a *logic* that, unlike most logics, does not become trivial starting from inconsistent premises.

**Foundations.** [Part IV](#) develops foundational concepts that underlie the theory of PDGs presented in the previous chapters. In [Chapter 11](#), we develop a generic framework for describing the notion of (*learner’s*) *confidence* that underlies the definition of a PDG—a concept that is fundamentally different from probability, but complements it. In [Chapter 12](#), we describe several ways in which PDGs fit into the general framework developed in [Chapter 11](#). PDGs also have a rich foundation in category theory. In fact, the PDG representation was inspired by categorical thinking. We describe this in [Chapter 13](#).

Finally, we conclude with discussion and open questions in [Chapter 14](#).

## 1.3 Themes and Motifs

Before embarking in earnest, we first prime ourselves to notice some important recurring themes. Some of these patterns have been important guiding principles; others surprised us by their recurrence. By recognizing these patterns, it may be easier to digest the material that succeeds them.

### 1.3.1 Mathematical Precision for Informal Reasoning

The hallmark of a good formalism is that it supports a user in manipulating the concepts quickly and intuitively, yet at the same time is precise enough to withstand critical scrutiny. Correspondingly, our representations (and tools for manipulating them) strive for two opposing virtues:

1. things that are different look different, and
2. things that are the same look the same.

The first aids precision, and the latter aids ease of use. It is common for computer scientists and mathematicians, in pursuit of precision, to focus on distinguishing objects that are even superficially different. Failure to do so called “abuse of notation”. But in some deep way, the notation is not really being abused if the two objects are really two different views of the same thing.

The idea can be described more precisely as etiquette for *implicit conversion*: when one defines a way of regarding one type of mathematical object as another that is so transparent that it can be implied. To avoid confusion, it is not strictly necessary to avoid implicit conversions—but it may require ensuring that all ways of implicitly converting an object of type  $A$  to one of type  $B$  are equivalent.

Doing so can be challenging, but it is often the best way to get representations that simultaneously attain both virtues 1 and 2. Avoiding the challenge altogether—that is, ensuring objects of different types are given different symbols—is a way of ensuring consistency by construction. This practice is precise and grants a certain peace of mind, but also makes it difficult to think. In keeping with the argument made in [Section 1.1](#), I have not shied away from defining the same symbol in more than one way—but I have tried to do **so** in a way that the definitions never disagree in any context.

\*consistently\*

PDGs, the representation at the heart of this thesis, enable this kind of thinking in the probabilistic setting. They allow us to manipulate probabilistic primitives in ways that look just like standard intuitive short-hand, yet the machinery underneath makes the reasoning completely precise. Some of the key reasoning principles have congealed to form [Chapter 10](#), but a keen reader will see examples of how PDGs formalize intuitive probabilistic reasoning throughout. But when we introduce probabilities, consistency is not just about ensuring paths are equal (as is the case for implicit conversions); it’s about something else. Making that something else precise is one objective of this thesis.

### 1.3.2 Qualitative and Quantitative Information

Broadly speaking, most modeling happens at two different levels: at an abstract conceptual level, and a concrete fully detailed one. The division between the two kinds of objects goes by different names in different communities. A few examples of the distinction we’re getting at are given below.

Qualitative	Quantitative
Structural	Observational
Types	Values
Theorems	Proofs
Definitions	Examples
Schemas	Instances

Some readers might immediately recognize this division; others may find the analogy between these concepts resonates less. By the end of the thesis, hopefully we will gain some appreciation for this concept and its role in probabilistic modeling. Following the nomenclature of the graphical models community, we will refer to it principally as the divide between *qualitative* and *quantitative* information, or alternatively as the divide between *structural* and *observational* information. Both kinds of information are important, and they are fundamentally different. The interplay between them will be a recurring theme.

### 1.3.3 Equivalent Representations

When working with abstract representations, often it is the case that one class of models (representation  $R_1$ ) seems clearly more general than another (representation  $R_2$ ). To prove this, you show that anything with representation  $R_2$  can be converted to  $R_1$ ; perhaps  $R_2$  is even literally a subset of  $R_1$ . You convince yourself that the relationship is strict, because very little of  $R_1$  is in the image of this transformation—and besides, how could you possibly represent a generic element of  $R_1$  with  $R_2$ ? Yet it is still possible that, sometime later, you discover that there is in fact a much more subtle way of capturing  $R_1$  with  $R_2$ . So, despite appearances,  $R_1$  may turn out to be equivalent to  $R_2$ .

We will see this happen almost immediately even in the preliminary material: with hypergraphs and bipartite graphs (Section 2.4), with variables, sets of variables, and random variables, (Sections 2.2 and 2.3.1), and with causal models and randomized causal models (Section 5.3), before we even get into our own contributions. But even against this backdrop, Probabilistic Dependency Graphs (PDGs), the central mathematical object whose theory that we develop in this dissertation, are particularly special in this regard. Not only can PDGs capture a great many things, but they also encompass many natural “obvious generalizations” of themselves (Section 3.2.1). If this abstract statement seems puzzling, recall that Turing machines (TMs) not only capture other notions of computation, but also capture many natural variants of TMs that seem on the surface to make them more expressive. Yet “generalized” TMs with multiple tapes, multiple heads, tapes of higher dimension, and even non-determinism—are all (semantically) equivalent to ordinary Turing Machines.

PDGs push this idea even further. Not only is it the case that there are many different ways of specifying a PDG that are all equivalent, but also there are also several different semantics (with seemingly different expressiveness), that also turn out to all be equivalent. This property, together with their ability to capture seemingly every other epistemic representation has led us to call PDGs, informally, *a universal modeling language*. Part I aims to justify this rather bold terminology. Part II explores a different aspect in which PDGs are universal: minimizing inconsistency appears to be a *universal objective*. The two aspects of PDGs are deeply related, as we will see in Part III.

But first, we should review some math.

## CHAPTER 2

### BACKGROUND AND MATHEMATICAL PRELIMINARIES

Although the high-level ideas in this thesis are often intuitive and philosophical, the most precise forms of those ideas are quite mathematical. In this chapter, we build up a library of standard mathematical concepts and notation needed to fully understand them. A reader with a firm grasp of probability, graph theory, probabilistic graphical models, convex geometry, and information theory, should already be equipped to understand the later chapters; these readers should feel free to skip to [Part I](#). Still, even for such readers, the way these concepts fit together formally may be less familiar. The notation is carefully selected to be precise, yet serve multiple purposes, and look indistinguishable from various standard notations in different contexts.

The chapter also serves a second purpose: to introduce the kinds of reasoning and representation patterns that appear in later chapters. The principle such pattern is the concept of *variable*, which has several closely related meanings. One typically takes care to distinguish these different meanings, but we do some extra leg work in this chapter to justify using them interchangeably. In this chapter, we will get a chance to work with these ideas in a familiar context.

#### 2.1 Basic Concepts and Notation

Let's start with the basics. There are two kinds of equalities: definitions, and assertions. We distinguish between the two. Namely, we write " $A := B$ " to (re)define the symbol  $A$  so as to stand for  $B$ , but write " $A = B$ " to assert that the (already defined) expressions  $A$  and  $B$  are equal.

**Sets, Maps, Numbers, and Logic.** A *set* is a collection of unique elements, such as  $\{1, \text{hello}, 7\}$ , the empty set  $\emptyset$ , the natural numbers  $\mathbb{N} = \{0, 1, 2, \dots\}$ , or the set  $\mathbb{R}$  of real numbers. A *singleton* is a set with precisely one element. We write  $a \in A$ , or  $a : A$ , to indicate that  $a$  is an element of the set  $A$  (and  $a \notin A$  to indicate otherwise). For brevity, we write  $a, b \in A$  to indicate that both  $a \in A$  and  $b \in A$ .  $A$  is a *subset* of  $B$  (written  $A \subseteq B$ ) means that every element of  $A$  is also an element of  $B$ , and  $2^A$  is the set of all subsets of  $A$ .

There are a number of important ways to combine two sets  $A$  and  $B$  to produce a third set. Their *union*  $A \cup B$  is the set of all elements contained in either  $A$  or  $B$  (or both), while their *intersection*  $A \cap B$  is the set of elements they have in common. The *product* of  $A$  and  $B$  is the set  $A \times B := \{(a, b) : a \in A \wedge b \in B\}$ . The *disjoint union* of  $A$  and  $B$  is a variant of a union in which all elements are forced to be distinct, by “tagging” them with the set they come from, i.e.,  $A \sqcup B := \{(A, a) : a \in A\} \cup \{(B, b) : b \in B\}$ . The elements must be distinct even if the two sets share a symbol; in this case, we instead tag elements based on whether they come from the left or the right:  $A \sqcup A = \{\text{inl}(a) : a \in A\} \cup \{\text{inr}(a) : a \in A\}$ .

We write  $|A|$  or  $\#A$  for the *cardinality* of a set  $A$ ; when we write  $\#A$ , that means  $A$  is *finite*, meaning it contains  $\#A = |A| \in \mathbb{N}$  distinct elements. When  $n \in \mathbb{N}$  is a natural number, let  $[n] := \{0, 1, \dots, n - 1\}$  denote the set of the first  $n$  natural numbers. When  $a$  and  $b$  are numbers (e.g., elements of  $\mathbb{N}$  or  $\mathbb{R}$ ), we denote their product by  $a \cdot b$ , or, more commonly, simply  $ab$ . We will also often consider extended numbers that include positive infinity ( $\infty$ ), and perform arithmetic with it by defining, for  $r \in \mathbb{R}$ ,

$$\infty + r := \infty, \quad \infty \cdot r := \infty \text{ for } r > 0, \quad \infty \cdot 0 := 0,$$

and leaving  $\infty \cdot r$  undefined when  $r < 0$ . We write  $\bar{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$  for the set of

extended real numbers, and  $[0, \infty]$  for just the non-negative ones.

We write  $\forall x \in X. \varphi(x)$  to indicate that a logical expression  $\varphi(x)$  is true *for all*  $x \in X$ , and  $\exists x \in X. \varphi(x)$  to indicate that there *exists some*  $x \in X$  such that  $\varphi(x)$  is true. We use both of the two standard “set-builder” notations  $\{x \in X \mid \varphi(x)\}$  and  $\{x \in X : \varphi(x)\}$  to construct the subset of  $X$  that satisfies the property  $\varphi$ ; the choice between the colon and the bar is made to enhance readability in context. We write  $\neg\varphi$  for the negation of an expression  $\varphi$ . If  $\varphi$  and  $\psi$  are both logical expressions, we write  $\varphi \wedge \psi$  for their conjunction (which is true only if  $\varphi$  and  $\psi$  are both true), and  $\varphi \vee \psi$  for their disjunction (which is true if and only if either  $\varphi$  or  $\psi$  are true). In some contexts, it is standard (and may feel more natural) to use a comma instead of “ $\wedge$ ” to conjoin logical expressions. The notation  $\varphi \implies \psi$  means that  $\psi$  holds whenever  $\varphi$  does, and can be viewed as an abbreviation for  $\neg\varphi \vee \psi$ ; the notation  $\varphi \iff \psi$  means that  $\varphi$  and  $\psi$  are logically equivalent (which can be viewed as an abbreviation for  $(\varphi \implies \psi) \wedge (\psi \implies \varphi)$ ). Speaking of logical equivalence, we often abbreviate “if and only if” by writing just “iff”.

A *map*, or *function*  $f : A \rightarrow B$  is an object that, given  $a \in A$ , produces a value  $f(a) \in B$ . If  $f, g : A \rightarrow B$  have the same behavior (i.e.,  $\forall a \in A. f(a) = g(a)$ ), then they are considered equal, and we write  $f = g$ . Thus, a function can be specified with the notation  $f = a \mapsto f(a)$ , which allows us to talk about functions without naming them. The *preimage* of a set  $V \subseteq B$  is the set  $f^{-1}(V) := \{a \in A : f(a) \in V\} \subseteq A$  of inputs to  $f$ , whose corresponding outputs are in  $V$ . We can *compose* the functions  $f : A \rightarrow B$  and  $g : B \rightarrow C$  to form a new function  $g \circ f : A \rightarrow C$ , defined by  $(g \circ f) := a \mapsto g(f(a))$ . When  $A \subseteq X$ , it is often very useful to refer to the *indicator function* for membership in  $A$ ,

$$\mathbb{1}[A] := x \mapsto \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} : X \rightarrow \{0, 1\}$$

As one might hope from the notation, “ $A \rightarrow B$ ” is the set of functions from  $A$  to  $B$ ; it is also written  $B^A$ . It is worth pausing here to reflect on some implications of this notation. First, if  $1 = \{\star\}$  is a singleton set, then an element  $a \in A$  is really no different from a map  $a \in A^1 = 1 \rightarrow A$ . Second, if  $2 = \{0, 1\}$  represents a set with 2 elements, then the elements of  $2^A$  are functions assigning either 0 (absence) or 1 (presence) to each element of  $A$ —that is, they are subsets of  $A$ . Finally,  $|A^B| = |A|^{|B|}$ .

If  $I$  is a set, then an *indexed set*  $\mathbf{a} = \{a_i\}_{i \in I}$  consists of a collection of objects, one for each  $i \in I$ . If  $\forall i. a_i \in A$ , then  $\mathbf{a}$  is just a function  $\mathbf{a} : I \rightarrow A$ . A *tuple* is the special case of a set indexed by  $[n]$  for some natural number  $n \in \mathbb{N}$ , and typically written with parentheses in order, as in  $(a_1, \dots, a_n)$ . A *sequence* is the special case of a set indexed by  $\mathbb{N}$ , and is typically also written with parentheses, as in  $(a_1, a_2, \dots)$ . For tuples and sequences, we sometimes use *slice notation*, writing, for example,  $a_{3:7}$  for the sub-sequence  $(a_3, a_4, \dots, a_7)$ .

**Linear Algebra.** For us, a *vector* is a map from a finite set  $I$ , called its *shape*, to the (possibly extended) reals. To simplify and match standard notation, we write  $\mathbb{R}^n$  instead of  $\mathbb{R}^{[n]}$ , and often write  $\mathbf{u} = [u_i]_{i \in I}$  to specify a vector  $\mathbf{u}$  by its components. Given a number  $\alpha \in \mathbb{R}$  and a vector  $\mathbf{u} \in \mathbb{R}^I$ , we can form the scaled vector  $\alpha\mathbf{u} := [\alpha u_i]_{i \in I} = i \mapsto \alpha \cdot \mathbf{u}(i)$ ; for this reason,  $\alpha$  is called a *scalar* in this context. Vectors of the same shape can be added (+) or multiplied ( $\odot$ ) pointwise as usual, and we say that  $\mathbf{u} \leq \mathbf{v}$  iff  $\mathbf{u}(i) \leq \mathbf{v}(i)$  for all  $i \in I$ . We write  $\mathbf{u} \propto \mathbf{v}$  iff there exists some a scalar  $\alpha \in \mathbb{R}$  such that  $\mathbf{u} = \alpha\mathbf{v}$ .  $\mathbf{1}$  and  $\mathbf{0}$  denote all-ones and all-zeros vectors, of a shape implied by context. If  $\mathbf{u}$  and  $\mathbf{v}$  are both of shape  $I$ , their *inner product*, denoted either  $\mathbf{u} \cdot \mathbf{v}$  or  $\mathbf{u}^\top \mathbf{v}$ , is the real number  $\sum_{i \in I} \mathbf{u}(i)\mathbf{v}(i)$ . The latter notation will become clear once we introduce the notion of a *matrix*.

A map  $f : \mathbb{R}^I \rightarrow \mathbb{R}^J$  is *linear* iff  $f(\alpha\mathbf{u} + \beta\mathbf{v}) = \alpha f(\mathbf{u}) + \beta f(\mathbf{v})$  for all  $\mathbf{u}, \mathbf{v} \in \mathbb{R}^I$ .

There is a natural 1-1 correspondence between linear maps  $f : \mathbb{R}^I \rightarrow \mathbb{R}^J$  and special vectors  $A \in \mathbb{R}^{J \times I}$  called *matrices*—a fact that is arguably the foundation of linear algebra. Specifically, the matrix  $A \in \mathbb{R}^{J \times I}$  represents the function  $\mathbf{u} \mapsto [\sum_{j \in J} \mathbf{u}(j)A(j, i)]_{i \in I}$ . Two matrices  $A \in \mathbb{R}^{I \times J}$  and  $B \in \mathbb{R}^{J \times K}$  can be multiplied to form a matrix  $AB = \left[ \sum_{j \in J} a_{i,j} \cdot b_{j,k} \right]_{(i,k) \in I \times K}$ ; remarkably, if  $f_A : \mathbb{R}^J \rightarrow \mathbb{R}^I$  and  $f_B : \mathbb{R}^K \rightarrow \mathbb{R}^J$  are the linear maps corresponding to  $A$  and  $B$ , respectively, then  $AB$  is the matrix corresponding to  $f_A \circ f_B$ . Observe that a vector  $\mu \in \mathbb{R}^I$  can be viewed as a matrix  $\mathbf{u} \in \mathbb{R}^{I \times 1}$  for the purposes of composition; in this form, it is called a *column vector*. This establishes that linear maps can be viewed as vectors, and vectors can be viewed as linear maps. The *transpose* of a matrix  $A \in \mathbb{R}^{I \times J}$  is the matrix  $A^\top \in \mathbb{R}^{J \times I}$  that results from swapping the rows and columns of  $A$ . In particular, the transpose  $\mathbf{u}^\top$  of a column vector  $\mathbf{u} \in \mathbb{R}^I$  is a *row vector* or *dual vector*, and represents a linear function  $\mathbf{v} \mapsto \mathbf{u}^\top \mathbf{v} : \mathbb{R}^I \rightarrow \mathbb{R}$ .

**Cones and Convexity.** A subset  $A \subseteq \bar{\mathbb{R}}^I$  of extended real space is a *convex set* iff it contains all of the line segments between pairs of its elements—that is, if it has the property that  $(1 - \lambda)\mathbf{u} + \lambda\mathbf{v} \in A$  for all  $\mathbf{u}, \mathbf{v} \in A$  and  $\lambda \in [0, 1]$ ,  $(1 - \lambda)$ . The probability simplex  $\Delta\Omega$  is the prototypical convex set. The *convex hull* of a subset of  $U \subseteq \bar{\mathbb{R}}^I$  is the smallest convex set that contains  $U$ . The *extreme points* or *vertices* of a convex set  $A$  is the smallest subset of  $A$  whose convex hull is  $A$ . For instance, the convex hull of  $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$  is the unit square  $[0, 1]^2$ , and the extreme points of the unit square are those four points.

A function  $f : \bar{\mathbb{R}}^I \rightarrow \bar{\mathbb{R}}$  is *convex* iff

$$\forall \mathbf{u}, \mathbf{v} \in \bar{\mathbb{R}}^I, \forall \lambda \in (0, 1). \quad f((1 - \lambda)\mathbf{u} + \lambda\mathbf{v}) \leq (1 - \lambda)f(\mathbf{u}) + \lambda f(\mathbf{v}).$$

Recall that a linear map is one for which this relationship always holds with

equality for all  $\mathbf{u}$  and  $\mathbf{v}$ . A *strictly convex* function, on the other hand, is one where it never does, except when  $\mathbf{u} = \mathbf{v}$ .

Convex functions and convex sets are closely related. The set of points that lie above a function  $f$  is a convex set if and only if  $f$  is convex. The level sets of a function  $f$  (i.e.,  $\{\mathbf{u} \in \bar{\mathbb{R}}^I : f(\mathbf{u}) \leq k\}$  for some  $k \in \bar{\mathbb{R}}$ ) are convex if  $f$  is convex. And  $\mathbf{u} \mapsto \infty \cdot \mathbb{1}[\mathbf{u} \in A]$  is a convex function iff  $A$  is a convex set.

A *cone* is something between a linear space and a convex set: a linear space is one closed under linear combination with arbitrary coefficients, while a cone is one closed under linear combination with non-negative coefficients, and a convex space is closed under linear combination with non-negative coefficients that sum to one. It follows that linear spaces are cones, and cones are convex.

## 2.2 Variables

Intuitively, a *variable* represents some feature of the world or some property of some object. Although variables are widely used across computer science—in programming languages, graphical models, causality, and probability theory, to name a few—the term is actually shared by several different formalisms. The account presented here simultaneously explicates many of them.

Mathematically, our notion of a variable exists on two levels. *Qualitatively*, a variable is just some unique identifier (the variable name), such as “Height”, or  $X$ . We typically use capital roman letters for variables. *Quantitatively*, a variable  $X$  is also associated with a set  $\mathcal{V}(X)$ , or simply  $\mathcal{V}X$ , of possible values. For example,  $\mathcal{V}(\text{Height})$  might be the set of positive real numbers, or the set  $\{\text{short}, \text{tall}\}$ . A *binary variable* is one whose possible values are  $\{0, 1\}$ , and a *real variable* is one

whose possible values are  $\mathbb{R}$ . A *constant* is a variable  $X$  that can only take on one possible value (i.e.,  $|\mathcal{V}X| = 1$ ).

**Joint Variables.** We can regard sets of variables  $\mathbf{X}$  as variables themselves, with  $\mathcal{V}\mathbf{X} = \prod_{X \in \mathbf{X}} \mathcal{V}X$ . However, when we do so, we must also be careful to remember  $\mathbf{X}$  the variable has a special relationship with  $\mathbf{X}$  the set of variables. By this definition, the empty set  $\emptyset$  is a variable, and it takes on a single value (i.e., the unique setting of no variables).<sup>1</sup> Those with keen attention to typography may have noticed we have used a different symbol ( $\emptyset$ ) for the empty set of variables, than we have for other empty sets ( $\emptyset$ ). This will be necessary to avoid a minor ambiguity, because the (standard) notation for probability has a different meaning applied to a set of variables than it does applied to a set of possible outcomes.

Similarly, tuples  $(X_1, \dots, X_n)$  of variables and indexed sets  $\{X_i\}_{i \in I}$  of variables, are also themselves variables, with

$$\mathcal{V}(X_1, \dots, X_n) = \{(x_1, \dots, x_n) : \forall i \in [n]. x_i \in \mathcal{V}(X_i)\},$$

and, more generally,  $\mathcal{V}\{X_i\}_{i \in I} := \prod_{i \in I} \mathcal{V}(X_i)$ . The distinction between sets and indexed sets (such as tuples) of variables is only relevant when variables are not unique. In this text, we implicitly convert between the two representations as is convenient, and often view a set of variables as a tuple with a specific order for the sake of presentation. Given a joint setting  $\mathbf{x} \in \mathcal{V}(\mathbf{X})$ , we write either  $\mathbf{x}[X]$  or  $X(\mathbf{x})$  for the value of the variable  $X$  in the joint setting  $\mathbf{x}$ . The former is familiar lookup notation from many programming languages, and the latter is random

---

<sup>1</sup>This might seem strange, but there is precisely one way of selecting a value for each variable in  $\emptyset$ : simply do nothing. To see this another way, observe that if  $\mathbf{X} = \{X_1, \dots, X_n\}$  is a set of binary variables, then  $|\mathcal{V}\mathbf{X}| = 2^n$ , so in the special case of  $n = 0$ , we have  $|\mathcal{V}\emptyset| = 2^0 = 1$ .

variable notation, which we will get to in the next section.

**Variable Equality, Truth, and Functions of Variables.** Variable notation is designed to be intuitive and easy to read. Yet beneath the surface, it is unlike almost any other standard mathematical notation, which can lead to confusion if closely scrutinized. For instance, if  $X$  is a variable and  $x \in \mathcal{V}X$ , then we might expect  $X=x$  to be a false statement, since  $X$  and  $x$  refer to very different objects; instead, it represents the possibility that  $X$  happens take on the value  $x$ , which we will soon see is called an *event* in the context of probability. Furthermore, if  $X$  and  $Y$  are both variables, one might expect  $X=Y$  to be true iff  $X$  and  $Y$  stand for the same variable, but here again it instead represents the possibility that  $X$  and  $Y$  happen to take on the same value. More generally, if  $X$  is a variable and  $f : \mathcal{V}X \rightarrow S$  is a function of the value of  $X$ , then we can regard  $f(X)$  itself as a variable, with  $\mathcal{V}(f(X)) = S$ . (The behavior of equality is just the special case when we view  $= : \mathcal{V}(X, Y) \rightarrow \{0, 1\}$  as such a function.) However, as when forming joint variables, we must also keep in mind when we do so that the variable  $f(X)$  and the variable  $X$  have a constrained relationship.

The solution requires us to keep track of these constraints. Let  $\Gamma(X)$  denote the set of constraints involved in forming the variable  $X$ . Now, when we write  $X = Y$  we primarily mean the possibility that  $X$  and  $Y$  take the same value, which is a rich mathematical object (e.g., an event). At the same time, we can also implicitly convert that object to a truth value, if a truth value is required. Specifically, if the construction of  $X$  involves constraints on the set of variables  $\mathbf{X}$ , and the construction of  $Y$  involves constraints on the variables  $\mathbf{Y}$ , then we say  $X=Y$  is *true* iff  $w[X] = w[Y]$  for all  $w \in \mathcal{V}(\mathbf{X} \cup \mathbf{Y})$  that satisfy the constraints of  $\Gamma(X)$  and  $\Gamma(Y)$ . A variable  $X$  *primitive* if  $\Gamma(X)$  is empty, and we typically write

$\mathcal{X}$  for the set of all primitive variables in the current context.

This approach gives us what we want in most cases. It is always true that  $X = X$ , but not in general true that  $X = X'$  (unless  $X$  and  $X'$  must be equal by construction, or represent the same constant). It is true that  $f(X) = g(X)$  iff  $f = g$ . If  $X$  and  $Y$  are real variables, then (it is true that)  $X + Y = Y + X$ . Yet these expressions also remain unambiguously meaningful for describing events (as we will soon see). Indeed expressions that are *true*, correspond to events that must necessarily occur.

Typically one deals with these problems by adopting only fragments of this picture, or by sweeping ambiguities under the rug. But it isn't necessary to do so; all of these standard concepts can coexist in one unambiguous formal context. The constraint-tracking resolution given here is just a taste of the general modeling approach at the heart of our theory. The biggest missing part, as one might guess, is probability.

## 2.3 Probability

Probability is a foundational concept across many disciplines, and especially for us. Unfortunately, its notation is almost as diverse as the people who use it. Our approach simultaneously enables some of the most common notations and uses the standard foundations. Thus, although all of the pieces are likely to be familiar, the precise formal way it fits together may not be. We start by giving the simpler picture of probability in the finite case ([Section 2.3.1](#)), which, for our purposes, is largely representative of how one should think of probability in the general case (which we return to [Section 2.3.2](#)).

### 2.3.1 Probability in the Finite Case

**Measures and Conditioning.** Suppose that  $\Omega$  is a finite set. A *probability distribution*  $\mu$  over  $\Omega$  is essentially a function  $\mu : \Omega \rightarrow [0, 1]$  such that  $\sum_{\omega \in \Omega} \mu(\omega) = 1$ . This means  $\mu$  is a vector of shape  $\Omega$  satisfying  $\mu \geq 0$  and  $\mu \cdot \mathbf{1} = 1$ . The distribution that places all mass on a single outcome  $\omega \in \Omega$  is written  $\delta_\omega$ . The *probability simplex*  $\Delta\Omega$  is defined as the set of all probability distributions over  $\Omega$ . The *support* of a distribution  $\mu$  is the set  $\text{Supp } \mu := \{\omega \in \Omega : \mu(\omega) > 0\}$  of outcomes that have positive probability. For  $U \subseteq \Omega$ , called *events*, define  $\mu(U) := \sum_{x \in U} \mu(x)$ . While we just defined  $\mu$  on events  $U$  of  $\Omega$  in terms of  $\mu(\omega)$  for elements  $\omega \in \Omega$ , we remark that, in general (i.e., beyond the finite case), it has to be the other way around: a probability distribution is an object that assigns probability to events in a way that is compatible with intuitions from the finite case we have just described. We return to this in [Section 2.3.2](#).

The standard way of updating a probability distribution based on new evidence is conditioning. Given a subset  $A \subseteq \Omega$ , the conditional measure  $\mu | A$  is defined by  $(\mu | A)(x) := \frac{1}{\mu(A)} \mathbb{1}[x \in A] \mu(x)$ . That quantity is also given a shorter and more common name:  $\mu(x | A)$ . From this definition, it is easy to recover the usual definition of conditional probability:  $\mu(U | A) = \mu(U \cap A)/\mu(A)$ . There is just one caveat: this formula is not meaningful when  $\mu(A) = 0$ ; in this case, we leave the conditional probability undefined. But it is not undefined because there is no value that makes sense—but rather because there is an equally strong argument suggesting that *every* value makes sense: if  $\mu(A) = 0$ , then for all  $p \in [0, 1]$  there is a sequence  $\{\mu_k\}_{k \in \mathbb{N}}$  of probability distributions converging to  $\mu$  with the property that  $\forall k. \mu_k(U | A) = p$ .

**Probability via Variables.** Our notation for representing and manipulating probability distributions will make heavy use of the notion of a variable, as developed in Section 2.2. Namely, we almost always work with probability distributions over the values of a variable. This is just a notational convenience, since we can always define a variable  $W$  with possible values  $\mathcal{V}W := \Omega$ . Working with variables confers us a number of benefits, the first of which is to enable standard notation in the AI community: given a distribution  $\mu \in \Delta^{\mathcal{V}X}$  over the values of  $X$ , we also write  $\mu(X)$  to specify  $\mu$  and its type in a compact way. Since  $\mu : \mathcal{V}W \rightarrow [0, 1]$  is a function, we have already defined  $\mu(W)$  to be a variable, but fortunately that variable encodes exactly the same information as  $\mu$  itself.

A *joint distribution*  $\mu(\mathbf{X}) = \mu(X_1, \dots, X_n)$  over a finite (indexed) set of variables  $\mathbf{X} = \{X_1, \dots, X_n\}$  is just a distribution over joint settings of all variables—i.e., a distribution over  $\mathcal{V}\mathbf{X}$ . If  $\mu(\mathbf{X})$  is a joint distribution and  $X \in \mathbf{X}$ , then we write  $\mu(X)$  for its *marginal* on  $X$ , which is given by

$$\mu(X=x) = \mu(X)(x) := \sum_{\substack{\mathbf{x} \in \mathcal{V}\mathbf{X} \\ \mathbf{x}[X]=x}} \mu(\mathbf{x}).$$

It is called a marginal because of the special case of two variables: a joint distribution  $\mu(X, Y)$  may be viewed as a grid of numbers summing to 1, whose rows correspond to values of  $X$ , and whose columns correspond to values of  $Y$ . If we were to write the sum  $\sum_{y \in \mathcal{V}Y} \mu(x, y)$  of each row  $x$  in the margin, this would give the marginal distribution on  $X$ .

To *condition* a joint distribution  $\mu(\mathbf{X})$  on a variable  $X \in \mathbf{X}$  is to form an indexed set of conditioned distributions  $\mu(\mathbf{X} \mid X) = \{(\mu|X=x)\}_{x \in \mathcal{V}X}$ , one for each possible value of  $X$ . We can also refer to such objects without reference to a joint distribution. A *conditional probability distribution (cpd)* on  $Y$  given  $X$  is a function  $p : \mathcal{V}X \rightarrow \Delta^{\mathcal{V}Y}$  that, for each  $x \in \mathcal{V}X$ , yields a probability distribution

over  $Y$ , which is written in any of the following equivalent notations:

$$p(x) = p|x = p(Y|X=x) = p(Y|x) \in \Delta \mathcal{V}Y.$$

Just like we write  $\mu(X)$  for a distribution over the values of  $X$ , we write  $p(Y|X)$  for a cpd on  $Y$  given  $X$ . If  $p(Y|x)$  gives probability one to a single value of  $Y$  for all  $x \in \mathcal{V}X$ , then  $p(Y|X)$  is *deterministic*, and has the same meaning as the function  $f : X \rightarrow Y$  that maps  $x$  to the value  $y$  for which  $p(Y|x)$  has probability one. Conversely, given a function  $f : \mathcal{X} \rightarrow \mathcal{Y}$ , we write  $\delta f(Y|X)$  for the corresponding deterministic cpd. In the finite case, cpds coincide with *stochastic matrices*: matrices whose columns correspond to values of  $Y$ , rows correspond to values of  $X$ , and such that the sum of each row equals 1. We will use **alternate notation** (such as  $\mathbf{P}_{Y|X}$ ) when we want to view  $p(Y|X)$  as a matrix.

**will we?**

We typically write  $\mathcal{X}$  for the set of all variables relevant to a given context. If  $X, Y \in \mathcal{X}$ , then we write  $\mu(Y|X)$  for the cpd given by first conditioning  $\mu(\mathcal{X})$  on  $X$ , and then marginalizing each conditional distribution to  $Y$ . By regarding  $\mu(Y|X)$  as a function of  $Y$  and  $X$ , it too can be viewed a real-valued variable (just like  $\mu(X)$ ). We can now state *chain rule for probability*, a simple but incredibly important fact:  $\mu(X, Y) = \mu(X)\mu(Y|X)$ . This standard formula is typically viewed as shorthand for

$$\forall (x, y) \in \mathcal{V}(X, Y). \quad \mu(X=x, Y=y) = \mu(X=x)\mu(Y=y|X=x),$$

a view that also demands a caveat: “whenever  $\mu(X=x) > 0$ , so that  $\mu(Y=y|X=x)$  is defined”. However, since  $\mu(X, Y)$ ,  $\mu(X)$ , and  $\mu(Y|X)$  are all variables, the equality  $\mu(X, Y) = \mu(X)\mu(Y|X)$  is simply true without caveats. Why? When  $\mu(X=x) = 0$ , then  $\mu(Y=y|X=x)$  is undefined, meaning it has no definitional constraint (i.e.,  $\Gamma(\mu(Y=y|X=x)))$ —but no matter what value we select for

Needs \Delta

$\mu(Y=y|X=x)$ , the result still holds, because the question is whether or not that value times zero equals zero.

**Variables and Random Variables.** In the context of a probability distribution  $\mu \in \Delta\Omega$ , a *random variable* is a function  $X : \Omega \rightarrow \mathcal{V}X$ , where  $\mathcal{V}X$  is just some set (unfortunately) called the *domain* of  $X$  (even though  $\Omega$  is the domain of the function). Despite the superficial similarity induced by our choice of notation, random variables and variables are different formal objects; hence the mantra, “a random variable is neither random nor a variable”. Yet at an even deeper level, variables and random variables are interchangeable, and fundamentally represent the same concept, as we now explore.

Whereas the traditional foundations of probability start with a sample space  $\Omega$  and defines random variables  $X : \Omega \rightarrow \mathcal{V}X$  in terms of it, probabilistic modeling is often done the other way around in practice: one starts with variables  $\mathcal{X}$  of interest, and defines  $\Omega := \mathcal{V}\mathcal{X}$  to be the joint settings of those variables. Under that definition of  $\Omega$  a variable  $X \in \mathcal{X}$  can be viewed as a random variable  $\hat{X} : \mathcal{V}\mathcal{X} \rightarrow \mathcal{V}X$  that selects the value of  $X$  from a joint setting of all variables. The converse not well appreciated, perhaps because it requires one to think about the constraints involved in defining a variable. If  $\Omega = \Delta\mathcal{V}X$ , and  $\hat{Z} : \mathcal{V}\mathcal{X} \rightarrow \mathcal{V}Z$  is an arbitrary random variable, we can regard  $\hat{Z}$  as a variable  $Z$ , and include it in the set  $\mathcal{X}' := \mathcal{X} \cup \{\hat{Z}\}$  of all variables. When we do so, there is a natural 1-1 correspondence between distributions  $\mu \in \Delta\mathcal{V}\mathcal{X}$  over the original variables, and “extended” distributions  $\bar{\mu} \in \Delta\mathcal{V}\mathcal{X}'$  in which variable  $Z$  and the function  $\hat{Z}$  applied to the other variables coincide (formally, where  $\bar{\mu}(\hat{Z}(\mathcal{X}) = Z) = 1$ ).

Although the two perspectives seem to be equally expressive, we opt for the

one in which the variables  $\mathcal{X}$  are primitive, rather than the one in which the sample space  $\Omega$  is primitive. After all, we are interested in modeling epistemic states that undergo structural changes, and it is more natural to add and remove variables of interest, than it is to change sample spaces and figure out how to properly translate the relevant random variables on them. We will see an even deeper connection between the two ways of thinking about (random) variables in Chapter 13.

**Independence.** Suppose that  $X, Y$ , and  $Z$  are (random) variables and  $\mu(X, Y, Z)$  is a distribution over them. The variables  $X$  and  $Y$  are *independent* (according to  $\mu$ ) iff  $\mu(X, Y) = \mu(X)\mu(Y)$ . Often it is the case that two variables are independent only when we control for something else. The variable  $X$  is *conditionally independent of  $Y$  given  $Z$*  (according to  $\mu$ ), denoted  $\mu \models X \perp\!\!\!\perp Y \mid Z$ , only if  $\mu(X|Z)\mu(Y|Z) = \mu(X, Y|Z)$ , or equivalently, if  $\mu(X, Z)\mu(Y, Z) = \mu(X, Y, Z)\mu(Z)$ . Precisely the same formulas define (conditional) independence when  $X, Y$ , and  $Z$  are not events, but rather events (modulo the fact that conditioning on an event of probability zero is undefined).

**Expectation.** If  $\mu$  is a probability over  $\Omega$ , and  $X$  is random variable whose possible values are vectors, then the *expectation* of  $X$  with respect to  $\mu$  is intuitively its mean or average value. Depending on which is clearer in context, an expectation may be notated in either of the following two ways:

$$\mathbb{E}_{\omega \sim \mu} [X(\omega)] = \mathbb{E}_\mu[X] := \sum_{\omega \in \Omega} \mu(\omega)X(\omega).$$

In some cases, we omit the brackets to reduce clutter and improve readability. For a fixed random variable  $X$ , the expectation operator  $\mu \mapsto \mathbb{E}_\mu[X] : [0, 1]^{\mathcal{V}X} \rightarrow \bar{\mathbb{R}}^I$  is a linear function of the probability distribution  $\mu$ .

Expectations and probabilities are closely related to convexity. For instance, the convex hull of  $A \subseteq \bar{\mathbb{R}}^I$  is set of possible expectations over the extreme points of  $A$ , or more formally,

$$\text{conv}(A) = \text{im} \left( \Delta\text{Ext}(A) \ni \mu \mapsto \mathbb{E}_{v \sim \mu}[v] \right).$$

Thus, if  $A$  is convex, to specify  $a \in A$ , it suffices to specify a probability distribution over  $\text{Ext}(A)$ . The components of a distribution  $\mu \in \Delta\text{Ext}(A)$  are called *barycentric coordinates* for the element  $a = \mathbb{E}_{v \sim \mu}[v] \in A$ . A cpd  $p(Y|X)$  can be viewed as providing a barycentric coordinate system for  $\Delta VY$  whose vertices correspond to  $VX$ . This way of thinking about conditional probability distributions will at times be quite useful (e.g., in [Section 4.3.1](#)).

### 2.3.2 Probability in General

The time has come for us to bite the bullet and describe how probability actually works, when a finite representation is insufficient. The material here can be interesting and instructive, and it is certainly a necessary foundation for our results that talk about distributions over real numbers. Still, we reiterate that the full account of measure theory needed to define probability in general is more than is necessary to get a deep conceptual understanding of the present work.

**Measures.** A *measurable space*  $(\Omega, \mathcal{F})$  is a set  $\Omega$ , called the *outcome space*, together with a collection  $\mathcal{F} \subseteq 2^\Omega$  of subsets of  $\Omega$  called *events*, or *measurable sets*. The elements of  $\Omega$ , should be thought of *possible outcomes* or *possible worlds* for in the model. The set  $\mathcal{F}$  of events must also be a  $\sigma$ -algebra, meaning that it contains  $\Omega$ , and is closed under complement and countable union—that is,  $\Omega \in \mathcal{F}$ , if  $U \in \mathcal{F}$

then  $\Omega \setminus U \in \mathcal{F}$ , and if  $U_1, U_2, \dots \in \mathcal{F}$ , then  $U_1 \cup U_2 \cup \dots \in \mathcal{F}$ . When  $\Omega$  is finite, we allow all subsets to be measured by defining  $\mathcal{F} := 2^\Omega$ , and there is also a standard choice for most other spaces of interest, such as when  $\Omega$  is a convex subset of real numbers, called its *Standard Borel Space*. Either way, we will typically leave  $\mathcal{F}$  implicit after this chapter, referring to a measurable space with the same symbol as its sample space  $\Omega$ .

A *measure*  $\mu$  over a measurable space  $(\Omega, \mathcal{F})$  is a function  $\mu : \mathcal{F} \rightarrow [0, \infty]$  with two additional properties:  $\mu(\emptyset) = 0$ , and, for every countable collection  $\{U_i\}_{i \in \mathbb{N}}$  of pairwise disjoint measurable sets (i.e., where each  $U_i \in \mathcal{F}$  and  $U_i \cap U_j = \emptyset$ ), we have that  $\sum_{i \in \mathbb{N}} \mu(U_i) = \lambda(\bigcup_{i \in \mathbb{N}} U_i)$ . A *probability distribution* (or probability measure) is a measure  $\mu$  with the additional property that  $\mu(\Omega) = 1$ . These three properties are quite intuitive:  $\emptyset$  is an event that cannot occur and hence has probability zero,  $\Omega$  is an event that always occurs and hence has probability 1, and if events cannot co-occur, then the probability that one of them occurs is the sum of their probabilities.

**Statistics and (Random) Variables.** If  $(\Omega, \mathcal{F})$  and  $(Y, \mathcal{G})$  are two measurable spaces, then a function  $f : \Omega \rightarrow Y$  is *measurable* if  $f^{-1}(U) \in \mathcal{F}$  for all  $U \in \mathcal{G}$ . In a context where one is interested in probabilities over  $\Omega$ , a measurable function  $X : \Omega \rightarrow V$  is also called a *random variable* or a *statistic*. For our purposes, measurable functions have two key properties: (1) they are closed under composition (so  $f \circ g$  is measurable if both  $f$  and  $g$  are), and (2) if  $f : (\Omega, \mathcal{F}) \rightarrow \bar{\mathbb{R}}^n$  is a measurable function,  $\mu$  is a measure on  $(\Omega, \mathcal{F})$ , and  $A \in \mathcal{F}$ , then we can form the (Lebesgue) integral  $\int_A f d\mu$  to obtain an element of  $\bar{\mathbb{R}}^n$ .<sup>2</sup>

---

<sup>2</sup>To be certain that nothing goes wrong, it is typically also necessary to assert that either  $f \geq 0$  or  $\int_A |f_i| < \infty$  for all  $i \in [n]$ .

We now present an important way of handling concerns about measurability, and bringing the formalism closer to the finite case. The idea is to endow each measurable space with a fixed “base measure”, and describe all measures on a space relative to that measure. Indeed, this is what we will need to develop the analogues of information theoretic quantities that extend beyond the finite case. To make this precise, we need:

**The Radon-Nikodym Derivative.** Suppose  $\mu$  and  $\nu$  are both measures over a measurable space  $(\Omega, \mathcal{F})$ . If  $\nu(U) = 0$  implies  $\mu(U) = 0$  for all  $U \in \mathcal{F}$ , then  $\mu$  is *absolutely continuous* with respect to  $\nu$ , also written  $\mu \ll \nu$ . When  $\mu \ll \nu$ , the Radon-Nikodym theorem ([Nikodym 1930](#)) states that there exists a unique measurable function  $f : \Omega \rightarrow \mathbb{R}$  such that, for all  $A \in \mathcal{F}$ ,  $\mu(A) = \int_A f d\nu$ . This function  $f$  is called the *Radon-Nikodym derivative* of  $\mu$  with respect to  $\nu$ , and is denoted  $\frac{d\mu}{d\nu} := f$ .

add: and otherwise define  $d\mu/d\nu = \infty$ ?

In this more general setting, we no longer require  $\mathcal{V}X$  to be a finite set, but in exchange, we require some structure:  $\mathcal{V}X$  has to be a measurable space (i.e., come equipped with a  $\sigma$ -algebra). Furthermore, we will also assume that  $\mathcal{V}X$  comes equipped with a measure  $\lambda_X$  that we call the *base measure*. In the case of a discrete variable  $X$  that takes on finitely or countably many values, we take  $\lambda_X$  to be the counting measure: for  $U \subseteq \mathcal{V}X$   $\lambda_X(U) = |U|$  is the number of elements in  $U$ . When  $\mathcal{V}X$  is a subset of real space  $\mathbb{R}^n$ , we use the *Lebesgue measure*, which agrees with the standard notions of length, area, and volume.

All of this new structure is compatible with our discussion of variables in [Section 2.2](#). This is because all of our ways of constructing variables—through (indexed) sets of variables, and by applying functions—work equally well over

measurable spaces (with a base measure).

1. Recall that a tuple  $\mathbf{X} = (X_1, \dots, X_n)$  of variables can itself be viewed as itself a variable, with  $\mathcal{V}\mathbf{X} = \prod_{i=1}^n \mathcal{V}X_i$ .  $\mathcal{V}\mathbf{X}$  is a measurable space, because if each  $\mathcal{V}(X_i)$  comes with a  $\sigma$ -algebra  $\mathcal{F}_i$ , then  $\mathcal{V}\mathbf{X}$  naturally inherits the product algebra  $\mathcal{F}_{1:n}$  generated by complements and countable unions of events of the form  $U = U_1 \times \dots \times U_n$  with each  $U_i \in \mathcal{F}_i$ . Similarly, it inherits the unique product measure  $\lambda_{\mathbf{X}}$  whose action on such primitive measurable sets is given by  $\lambda_{\mathbf{X}}(U_1 \times \dots \times U_n) := \lambda_{X_1}(U_1)\lambda_{X_2}(U_2) \cdots \lambda_{X_n}(U_n)$ .
2. Similarly, recall that when  $X$  is a variable and  $f : \mathcal{V}X \rightarrow S$  is a function, we regard  $f(X)$  as a variable. In this setting, we require that  $S$  come equipped with a  $\sigma$ -algebra  $\mathcal{G}$  with respect to which  $f$  is a measurable function. The base measure  $\lambda_{f(X)}$  is then the *pushforward* of  $\lambda_X$  through  $f$ , given by  $\lambda_{f(X)}(U) := \lambda_X(f^{-1}(U))$ , for  $U \in \mathcal{G}$ .

Why do we need all of this? Recall in the finite case, that when  $X$  is a variable and  $\mu \in \Delta \mathcal{V}X$ , then  $\mu(X) : \mathcal{V}X \rightarrow \mathbb{R}$  was a random variable. In the more general setting, the appropriate analogue is  $\frac{d\mu(X)}{d\lambda_X}$ , the Radon-Nikodym derivative of  $\mu$  with respect to the base measure, which is the random variable of interest. For finite variables, this definition coincides with the *probability mass function* (*pmf*) that we were using before; for continuous variables, it coincides with the *probability density function* (*pdf*) of the distribution. Having such an analogue of a random variable corresponding to a measure  $\mu$  will be particularly important for generalizing information-theoretic quantities to continuous variables (Section 2.6).

**Markov Kernels.** Finally, we get to the precise notion of a conditional probability distribution (cpd) in this more careful framework regarding measurability. If  $(\Omega, \mathcal{F})$ , and  $(\Omega', \mathcal{G})$  are two measurable spaces, a *Markov Kernel*  $\kappa$  from  $(\Omega, \mathcal{F})$  to  $(\Omega', \mathcal{G})$  is a function  $\kappa : \Omega \times \mathcal{G} \rightarrow \mathbb{R}$ , such that

1. For every  $\omega \in \Omega$ , the map  $\kappa(\omega, -) : \mathcal{G} \rightarrow [0, 1]$  is a probability measure on  $\Omega'$ .  
(So  $\kappa$  is also a cpd.)
2. For every  $U \in \mathcal{G}$ , the map  $\kappa(-, U) : \Omega \rightarrow [0, 1]$  is a measurable function from  $\Omega$  to the Borel space  $[0, 1]$ . Or more explicitly: for every open set  $S \subseteq [0, 1]$ , and  $U \in \mathcal{G}$ , we have that  $\{\omega \in \Omega : \kappa(\omega, U) \in S\} \in \mathcal{F}$ .

The second condition is the new one, and it is necessary to ensure that we can compose them and treat them as models of randomized functions, just as in the finite case. The analogue of viewing a cpd  $p(Y|X)$  in the finite case as a random variable, is to implicitly convert the Markov Kernel  $\kappa(Y|X)$  to the measurable function

$$(x, y) \mapsto \frac{d\kappa(x, -)}{d\lambda_Y}(y) : \mathcal{V}(X, Y) \rightarrow [0, \infty].$$

## 2.4 The Basic Theory of Graphs and Hypergraphs

In this dissertation (as in much of mathematics and computer science), the term *graph* refers not to a *graphical* depiction of a function or data, but rather to what is more commonly known as a *network*.

Graphs come in two flavors: directed and undirected.

**Definition 2.1.** A (*directed*) (*multi*)graph  $G = (N, A)$ , or simply a *graph*, is a set  $N$  of nodes, and a collection  $A$  of *directed edges* (or *arcs*), such that each  $a \in A$  has a source node  $S_a \in N$  and a target node  $T_a \in N$ . So, formally, the definition is  $G = (N, A, S, T)$ , with  $S, T : A \rightarrow N$  often left implicit. We write  $u \xrightarrow{a} v \in A$  to indicate that there is some  $a \in A$  with source  $S_a = u$  and target  $T_a = v$ , and we call  $a$  an arc “from  $u$  to  $v$ ”.  $\square$

In a graph  $G = (N, A)$ , the *parents* of a node  $u \in N$  are the nodes

$$\mathbf{Pa}_G(u) := \left\{ v \in N : \exists u \xrightarrow{a} v \in A \right\} \subseteq N$$

that are sources of arcs leading to  $u$ . Dually, the *children* of  $u \in N$  are the nodes  $\mathbf{Ch}_G(u)$  that are targets of some arc whose source is  $u$ . There is no distinction between the two notions if the edges do not have direction.

**Definition 2.2.** An *undirected* (*multi*)graph  $G = (N, E)$  is a set  $N$  of vertices (or nodes) and a set  $E$  of edges, each element  $e \in E$  of which corresponds to an unordered pair of vertices  $\{u, v\}$ . More formally, there is a map

$$\iota : E \rightarrow V \times V \setminus \{(v, v) : v \in N\} / \{(u, v) \sim (v, u) \mid (u, v) \in N \times N\}.$$

implicit in the definition of  $G$ , which we will write  $G = (N, E, \iota)$  only when being extra careful.  $\square$

A *clique* of an undirected graph  $G = (N, E)$  is a subset  $C \subseteq N$  of its vertices that are all connected by edges—that is, such that every distinct pair  $\{x, y\}$  of vertices  $x, y \in C$  is a member of  $E$ . We write  $\text{cliques}(G)$  for the set of all cliques in a graph. We often write  $u - v \in E$  as a slightly more visual alternative to  $\{u, v\} \in E$ , and one that can be chained together. In an undirected graph, a *path* from  $u$  to  $v$  is a sequence of edges  $u = u_1 - u_2 - \cdots - u_n = v$  that connect  $u$  and  $v$ .

A path in a directed graph has the additional requirement that every arc along the path must point from  $u_i$  to  $u_{i+1}$ . A *cycle* is a path that starts and ends at the same node. A *directed acyclic graph*, or simply *dag* for short, is a directed graph without any cycles.

It is common to identify a graph  $H = (N, A)$  (or an undirected graph  $G = (N, E)$ ) with its (symmetric) adjacency matrix

$$\mathbb{A}_H = \left[ \# \left\{ a \in A : \begin{array}{l} S_a = u, \\ T_a = v \end{array} \right\} \right]_{(u,v) \in N \times N} \quad \mathbb{A}_G = \left[ \# \{ e \in E : \iota(e) = \{u, v\} \} \right]_{(u,v) \in N \times N},$$

in part because there is a natural bijection between (undirected) multigraphs and (symmetric) square matrices over the natural numbers. For example:

The diagram shows a red-bordered box containing the text "good place to talk about weights!". To its right is a 3x3 matrix with yellow entries:

$$\begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 0 \\ 2 & 0 & 1 \end{bmatrix}$$

Below the matrix is a small wavy arrow pointing to a graph diagram. The graph consists of three vertices labeled 'c' (top), 'a' (bottom-left), and 'b' (bottom-right). Vertex 'c' is connected to both 'a' and 'b' by edges, forming a triangle.

**There** is an important special class of (undirected) graphs that commonly arises in modeling the relationships between two distinct kinds of objects.

**Definition 2.3.** A bipartite graph  $G = (L, R, E)$  is an undirected graph  $(L \sqcup R, E)$  whose vertices are partitioned into two components  $V = L \sqcup R$  (the *left part* and *right part*) with the property that every edge crosses the partition:  $\forall e \in E. |\iota(e) \cap L| = |\iota(e) \cap R| = 1$ . □

The notion of a graph captures only pairwise relationships between vertices. If we also want to model relationships between triples of vertices, it would appear that we need something more general than a graph.

**Definition 2.4.** A *hypergraph*  $G = (V, \mathcal{E})$  is a set  $V$  of vertices, and a collection  $\mathcal{E}$  of *hyperedges*, which correspond to finite subsets of  $V$ . □

An undirected graph is the special case of a hypergraph in which every hyperedge contains two vertices. In turn, a bipartite graph is a (very) special case of an ordinary undirected graph. By transitivity, one might expect bipartite graphs to naturally be an extremely strict special case of hypergraphs—yet, perhaps counter-intuitively, they are in fact naturally isomorphic.

**Proposition 2.1.** *The functions*

$$h2b(V, \mathcal{E}) := (V, \mathcal{E}, \{(v, E) \in V \times \mathcal{E} : v \in E\})$$

$$h2b^{-1}(L, R, E) = b2h(L, R, E) := (L, \{\{v \in L : (v, r) \in E\} : r \in R\})$$

define an isomorphism between hypergraphs and bipartite graphs.

A second isomorphism can be obtained by swapping roles of the left and right parts of the bipartite graph. This is because there is an obvious symmetry in the definition of a bipartite graph: by swapping  $L$  and  $R$ , one obtains another bipartite graph that is in many ways no different from the original. The corresponding symmetry for hypergraphs, known as duality, is less obvious. The *dual* of the hypergraph  $G = (V, \mathcal{E})$  is the hypergraph  $\check{G} := (\mathcal{E}, \{\{e \in \mathcal{E} : v \in e\} : v \in V\})$ , which in some sense, has swapped the roles of the vertices and the hyperedges. These two views, and the symmetry between them, plays an important role in traditional graphical models, as we will see in [Section 2.5.2](#).

For the development of the ideas in this thesis, we are even more interested in the *directed* analogue of a hypergraph. While these objects have been defined and studied before ([Gallo et al. 1993](#)), they are far less common in computer science than (ordinary) directed graphs, or (undirected) hypergraphs. These directed hypergraphs will form the basis of the representations at the heart of this thesis ([Chapters 3 and 5](#)).

**Definition 2.5.** A *directed hypergraph* consists of a set  $\mathcal{N}$  of nodes and a set  $\mathcal{A}$  of directed hyperedges, or *hyperarcs*; each hyperarc  $a \in \mathcal{A}$  is associated with a set  $S_a \subseteq \mathcal{N}$  of source nodes and a set  $T_a \subseteq \mathcal{N}$  of target nodes. We write  $S \xrightarrow{a} T \in \mathcal{A}$  to specify a hyperarc  $a \in \mathcal{A}$  together with its sources  $S = S_a$  and targets  $T = T_a$ .  $\square$

Nodes that are neither a source nor a target of any hyperarc will seldom have any effect on our constructions; the other nodes can be recovered from the hyperarcs by selecting  $\mathcal{N}_{\mathcal{A}} := \bigcup_{a \in \mathcal{A}} S_a \cup T_a$ . If necessary, we can often add an “identity” arc  $N \rightarrow N$  for any node  $N \in \mathcal{N} \setminus \mathcal{N}_{\mathcal{A}}$ . For these reasons, we often leave  $\mathcal{N}$  implicit, referring to the directed hypergraph simply as  $\mathcal{A}$ .

A directed hypergraph  $(N, \mathcal{A})$  can be equivalently defined as an (ordinary) directed graph  $(2^N, \mathcal{A})$  whose set of nodes is the powerset of some set  $N$ . Just as hypergraphs and bipartite graphs are isomorphic, so too are directed hypergraphs and directed bipartite graphs:

$$(N, \mathcal{A}) \mapsto \left( N, \mathcal{A}, \{u \rightarrow a : a \in \mathcal{A}, u \in S_a\} \cup \{a \rightarrow v : a \in \mathcal{A}, v \in T_a\} \right)$$

$$(L, R, A) \mapsto (L, \{\mathbf{Pa}(r) \rightarrow \mathbf{Ch}(r) : r \in R\}).$$

Interestingly, the corresponding concept of duality in directed hypergraphs is clearer than this isomorphism, and even more symmetric than its undirected counterpart. Despite this, we are not aware of any thorough treatment of duality in directed hypergraphs.

**Definition 2.6.** The *dual* of a directed hypergraph  $\mathcal{H} = (N, \mathcal{A})$  is  $\check{\mathcal{H}} := (\mathcal{A}, N)$ , where  $\check{S}_n = \{a \in \mathcal{A} : n \in T_a\}$  and  $\check{T}_n = \{a \in \mathcal{A} : n \in S_a\}$ .  $\square$

It is easy to verify that, as with undirected hypergraphs, the dual of the dual

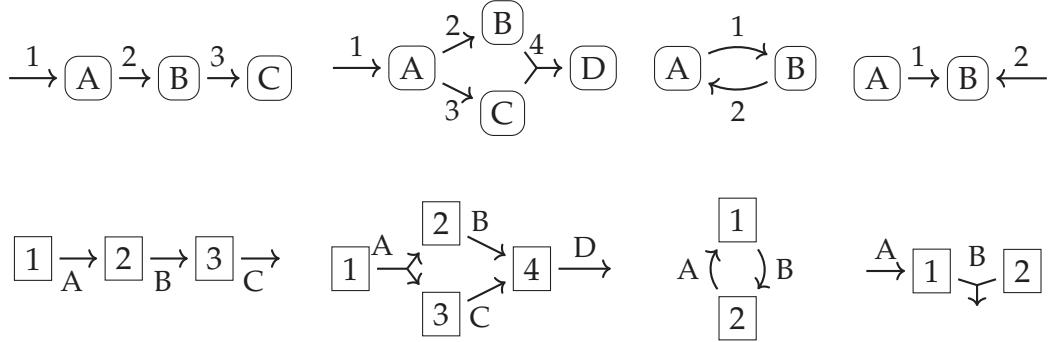


Figure 2.1: Examples of directed hypergraphs (one row) and their duals (the other row).

of a directed hypergraph is the original.<sup>3</sup> The operation is worth visualizing. A few hypergraphs and their duals are depicted in Figure 2.1. A hypergraph and its dual correspond to two different ways of notating a composite process: as a *flowchart* or as a *circuit* (see [Kasangian and Walters \(1990\)](#) for a discussion of this duality in a related context). In a flowchart, nodes represent state and arcs represent transformations between them. In a circuit, nodes represent operations, and arcs, thought of as “wires” that connect them, carry state. So, if capital letters represent variables and carry state, while numbers represent (randomized) transitions between variables (i.e., cpds), then the top row of Figure 2.1 is the flowchart representation and the bottom row is the circuit representation.

In either representation, the two hypergraphs on the left of Figure 2.1 represent unambiguous composite processes. The far left one is a chain: operation  $1$  ( $op_1$ ) produces  $X$ , which is fed to  $op_2$  which produces  $Y$ , and finally fed to  $op_3$  to produce  $Z$ . The second hypergraph from the left involves a split and a merge:  $op_1$  produces  $A$ , which is copied and sent to  $op_2$  and  $op_3$ , which use it to produce  $B$  and  $C$ , respectively. Then,  $op_4$  uses the values of  $B$  and  $C$  to produce  $D$ . The middle right hypergraph represents a cyclic process, in which  $op_1$  produces  $B$

---

<sup>3</sup> $\check{S}_a = \{n \in N : a \in \check{T}_n\} = \{n \in N : a \in \{a' \in \mathcal{A} : n \in S_{a'}\}\} = \{n \in N : n \in S_a\} = S_a$ , and a symmetric argument establishes that  $T_a = \check{\check{T}}_a$ .

from A and  $op_2$  produces A from B. In each of these three cases, every variable is produced by a single operation. (On the top, this means one hyperarc goes into every node; on the bottom, it means one hyperarc comes out of every node e). But what does the hypergraph on the far right represent?  $Op_1$  produces B from A, but  $op_2$  somehow also produces B. Is this meaningful? Previous approaches to modeling processes rule out situations like these, but in [Chapter 3](#), we will use them to capture (possibly) conflicted beliefs.

For representing processes like the ones on the left, both circuit-style and flowchart-style diagrams are common in computer science. Focusing specifically on machine learning, architecture diagrams<sup>4</sup> are circuit representations that have become the standard way to communicate the architecture of a neural network. Meanwhile, the duals of these hypergraphs (i.e., flowchart variants) are closely related to a standard representation called a *Bayesian Network*, which is a primary focus of the next section.

## 2.5 Probabilistic Graphical Models

Like variables themselves, graphical models exist on two levels: a qualitative or structural level, and a quantitative or observational one. A *graphical model structure*, or a *qualitative graphical model* is a (directed) (hyper)graph whose vertices  $\mathcal{X}$  are variables, and whose (hyper)edges somehow indicate local influences between variables. The quantitative form then annotates those concepts with data. A *probabilistic graphical model*, or simply “graphical model”, is a graphical model structure together with *quantitative* or probabilistic information about

---

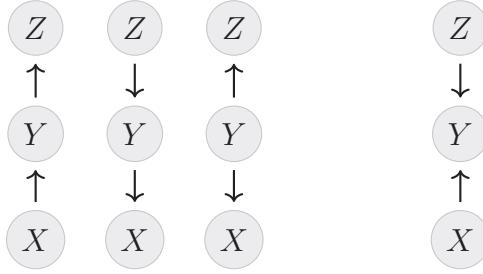
<sup>4</sup>as produced by this library, for example <https://github.com/HarisIqbal88/PlotNeuralNet>

these local influences. The purpose of a graphical model, at least historically, is to be a compact representation of a probability distribution (by exploiting independencies). Thus, a graphical model  $\mathcal{M}$  typically represents a joint probability distribution  $\text{Pr}_{\mathcal{M}} \in \Delta^{\mathcal{V}\mathcal{X}}$  over the values its variables  $\mathcal{X}$ .

We now review the basics of the most important classes of graphical models used in practice: Bayesian Networks, which are based on directed (acyclic) graphs, and what we will call factor graphs, which are based on undirected (hyper)graphs. What follows is really a barebones sketch of the graphical models context for the ideas we present in this dissertation; for a proper treatment, we refer readers to [Koller and Friedman \(2009\)](#).

### 2.5.1 Bayesian Networks and Variants

A *qualitative Bayesian Network* (qualitative BN) is an acyclic directed graph  $G = (\mathcal{X}, A)$  over a set of variables  $\mathcal{X}$ . An arc  $X \rightarrow Y \in A$  intuitively reflects a possibility that  $X$  can influence  $Y$ . Intuitively, each variable should be thought of as being associated with a process that produces it in a way that depends jointly on all of its parents. The absence of an arc, then, in context, reflects a kind of independence. More precisely, a qualitative BN states that *every variable is conditionally independent of its non-descendents given its parents*. Another famous (and more algorithmic) characterization of the same conditional independences is the *d-separation* criterion ([Geiger et al. 1990](#)). It may be helpful to see some examples:



The three qualitative BNs on the left all encode the same independence information:  $X$  and  $Z$  are conditionally independent given  $Y$ . The last one, however, is different: it states that  $X$  and  $Y$  are independent (but  $Z$  can depend arbitrarily on both  $X$  and  $Y$ ). Note also that, while the left three qualitative BNs all are (almost) flowchart representations, the one on the right is not; the two arrows do not represent two different processes, but rather a single joint dependence. This intuition can be further clarified once we add probabilities.

A (*quantitative*) BN  $\mathcal{B} = (\mathcal{X}, A, \mathbb{P})$  is a qualitative BN  $(\mathcal{X}, A)$  together with a collection of cpds  $\mathbb{P} = \{P_X(X \mid \text{Pa}(X))\}_{X \in \text{Pa}(X)}$ , one for each variable given its parents. A BN then induces a joint probability distribution

$$\Pr_{\mathcal{B}}(\mathcal{X}) := \prod_{X \in \mathcal{X}} P_X(X \mid \text{Pa}(X)), \quad (2.1)$$

which indeed satisfies the independence properties specified by its underlying qualitative BN. Furthermore,  $\Pr_{\mathcal{B}}$  has every conditional marginal specified by  $\mathbb{P}$ . Indeed, the factorization (2.1) is yet another equivalent characterization of the conditional independencies of this graph. This quantitative view (2.1) is the primary way that the broader modern machine learning community views BNs. It also suggests a way of generalizing BNs to handle cycles and more: simply use the same equation. However, the resulting object will typically not be a probability distribution: its sum over all values of  $\mathcal{X}$  may not equal 1. But, in principle, it could be normalized; this approach leads to the second major class of graphical models.

### 2.5.2 Markov Random Fields, Factor Graphs, and Variants

Now, suppose  $G = (\mathcal{X}, E)$  is an undirected graph over a set of variables. By interpreting an edge  $X - Y \in E$  as a possibility of (bi-directional) influence between  $X$  and  $Y$ . A distribution  $\mu$  is said to be a *Markov Random Field* for  $G$  if it satisfies one of the following properties:

**(pairwise)** If  $X, Y \in \mathcal{X}$  with  $X \neq Y$  and  $\{X, Y\} \notin E$ , then  $\mu \models X \perp\!\!\!\perp Y \mid \mathcal{X} \setminus \{X, Y\}$ ;

**(local)** For each  $X \in \mathcal{X}$ , we have  $\mu \models X \perp\!\!\!\perp \mathcal{X} - \{X\} \mid \partial_G(X)$ , where  $\partial_G(X) := \{Y \in \mathcal{X} : \{Y, X\} \in E\}$  is the set of variables neighboring  $X$ ;

**(global)** For all  $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathcal{X}$ , if every path from some  $X \in \mathbf{X}$  to some  $Y \in \mathbf{Y}$  goes through some  $Z \in \mathbf{Z}$ , then  $\mu \models \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ ;

**(factorization)** There exists a collection of functions  $\{\phi_{\mathbf{X}} : \mathcal{V}\mathbf{X} \rightarrow \mathbb{R}_{\geq 0}\}_{\mathbf{X} \in \text{cliques}(G)}$  indexed by the cliques of the graph  $G$ , whose product is the distribution  $\mu$ .

$$\text{That is, } \mu(\mathcal{X}) = \prod_{\mathbf{X} \in \text{cliques}(G)} \phi_K(\mathbf{X}).$$

In general, the conditions are distinct, with (global)  $\Rightarrow$  (local)  $\Rightarrow$  (pairwise). However, if we restrict to *positive* distributions (i.e.,  $\forall \omega \in \mathcal{V}\mathcal{X}. \mu(\omega) > 0$ ), then all four conditions are equivalent. This result is known as the *Clifford-Hammersley Theorem*, or the *fundamental theorem of graphical models*. cite!

If we care only about such distributions, this suggests another approach to the same idea: throw away the graph  $G$ , multiply together collection of functions  $\{\phi_j : \mathcal{V}\mathbf{X}_j \rightarrow \mathbb{R}_{\geq 0}\}_{j \in J}$  indexed by some arbitrary set  $J$ , and then renormalize; the subset  $\mathbf{X}_j \subseteq \mathcal{X}$  is called the *scope* of factor  $j$ . Although we have described an annotation of a hypergraph, it is more standard and easier to draw in another form. The bipartite graph that corresponds naturally to this hypergraph is called

a *factor graph*. More precisely, a factor graph has two kinds of nodes: factors (drawn as squares), and variables (circles), with an edge between them if the factor depends on the variable. Some authors reserve the term “factor graph” just for the bipartite graph representation, but we will also refer to the hypergraph and collection of factors this way, as they are essentially the same object. We will see more details in [Section 3.4.2](#).

### 2.5.3 Other Graphical Models

Many other graphical models have been proposed as well, although the vast majority are variants on the two we have already described. We now give a quick overview of some of these representations, although this list is far from complete. These models make only minor appearances in the text, but knowing about them may nonetheless be helpful for situating PDGs in the literature.

The most important for our purposes, will be *conditional* Bayesian Networks and *causal* Bayesian Networks, both of which are called CBNs, and, fortuantely, both of which make the same modification: marking some variables as “observed” so that they do not require probability distributions. Writing  $\mathcal{U}$  for the observed variables and  $\mathcal{V}$  for the unobserved ones, a CBN  $\mathcal{B}$  represents a cpd  $\text{Pr}_{\mathcal{B}}(\mathcal{V} \mid \mathcal{U})$ . *Conditional random fields* (CRFs) are an analogue for factor graphs/MRFs.

Graphs with both directed and undirected edges have also been used as the basis for probabilistic models. The most widely known graphical model of this kind is called a *chain graph*, in part because it can be seen as a way of connecting conneted components (themselves factor graphs) along a directed acylcic chain. People have also considered other ways of making precise the notion of a cyclic

baysian network (aside from multiply together factors for a cyclic network), including unrolling it and looking at fixed points—such as dyanmic Bayesian Networks ([Dean and Kanazawa 1989](#))([Koller and Friedman 2009, §6.2](#)), and a related cut-parameterized approaches to looking at fixed points ([Baier et al. 2022](#)). *Directed factor graphs* ([Frey 2012](#)) combine the unify the qualitative expressiveness of directed graphical models with the (undirected) factor graph formalism, but are quantitatively no different from ordinary factor graphs.

A *Dependency Network* (DN) is another (much less popular) kind of directed graphical model. Syntactically it is like a Bayesian Network, but semantically, it is closer to an undirected graphical model. It is typically cyclic, and defines a probability distribuiton as a fixed point of a Markov Chain. “Consistent” DNs turn out to be an alternate way of parameterizing undirected graphical models. There is also a notion of an *inconsistent* DN, which turns out to be a very special case of a representation developed in this thesis ([Chapter 12](#)).

## 2.6 Information Theory

Information theory is the study of when computational tasks (such as communicating in a noisy environment, or ) are possible, and when they are not. The field is vast and heterogenous, yet much of it is built using a small set of particularly nice ways of quantifying (un)certainty in a probabilistic context, which we will refer to throughout as *information-theoretic primitives*. These quantities are all based on the concept of log probability, or *surprisal*  $I_\mu[X=x]$ , which is

- equal to zero if the event  $X=x$  was certain to happen (i.e., totally unsurprising) but otherwise strictly positive, and

- combines additively over independent events, just like most quantities in science and everyday life (meters, seconds, etc.).

Although not typically thought of in this way, the basic information-theoretic quantities come in two flavors, mirroring the divide between qualitative and quantitative information that we saw in the presentation of graphical models. With the exception of these editorial comments on the theme of qualitative vs quantitative concepts, all of this material can be found in any introduction to information theory (e.g., [MacKay \(2003, Chapter 1\)](#) or [Cover and Thomas \(1991\)](#)).

### 2.6.1 Shannon Entropy and the Information Profile

Let  $\mu$  be a probability distribution, and let  $X, Y, Z$  be (sets of) discrete (random) variables. The *information content* or *surprisal*  $I_\mu[X=x] := \log \frac{1}{\mu(X=x)}$  of an event  $X=x$  quantifies one's surprise at learning  $X=x$ , knowing that  $X \sim \mu(X)$ .

The *entropy* of  $X$  is the uncertainty in  $X$ , when it is distributed according to  $\mu$ , as measured by the number of bits of information needed (in expectation) needed to determine it, if the distribution  $\mu$  is known. It is given by

$$H_\mu(X) := \sum_{x \in \mathcal{V}(X)} \mu(X=x) \log \frac{1}{\mu(X=x)} = -\mathbb{E}_\mu[\log \mu(X)],$$

and has several very important properties. Chief among them,  $H_\mu(X)$  is non-negative, and equal to zero iff  $X$  is a constant according to  $\mu$ .

The “joint entropy”  $H(X, Y)$  is just the entropy of the combined variable  $(X, Y)$  whose values are pairs  $(x, y)$  for  $x \in \mathcal{V}X, y \in \mathcal{V}Y$ ; this is the same as the entropy of the variable  $X \cup Y$  when  $X$  and  $Y$  are themselves sets of variables.

The *conditional entropy* of  $Y$  given  $X$  measures the uncertainty present in  $Y$

if one knows the value of  $X$  (think: the information in  $Y$  but not  $X$ ), and is equivalently defined as any of the following three quantities:

$$H_\mu(Y|X) := \mathbb{E}_\mu[\log^{1/\mu(Y|X)}] = H_\mu(X, Y) - H_\mu(X) = \mathbb{E}_{x \sim \mu(X)}[H_{\mu|X=x}(Y)].$$

The fact that these three quantities coincide is a remarkable fact, and one reason that entropy is particularly special.

**Fact 2.2** (Entropy Chain Rule). *If  $X$  and  $Y$  are random variables, then the entropy of the joint variable  $(X, Y)$  can be written as  $H_\mu(X, Y) = H_\mu(Y | X) + H_\mu(X)$ . It follows that if  $\mu$  is a distribution over the  $n$  variables  $X_1, \dots, X_n$ , then*

$$H(\mu) = \sum_{i=1}^n H_\mu(X_i | X_1, \dots, X_{i-1}).$$

The *mutual information*  $I(X; Y)$ , and its conditional variant  $I(X; Y|Z)$ , are given, respectively, by

$$I_\mu(X; Y) := \mathbb{E}_\mu \left[ \log \frac{\mu(X, Y)}{\mu(X)\mu(Y)} \right], \quad \text{and} \quad I(X; Y|Z) := \mathbb{E}_\mu \left[ \log \frac{\mu(X, Y, Z)\mu(Z)}{\mu(X, Z)\mu(Y, Z)} \right].$$

Both quantities are non-negative. The former is equal to zero iff  $\mu \models X \perp\!\!\!\perp Y$ , and the latter is equal to zero iff  $\mu \models X \perp\!\!\!\perp Y | Z$ ; thus, they measure distance to independence.

Just as conditional entropy can be written as a linear combination of unconditional entropies, so too can conditional mutual information be written as a linear combination of unconditional mutual informations:  $I(X; Y|Z) = I(X; (Y, Z)) - I(X; Z)$ . Thus conditional quantities are easily derived from the unconditional ones. But at the same time, the unconditional versions are clearly special cases of the conditional ones; for example,  $H_\mu(X)$  is clearly the special case of  $H(X|Z)$  when  $Z$  is a constant (e.g.,  $Z = \emptyset$ ). Furthermore, entropy and

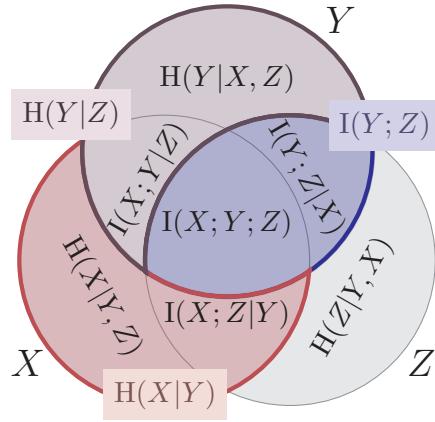


Figure 2.2: The components of the information profile  $\mathbf{I}_\mu$  for three variables  $\mathcal{X} = \{X, Y, Z\}$ .

mutual information are also interdefinable and generated by linear combinations of one another. It is easy to verify that

$$\begin{aligned}\mathbf{I}_\mu(X; Y) &= H_\mu(X) + H_\mu(Y) - H_\mu(X, Y) \\ &= H_\mu(X) - H_\mu(X|Y) \\ \text{and } \mathbf{I}_\mu(X; Y|Z) &= H_\mu(X|Z) + H_\mu(Y|Z) - H_\mu(X, Y|Z) \\ &= H_\mu(X | Y) - H_\mu(X | Y, Z);\end{aligned}$$

thus mutual information is derived from entropy. Yet on the other hand,  $I_\mu(Y; Y) = H_\mu(Y)$  and  $I_\mu(Y; Y|X) = H_\mu(Y|X)$ , at least in the discrete case—thus entropy is also a special case of mutual information.

**Information Diagrams and the Information Profile.** Information theoretic quantities such as (conditional) entropy and (conditional) mutual information for different subsets of a set of variables  $\mathcal{X}$  fit together with an inclusion-exclusion rule. There leads to an easy graphical way to quickly read off identities such as the ones above from an *information diagram*, as depicted in Figure 2.2.

**Entropy for Continuous Variables.** So far, everything we have said relies on intuitions that the variables are discrete. If we continuous variables by discretizing them, it is not hard to see that the entropy as defined goes to  $\infty$  as the discretization becomes fine. Indeed, this is appropriate, because there is a sense in which a real number sampled from  $[0, 1]$  contains infinite information. However, it is difficult to work with divergent limits in this way; instead, we define a much more useful quantity.

In the general case, we need only replace instances of  $\mu(X)$  used as a random variable with its derivative  $\frac{d\mu}{d\lambda}$  with respect to  $X$ 's base measure  $\lambda_X$ . So, for instance, the entropy of a variable  $X$  in the general case is

$$H_\mu(X) = \mathbb{E}_\mu \left[ \log \frac{d\lambda_X}{d\mu(X)} \right] = \int_{\mathcal{V}\mathcal{X}} \log \frac{d\lambda_X}{d\mu(X)} d\mu. \quad (2.2)$$

In the finite case, this is just what we had before. However, it is important to keep in mind that this quantity can be negative in general (because, outside of the finite case, it is possible that  $\frac{d\mu(X)}{d\lambda_X}(x) > 1$ ). The reader should also be aware that, for real-valued variables,  $H(X) = 0$  no longer means that  $X$  is a constant, but rather that  $X$  has expected unit density (like the uniform distribution over  $[0, 1]$ ). Aside from these unwelcome developments, this analogue of entropy acts otherwise just like the original, but shifted by an additive constant to keep it in a more useful range. It still satisfies all of the other properties (such as the chain rule), has the same minima and gradients as the limit of the discrete entropy at infinitely fine discretization. Fortunately, we do not have to lower our standards for (conditional) mutual information, which works just as before: it is non-negative and zero iff there is a (conditional) independence.

**The Principle of Maximum Entropy.** Suppose that, with rigorous empirical investigation, we have determined some statistics about a variable  $X$ —perhaps its mean, and the proportion of the time it takes on a certain value—but have not (yet) collected enough information to uniquely determine the distribution of  $X$ . How should we proceed? The most faithful representation of our knowledge, of course, is just the information we have—those statistics, and perhaps the set of probability distributions consistent with them. But what should we do if we are obligated to give a probability distribution (e.g., to satisfy the axioms of rationality (Vineberg 2022; Halpern 2017))? There is a well-established principle stating that among all possible distributions, the one with maximum entropy is most appropriate. At a high level, this is because entropy measures uncertainty in a distribution, and we do not want to assume things that we do not know.

*“This is the only unbiased assignment we can make; to use any other would amount to arbitrary assumption of information which by hypothesis we do not have.” (Jaynes 1957, p.623)*

Imagine, for example, that our friend is undergoing a secret but important evaluation, and we know nothing about her score except that it will be a number between 1 and 5. It would be inappropriate to assume that she will score a 5 (or a 1) simply because it’s consistent with what we know, because that does not adequately represent our ignorance. The principle of maximum entropy implies that we should assume a uniform distribution over scores until we learn more.

awkward

The principle of maximum entropy has also faced significant criticism, from a number of different angles (Seidenfeld 1986; Cardoso Dias and Shimony 1981; Friedman and Shimony 1971). An recurring theme in these works is the dependence on representation. For example, had we said only “our friend either will

or will not score a 5”, then maximum entropy would have suggested that we should presume probability  $\frac{1}{2}$  that she scores a 5, which is incompatible with the probability  $\frac{1}{5}$  that we were supposed to presume before. Without wading too far into this debate, we point out that it may be problematic to compare our intuition, which has been primed with the context of all five possible scores, with a picture that has willfully ignored that extra context. Whether or not it is appropriate as a general rule, the principle of maximum entropy does not apply locally; if we expect it to work at all, we must represent the entirety of the relevant information.

Many decades later, there is an empirical case in favor of the principle of maximum entropy: it validates many standard and intuitively natural choices made in the literature. The uniform distribution over an interval  $[a, b]$  is the maximum entropy distribution with that support. A gaussian with mean  $m$  and variance  $\sigma^2$  is the maximum entropy distribution with that mean and variance. Among distributions supported on  $[0, \infty]$  with mean  $\lambda$ , the exponential distribution  $\mu(X) = \lambda e^{-\lambda X}$  is the one that maximizes entropy. Logistic regression outputs the maximum entropy distribution over class labels, among those that share certain means with the training data. More generally, *exponential families*, important classes of models that subsume the examples above as well as undirected graphical models, also maximize entropy subject to an expectation matching constraint. add citation!

But what about directed graphical models? Bayesian Networks, so far, have resisted such a clean characterization in terms of maximum entropy; maximum entropy approaches seem to require some sort of deference to a causal structure ([Williamson 2001](#)). A related concern with maximum entropy in the face of conditional information is addressed in ([Grove and Halpern 1997](#)). One contribution

of this thesis is to expand the principle of maximum entropy to account for causal information and conditional constraints ([Section 3.4.1](#)).

**Entropy as a Structural Information Measure.** Throughout this thesis, we will argue that entropy and its ilk are purely “qualitative” or “structural” aspects of a probability distribution (although perhaps the latter term is clearer in this context, because we are still talking about real numbers, i.e., “quantities”). Here is one important reason why: we do not need to know anything about  $\mathcal{V}X$  in order make full use of these quantities. Just like the qualitative notion of independence  $X \perp\!\!\!\perp Y$ , the mutual information  $I(X; Y)$  is fully defined and makes complete sense without knowing what values the variables can take on. Even the notation reflects this. Recall that, at least in the discrete case,  $\mu(X)$ ,  $\mu(X, Y)$  and  $\mu(Y|X)$  can all be viewed as random variables, taking as input values of  $X$  and  $Y$ .  $H_\mu(X)$ ,  $H_\mu(X, Y)$ , and  $H_\mu(Y|X)$ , however, are not functions of the values of  $X$  and  $Y$ , but rather functions of (the identities of) the variables themselves. Indeed, it makes no sense to feed values  $(x, y) \in \mathcal{V}(X, Y)$  to these concepts. In other words, they operate on the qualitative aspect of a variable, not its quantitative aspect.

A second reason to view entropy as structural appears in its general definition ([2.2](#)). There, the derivative is taken with respect to structural information about the variable: the base measure, which describes only agreed-upon coordinates, not observational data about what is likely. But if we replace that base measure with a *probability measure* of our choosing, we get a closely related (but quite different) kind of entropy that is appropriate for working with observational information.

### 2.6.2 Relative Entropy

We now move on to a second kind of information-theoretic primitive, of a more quantitative flavor. If  $\mu, \nu \in \Delta\Omega$ , then the *relative entropy* of  $\nu$  with respect to  $\mu$  is given by:

$$D(\mu \parallel \nu) := \sum_{\omega \in \Omega} \mu(\omega) \log \frac{\mu(\omega)}{\nu(\omega)} = \mathbb{E}_{\omega \sim \mu} \left[ \log \frac{d\mu}{d\nu}(\omega) \right].$$

One critical property of relative entropy, known as *Gibbs' inequality*, is that  $D(\mu \parallel \nu) \geq 0$ , with equality if and only if  $\mu = \nu$  (with probability 1). This is why  $D$  is called a “divergence”. It is not, however, a “distance measure” in the standard sense, because it is not symmetric: typically  $D(\mu \parallel \nu) \neq D(\nu \parallel \mu)$ . But this is a good thing, because the task of discerning one probability distribution from another is inherently asymmetric.

according to \\mu

add link to other example!

Explain principle of minimum discrimination information.

**Example 2.1.** Suppose that there are two coins: a fair one, and a double headed one. One of them (although you do not know which) has been selected, and flipped; you observe the result.

First, suppose you believe the coin is fair, but in fact it is double-headed. You observe the only possible outcome: heads. This is totally within the realm of possibility. It would take many flips to convince yourself the coin was not fair, and even then you could not be completely sure. Now reverse the roles: suppose that in reality, the coin is fair, but you believe it is double-headed. With probability 1/2, you observe a tails and are immediately persuaded (with certainty) that your belief is incorrect.

There is an inherent asymmetry: it is far easier to “distinguish the double-headed coin from the fair one”, than it is to “distinguish the fair coin from the

double-headed one".

△

Relative entropy is arguably exactly the appropriate measure of discrepancy between belief and reality. The most obvious argument for this is the empirical one: relative entropy has become the overwhelmingly standard measure of discrepancy between distributions used to train machine learning systems, where it is standardly known as *Kullback Leibler (KL) divergence*. Chapter 6 can be viewed as a significant strengthening of this argument.

nod to below  
to head off

Here is a more theoretical argument, that allows us to interpret relative entropy in terms of coding theory. The length of an optimal code for a sample  $\omega \sim \nu$  is proportional to its surprisal  $\log 1/\nu(\omega)$ . Imagine that you have belief  $\nu$  and the codes you are using reflect this, while in fact reality is distributed according to  $\mu$ . Upon seeing  $\omega$ , the length of your code is  $\log \frac{1}{\nu(\omega)}$ , while the length of an optimal one would be  $\log \frac{1}{\mu(\omega)}$ . In expectation, the difference between these quantities, which is the overhead of using your mis-specified codes, is  $\mathbb{E}_{\omega \sim \mu}[\log \frac{\mu(\omega)}{\nu(\omega)}] = D(\mu \parallel \nu)$ .

There are a number of other results showing that relative entropy is the unique quantity with important properties of interest; treating them all is beyond the scope of this thesis.<sup>5</sup> Suffice it to say that relative entropy is an extremely special, and a natural measure of quantitative discrepancy between belief. A great deal of this thesis can be viewed as further evidence of this.

---

<sup>5</sup>There are many nice axiomatizations of relative entropy; some important ones include Rényi's ([Rényi 1961](#)), Fadeev's ([Fadeev 1957](#)), Leinster's result showing that it is the only function satisfying three trivial regularity properties and the chain rule ([Leinster 2017](#)), and a characterization as the unique functor from statistical maps to the additive monoid  $([0, \infty], +)$  ([Baez and Fritz 2014](#)). It is the only statistical divergence that is both a Bregman divergence and an f-divergence. The Hessian of relative entropy is the Fisher metric, which is the unique metric tensor that is invariant under sufficient statistics. In that geometry, we will see ([Chapter 12](#)) that the gradient flow of relative entropy with respect to its first ("belief") argument amounts to multiplicative interpolation of probability measures, while gradient flow of relative entropy with respect to its second ("reality") argument amounts to additive interpolation.

**Cross Entropy.** Finally, to head off confusion, we mention that the *cross entropy*

$$\text{CrossEntropy}(p, q) := \mathbb{E}_p \left[ \log \frac{1}{q} \right] = \mathbb{E}_{\omega \sim p} \left[ \log \frac{d\lambda_\Omega}{dq} \right] = D(p \parallel q) + H(p),$$

the standard function used to train a statistical model  $q$  using data  $p$ , is not the same as relative entropy  $D(p \parallel q)$ . However, as can be seen on the right-hand side of the equation above, minimizing it is equivalent to minimizing  $D(p \parallel q)$ , because the entropy  $H(p)$  of the data is a constant and does not affect the optimization process. We view use cross entropy as a small shortcut that simplifies computing with relative entropy in some contexts; we will only use it when comparing with standard learning objectives (in [Chapters 6 and 7](#)).



## **Part I**

# **A Universal Modeling Language**

## CHAPTER 3

### PROBABILISTIC DEPENDENCY GRAPHS (PDGS)

We introduce Probabilistic Dependency Graphs (PDGs), a new class of directed graphical models. PDGs can capture inconsistent beliefs in a natural way and are more modular than Bayesian Networks (BNs), in that they make it easier to incorporate new information and restructure the representation. We show by example how PDGs are an especially natural modeling tool. We provide three semantics for PDGs, each of which can be derived from a scoring function (on joint distributions over the variables in the network) that can be viewed as representing a distribution's incompatibility with the PDG. For the PDG corresponding to a BN, this function is uniquely minimized by the distribution the BN represents, showing that PDG semantics extend BN semantics. We show further that factor graphs and their exponential families can also be faithfully represented as PDGs, while there are significant barriers to modeling a PDG with a factor graph.

#### 3.1 Introduction and Examples

In this chapter we introduce yet another graphical tool for modeling beliefs, called *Probabilistic Dependency Graphs* (PDGs). There are already many such models in the literature, including Bayesian networks (BNs) and factor graphs. (For an overview, see [Section 2.5](#), or better yet, [Koller and Friedman \(2009\)](#).) Why does the world need one more?

Our original motivation for introducing PDGs was to be able capture inconsistency. We want to be able to model the process of resolving inconsistency; to do so, we have to model the inconsistency itself. But our approach to modeling

inconsistency has many other advantages. In particular, PDGs are significantly more modular than other directed graphical models: operations like restriction and combination (Section 3.2.2) that are easily done with PDGs are difficult or impossible to do with other representations. The following examples motivate PDGs and illustrate some of their advantages.

**Example 3.1.** Grok is visiting a neighboring district. From prior reading, she thinks it likely (probability .95) that guns are illegal here. Some brief conversations with locals lead her to believe that, with probability .1, the law prohibits floomps.

add  
values \W  
notation

The obvious way to represent this as a BN is to use two variables  $F$  and  $G$  (respectively taking values  $\{f, \neg f\}$  and  $\{g, \neg g\}$ ), indicating whether floomps and guns are prohibited. The semantics of a BN offer her two choices: either assume that  $F$  and  $G$  to be independent and give (unconditional) probabilities of  $F$  and  $G$ , or choose a direction of dependency, and give one of the two unconditional probabilities and a conditional probability distribution. As there is no reason to choose either direction of dependence, the natural choice is to assume independence, giving her the BN on the left of Figure 3.1.

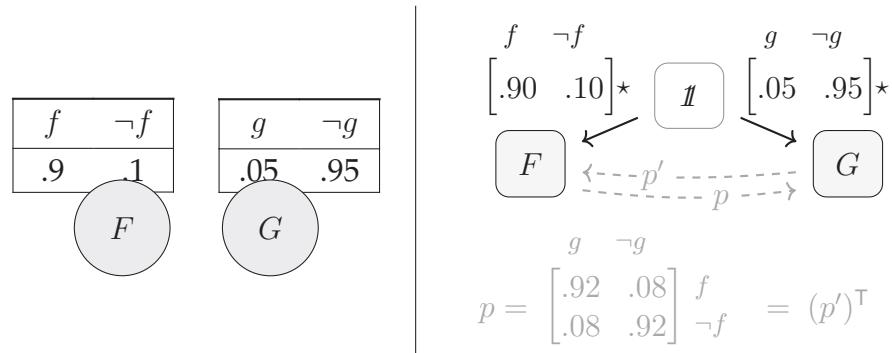


Figure 3.1: A BN (left) and corresponding PDG (right), which can be augmented with additional cpds. The cpds  $p$  and/or  $p'$  make it inconsistent.

A traumatic experience a few hours later leaves Grok believing that “floomp”

is likely (probability .92) to be another word for gun. Let  $p(G \mid F)$  be the conditional probability distribution (cpd) that describes the belief that if floomps are legal (resp., illegal), then with probability .92, guns are as well, and  $p'(F \mid G)$  be the reverse. Starting with  $p$ , Grok's first instinct is to simply incorporate the conditional information by adding  $F$  as a parent of  $G$ , and then associating the cpd  $p$  with  $G$ . But then what should she do with the original probability she had for  $G$ ? Should she just discard it? It is easy to check that there is no joint distribution that is consistent with both the two original priors on  $F$  and  $G$  and also  $p$ . So if she is to represent the information with a BN, which always represents a consistent distribution, she must resolve the inconsistency.

However, sorting this out immediately may not be ideal. For instance, if the inconsistency arises from a conflation between two definitions of "gun", a resolution will have destroyed the original cpds. A better use of computation may be to notice the inconsistency and look up the actual law. By way of contrast, consider the corresponding PDG. In a PDG, the cpds are attached to arcs, rather than nodes of the graph. In order to represent unconditional probabilities, we introduce a *unit variable*  $\mathbb{1}$  which takes only one value, denoted  $\star$ . This leads Grok to the PDG depicted in [Figure 3.1](#), where the arcs from  $\mathbb{1}$  to  $F$  and  $G$  are associated with the unconditional probabilities of  $F$  and  $G$ , and the arcs between  $F$  and  $G$  are associated with  $p$  and  $p'$ .

The original state of knowledge consists of all three nodes and the two solid arcs from  $\mathbb{1}$ . This is like Bayes Net that we considered above, except that we no longer explicitly take  $F$  and  $G$  to be independent; we merely record the constraints imposed by the given probabilities.

The key point is that we can incorporate the new information into our original

representation (the graph in [Figure 3.1](#) without the arc from  $F$  to  $G$ ) simply by adding the arc from  $F$  to  $G$  and the associated cpd  $p$  (the new information is shown in blue). Doing so does not change the meaning of the original arcs.<sup>1</sup> Unlike a Bayesian update, the operation is even reversible: all we need to do recover our original belief state is delete the new arc, making it possible to mull over and then reject an observation.  $\triangle$

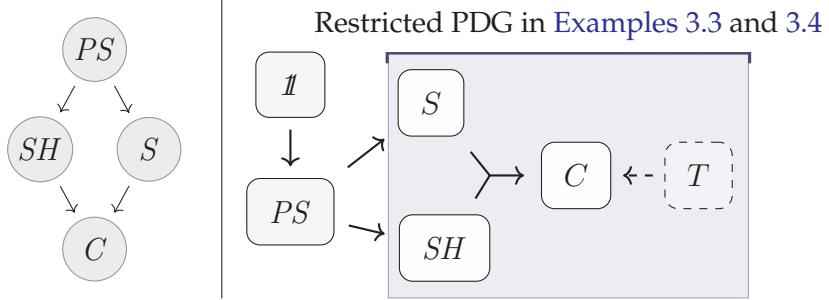
The ability of PDGs to model inconsistency, as illustrated in [Example 3.1](#), appears to have come at a significant cost. We seem to have lost a key benefit of BNs: the ease with which they can capture (conditional) independencies, which, as Pearl ([1988](#)) has argued forcefully, are omnipresent.

**Example 3.2** (emulating a BN). We now consider the classic (quantitative) Bayesian network  $\mathcal{B}$ , which has four binary variables indicating whether a person ( $C$ ) develops cancer, ( $S$ ) smokes, ( $SH$ ) is exposed to second-hand smoke, and ( $PS$ ) has parents who smoke, presented graphically in [Figure 3.2a](#). We now walk through what is required to represent  $\mathcal{B}$  as a PDG, which we call  $m_{\mathcal{B}}$ , shown as the solid nodes and arcs in [Figure 3.2b](#).

We start with the nodes corresponding to the variables in  $\mathcal{B}$ , together with the special node  $\mathbb{1}$  from [Example 3.1](#); we add an arc from  $\mathbb{1}$  to  $PS$ , to which we associate the unconditional probability given by the cpd for  $PS$  in  $\mathcal{B}$ . We can also re-use the cpds for  $S$  and  $SH$ , assigning them, respectively, to the arcs  $PS \rightarrow S$  and  $PS \rightarrow SH$  in  $m_{\mathcal{B}}$ . There are two remaining problems: (1) modeling the remaining table in  $\mathcal{B}$ , which corresponds to the conditional probability of  $C$  given  $S$  and  $SH$ ; and (2) recovering the additional conditional independence

---

<sup>1</sup>While the meaning of the original arcs does not change, one might want something different at a qualitative level, as we will explore in [Section 3.2.2](#).



*Figure 3.2:* (a) The Bayesian Network  $\mathcal{B}$  in [Example 3.2](#) (left), and (b)  $m_{\mathcal{B}}$ , its corresponding PDG (right). The shaded box indicates a restriction of  $m_{\mathcal{B}}$  to only the nodes and arcs it contains, and the dashed node  $T$  and its arrow to  $C$  can be added in the PDG, without taking into account  $S$  and  $SH$ .

assumptions in the BN.

For (1), we cannot just add the arcs  $S \rightarrow C$  and  $SH \rightarrow C$  that are present in  $\mathcal{B}$ . As we saw in [Example 3.1](#), this would mean supplying two *separate* tables, one indicating the probability of  $C$  given  $S$ , and the other indicating the probability of  $C$  given  $SH$ . We would lose significant information that is present in  $\mathcal{B}$  about how  $C$  depends jointly on  $S$  and  $SH$ . To distinguish the joint dependence on  $S$  and  $SH$ , for now, we draw an arc with two tails—a (directed) *hyperarc*—that completes the diagram in [Figure 3.2b](#). With regard to (2), there are many distributions consistent with the conditional marginal probabilities in the cpds, and the independencies presumed by  $\mathcal{B}$  need not hold for them. Rather than trying to distinguish between them with additional constraints, we develop a scoring-function semantics for PDGs which is in this case uniquely minimized by the distribution specified by  $\mathcal{B}$  ([Theorem 3.5](#)). This allows us to recover the semantics of Bayesian networks without requiring the independencies that they assume.

Next suppose that we get information beyond that captured by the original BN. Specifically, we read a thorough empirical study demonstrating that people

who use tanning beds have a 10% incidence of cancer, compared with 1% in the control group (call the cpd for this  $p$ ); we would like to add this information to  $\mathcal{B}$ . The first step is clearly to add a new node labeled  $T$ , for “tanning bed use”. But simply making  $T$  a parent of  $C$  (as clearly seems appropriate, given that the incidence of cancer depends on tanning bed use) requires a substantial expansion of the cpd; in particular, it requires us to make assumptions about the interactions between tanning beds and smoking. The corresponding PDG,  $m_{\mathcal{B}}$ , on the other hand, has no trouble: We can simply add the node  $T$  with an arc to  $C$  that is associated with  $p$ . But note that doing this makes it possible for our knowledge to be inconsistent. To take a simple example, if the distribution on  $C$  given  $S$  and  $H$  encoded in the original cpd was always deterministically “has cancer” for every possible value of  $S$  and  $H$ , but the distribution according to the new cpd from  $T$  was deterministically “no cancer”, the resulting PDG would be inconsistent.  $\triangle$

We have seen that we can easily add information to PDGs; removing information is equally painless.

**Example 3.3** (restriction). After the Communist party came to power, children were raised communally, and so parents’ smoking habits no longer had any impact on them. Grok is reading her favorite book on graphical models, and she realizes that while the node  $PS$  in Figure 3.2a has lost its usefulness, and nodes  $S$  and  $SH$  no longer ought to have  $PS$  as a parent, the other half of the diagram—that is, the node  $C$  and its dependence on  $S$  and  $SH$ —should apply as before. Grok has identified two obstacles to modeling deletion of information from a BN by simply deleting nodes and their associated cpds. First, this restricted model is technically no longer a BN (which in this case would require unconditional

distributions on  $S$  and  $SH$ ), but rather a *conditional* BN (Koller and Friedman 2009), which allows for these nodes to be marked as observations; observation nodes do not have associated beliefs. Second, even regarded as a conditional BN, the result of deleting a node may introduce *new* independence information, incompatible with the original BN. For instance, by deleting the node  $B$  in a chain  $A \rightarrow B \rightarrow C$ , one concludes that  $A$  and  $C$  are independent, a conclusion incompatible with the original BN containing all three nodes. PDGs do not suffer from either problem. We can easily delete the nodes labeled 1 and  $PS$  in Figure 3.2b to get the restricted PDG shown in the figure, which captures Grok's updated information. The resulting PDG has no arcs leading to  $S$  or  $SH$ , and hence no distributions specified on them; no special modeling distinction between observation nodes and other nodes are required. Because PDGs do not directly make independence assumptions, the information in this fragment is truly a subset of the information in the whole PDG.  $\triangle$

The ability to form restrict to arbitrary subsets of information is useful, and closely related to an even more compelling reason to use PDGs: they make it equally easy to aggregate arbitrary probabilistic information.

**Example 3.4.** Grok dreams of becoming Supreme Leader ( $SL$ ), and has come up with a plan. She has noticed that people who use tanning beds have significantly more power than those who don't. Unfortunately, her mom has always told her that tanning beds cause cancer; specifically, that 15% of people who use tanning beds get it, compared to the baseline of 2%. Call this cpd  $q$ . Grok thinks people will make fun of her if she uses a tanning bed and gets cancer, making becoming Supreme Leader impossible. This mental state is depicted as a PDG on the left of Figure 3.3.

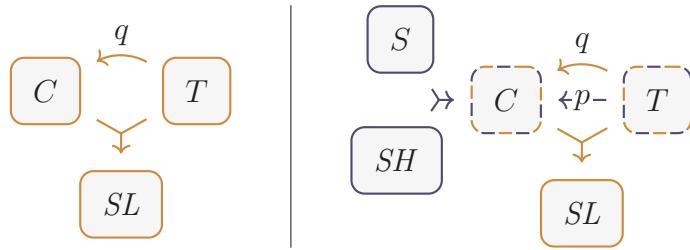


Figure 3.3: Grok's prior (left) and combined (right) knowledge.

Grok is reading about graphical models because she vaguely remembers that the variables in [Example 3.2](#) match the ones she already knows about. When she finishes reading the statistics on smoking and the original study on tanning beds (associated to a cpd  $p$  in [Example 3.2](#)), but before she has time to reflect, we can represent her (conflicted) knowledge state as the sum of the two graphs, depicted graphically on the right of [Figure 3.3](#).

The union of the two PDGs, even with overlapping nodes, is still a PDG. This is not the case in general for BNs. Note that the PDG that Grok used to represent her two different **sources** of information (the mother's wisdom and the study) regarding the distribution of  $C$  is a *multigraph*: there are two arcs from  $T$  to  $C$ , with inconsistent information. Had we not allowed multigraphs, we would have needed to choose between the two arcs, or represent the information some other (arguably less natural) way. As we are already allowing inconsistency, merely recording both is much more in keeping with the way we have handled other types of uncertainty.  $\triangle$

Not all inconsistencies are equally egregious. For example, even though the cpds  $p$  and  $q$  are different, they are numerically close, so, intuitively, the PDG on the right in [Figure 3.3](#) is not very inconsistent. Making this precise is the focus of [Section 3.3.2](#).

These examples give a taste of the power of PDGs. In the coming sections, we formalize PDGs and relate them to other approaches.

### 3.2 Syntax

We now provide our first formal definition of a PDG.

**Definition 3.1.** A *Probabilistic Dependency Graph* is a tuple  $\mathcal{M} = (\mathcal{N}, \mathcal{A}, \mathcal{V}, \mathbb{P}, \alpha, \beta)$ , where

$\mathcal{N}$  is a finite set of nodes, corresponding to the identities of variables;

$\mathcal{A}$  is a collection of (hyper)arcs each  $a \in \mathcal{A}$  of which has source(s)  $S$  and target(s)  
 $T$  in  $\mathcal{N}$ ;

$\mathcal{V}$  associates each node  $N \in \mathcal{N}$  with a set  $\mathcal{V}(N)$  of possible values, allowing us to view it as a variable;

$\mathbb{P}$  associates to each arc  $X \xrightarrow{a} Y \in \mathcal{A}$  a cpd  $\mathbb{P}_a(Y|X)$ ;

$\alpha$  associates to each arc  $X \xrightarrow{a} Y$  a number  $\alpha_a$  which, roughly speaking, is the modeler's confidence in the functional dependence of  $Y$  on  $X$  implicit in  $a$ ;

$\beta$  associates to each arc  $a \in \mathcal{A}$  a (possibly infinite) real number  $\beta_a$ , the modeler's subjective confidence in the reliability of  $\mathbb{P}$ .

The definition is ambiguous about whether  $(\mathcal{N}, \mathcal{A})$  is a directed graph, multi-graph, or hypergraph. This is because, perhaps counter-intuitively, all of these variants will turn out to be equivalent. Since  $\mathcal{A}$  can be extended with identity arcs  $\{X \xrightarrow{id} X\}_{X \in \mathcal{N}}$ , we make our usual assumption that  $\mathcal{N}$  is implicit in  $\mathcal{A}$ . Furthermore, in this case, both  $\mathcal{A}$  and  $\mathcal{V}$  are then implicit in  $\mathbb{P} = \{\mathbb{P}_a : \mathcal{V}S_a \rightarrow$

$\Delta\mathcal{V}T_a\}_{a \in \mathcal{A}}$ . Thus, we refer to  $\mathbb{P}$  as an *unweighted PDG*; we give it semantics as though it were the weighted PDG that has  $\beta = \alpha = 1$ .  $\square$

If  $\mathbf{m}$  is a PDG, we reserve the names  $\mathcal{N}^{\mathbf{m}}, \mathcal{A}^{\mathbf{m}}, \dots$ , for the components of  $\mathbf{m}$ , so that we may reference one without naming them all explicitly. The pair  $(\mathcal{N}, \mathcal{V})$  describes a set of nodes  $\mathcal{N}$  and a set of values  $\mathcal{V}X$  for each node  $X \in \mathcal{N}$ ; thus  $(\mathcal{N}, \mathcal{V})$  is equivalent to specifying a set  $\mathcal{X}$  of variables. For this reason, we write  $\mathcal{V}\mathcal{X} := \prod_{X \in \mathcal{N}} \mathcal{V}X$  for the set of all joint settings of the variables (which will be our outcome space  $\Omega$ ); we refer to  $\omega \in \mathcal{V}\mathcal{X}$  as “worlds”. Rather than specifying  $\mathcal{N}$  and  $\mathcal{V}$  separately, it is common practice in the graphical models community to instead specify a set  $\mathcal{X}$  of variables directly (Koller and Friedman 2009). We have kept them them separate in Definition 3.1 to emphasize the separation between the qualitative information  $(\mathcal{N}, \mathcal{A}, \alpha)$  and the quantitative information  $(\mathcal{V}, \mathbb{P}, \beta)$  that annotates it.

We now present some shorthand to clarify the presentation. We typically conflate a cpd’s symbol with its edge label, thus drawing the PDG with a single edge attached to  $f(Y|X)$  as  $[X] \xrightarrow{f} [Y]$ . Regardless of the underlying definition of a PDG, we will use hypergraph notation, which is typically more compact and clearer. Thus, we will indicate joint dependence with multi-tailed arcs, joint distributions with multi-headed arcs, and unconditional distributions with nothing at the tail. For instance, we draw

$$p(Y|X, Z) \text{ as } \begin{array}{c} Z \\ \nearrow p \\ X \end{array} \rightarrow [Y], \text{ and } q(A, B) \text{ as } \begin{array}{c} q \\ \swarrow \nearrow \\ A \quad B \end{array}.$$

To emphasize that a cpd  $f(Y|X)$  is deterministic (i.e., a function  $f : X \rightarrow Y$ ), we will draw its arc with two heads, as in:  $[X] \xrightarrow{f} [Y]$ . We identify an event  $X=x$  with the degenerate unconditional distribution  $\delta_x(X)$  that places all mass

on  $x$ ; hence it may be associated to an edge and drawn simply as  $\xrightarrow{x} [X]$ . To specify a confidence  $\beta \neq 1$ , we place the value near the edge, lightly colored and parenthesized, as in:  $\xrightarrow{(\beta)} [X]$ , and we write  $(\infty)$  for the limit of high confidence ( $\beta \rightarrow \infty$ ).

### 3.2.1 Alternate Equivalent Definitions of PDGs

[Definition 3.1](#) is not the only way to define a PDG. Indeed, there are many definitions of PDGs, and seeming generalizations or restrictions of them, all of which turn out to be essentially equivalent. We now give an overview of important alternative variants.

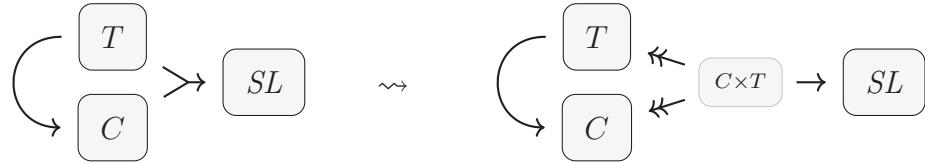
**PDG Variant 1: Segregated PDGs.** According to [Definition 3.1](#), each (hyper)arc  $X \xrightarrow{a} Y \in \mathcal{A}$  of PDG plays two distinct roles: qualitatively, it says something about causality and functional dependence through the parameter  $\alpha_a$ ; quantitatively, it specifies a cpd  $\mathbb{P}_a$  and a confidence  $\beta_a$  in it. Conceptually, however, these two aspects of an arc are rather separate. Indeed, a PDG can be equivalently defined with two distinct kinds of arcs, each of which only play one role. An arc that plays both roles can be separated into two separate arcs, one for each kind of information. Conversely, an arc with only one kind of information can be specified by spacing  $\alpha = 0$  or  $\beta = 0$  to nullify the other kind of information, as appropriate. We have opted for combined edges because certain semantic and inference properties of PDGs are easier to describe in this presentation (e.g., [Theorems 3.3, 3.7](#) and [9.1](#)).

**PDG Variant 2: Probabilistic Dependency Multigraphs.** As mentioned previously, the structure  $(\mathcal{N}, \mathcal{A})$  of a PDG can be taken to directly be an (ordi-

nary) directed multigraph rather than the (seemingly more expressive) directed *hypergraph*. This approach is an instance of the approach to probabilistic modeling that makes variables primitive and adds constraints, as first described in

[Section 2.3.1](#). Elaborate slightly. Then say: Formally, that means using the following construction to compile a hypergraph to a graph.

**Construction 3.2.** We can capture multi-tailed arcs and joint dependence, even with ordinary graphs, at the cost of an extra node and a few extra arcs. For instance, the diagram displaying Grok's prior knowledge in [Example 3.4](#), on the left of [Figure 3.3](#) can be viewed shorthand for the following PDG, where we insert a node labeled  $C \times T$  at the junction:



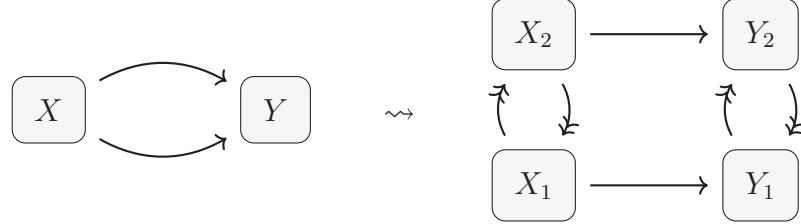
As the notation suggests,  $\mathcal{V}(C \times T) = \mathcal{V}(C) \times \mathcal{V}(T)$ . For all  $(c, t) \in \mathcal{V}(C \times T)$  the cpd for the arc from  $C \times T$  to  $C$  gives probability 1 to  $c$ ; similarly, the cpd for the arc from  $C \times T$  to  $T$  gives probability 1 to  $t$ .

More generally, a hyperarc  $S \rightarrow T$ , where  $S, T \subseteq \mathcal{N}$ , can be compiled to a fragment of an ordinary graph, containing:

- (a) two additional “super-nodes” that are not elements of  $\mathcal{N}$ , corresponding to  $S$  and  $T$  respectively, whose possible values are joint settings of their constituent variables;
- (b) projections  $S \rightarrow X$  for each  $X \in S$ , and projections  $T \rightarrow Y$  for each  $Y \in T$ , all with high confidence ( $\alpha = \beta = \infty$ );
- (c) an arc  $S \rightarrow T$ , associated with the original cpd and confidences.

The special case of a hyperarc with no tails leads to the special variable  $\mathbb{1}$  such that  $\mathcal{V}(\mathbb{1}) = \{\star\}$ , as introduced in [Examples 3.1](#) and [3.2](#).  $\square$

**PDG Variant 3: Strict PDGs.** Going even further, the restriction to PDGs in which  $(\mathcal{N}, \mathcal{A})$  is not even a multi-graph, but rather an ordinary directed graph, is not a restriction at all. It is not hard to see that a multi-graph could be simulated in that formalism with the following widget:



where the vertical arcs enforce equality.

**PDG Variant 4: Lax PDGs.** The confidence  $\beta_a$  in the arc  $X \xrightarrow{a} Y$  may be replaced with a separate confidence  $\beta_{a,x}$  for each possible input  $x \in \mathcal{V}(X)$ . More generally, one might imagine a variant in which one needn't specify a complete conditional probability table, but could leave some entries blank. We show in [Section 4.2.4](#) that these two seeming generalizations of a PDG are equivalent to the definition we already have.

**PDG Variant 5: Parametric PDGs.** One might want to also consider a *parametric* variant of a PDG, in which each cpd is not fixed, but dependent on some hyperparameters—like a neural network. This approach, which we take in [Chapter 7](#), may be especially appropriate if the cpds represent parametric models. Like the others, this parametric variant appears to be a more general object—yet, once again, it turns out to be equivalent to all of the other variants of PDGs we have considered so far.

### 3.2.2 Combining PDGs

How should we combine independent sources of information? As illustrated by [Example 3.4](#), a significant advantage of PDGs is their modularity: there is a

simple way to combine the information in  $m_1$  and  $m_2$  independently: take the union of their variables and the disjoint union of their arcs (and associated data) to get a new PDG, denoted  $m_1 + m_2$ .

**An Important Caveat For Qualitative (Structural) Information.** For the quantitative data (i.e., the parameters  $\mathbb{P}$  and  $\beta$ ), this is precisely how one would want to combine independent observations. Independently combining the qualitative information, however, may not quite be what is intended if there is any overlap. To illustrate, suppose that  $m_1$  and  $m_2$  both contain a single arc from  $X$  to  $Y$ , each specifying a single causal mechanism by which  $X$  gives rise to  $Y$ . The combined PDG  $m_1 + m_2$  contains two *distinct* mechanisms by which  $X$  gives rise to  $Y$  (see [Chapter 5](#)). Rather than adopting two distinct mechanisms, one might well have intended to merge them, viewing the two arcs as independent accounts that there is a *single* such causal mechanism. Yet PDG union performs no such merge. Indeed, the “appropriate thing” to do seems to depend on context, and so we urge the modeler to think about which (if any) qualitative mechanisms ought to be merged together, after this operation. We will return to this point in [Example 3.6](#), and again in [Section 6.7](#). In that context, we will also see that the analogue of this problem is even bigger for factor graphs, because  $\alpha$  and  $\beta$  are essentially “fused together” in these models.

It is worth mentioning that *convex combinations* of qualitative parameters  $\alpha$  actually *do* work as one might hope. If you believe that there’s a 50% chance that the causal graph is  $G_1$ , and a 50% chance that the causal graph is  $G_2$ , there is no problem with taking  $\alpha$  to be  $.5\alpha_1 + .5\alpha_2$ , where  $\alpha_1$  and  $\alpha_2$  represent the causal graphs  $G_1$  and  $G_2$ , respectively. For instance, if  $G_1$  and  $G_2$  are distinct causal graphs that represent the same independencies, then  $\alpha$  will encode those shared

independencies.

Still, despite the caution, there are also some cases where independent combination is appropriate even for causal information. We now proceed with the formal details.

**Formal Details of PDG Sums.** To be completely precise, it's helpful to first make a preliminary definition:  $m_1$  and  $m_2$  are  $\mathcal{V}$ -compatible iff the variables they have in common take the same sets of values—that is, if the nodes and values of  $m_1$  and  $m_2$ , are  $\mathcal{N}_1, \mathcal{V}_1$  and  $\mathcal{N}_2, \mathcal{V}_2$ , respectively, then  $\forall X \in \mathcal{N}_1 \cap \mathcal{N}_2. \mathcal{V}_1(X) = \mathcal{V}_2(X)$ . We can only combine PDGs that are  $\mathcal{V}$ -compatible—but this is not a meaningful restriction. If in place of the pair  $(\mathcal{N}, \mathcal{V})$ , PDGs are specified directly using a set  $\mathcal{X}$  of variables, then this definition is unnecessary, as its sole purpose is to ensure that we do not accidentally identify variables in different PDGs that do not take on the same values. In any case, the point is the following formal definition for how to combine independent information:

**Definition 3.3.** The sum of  $\mathcal{V}$ -compatible PDGs is given by taking the disjoint union of their arcs and associated data. Explicitly, if  $m_1 = (\mathcal{N}_1, \mathcal{V}_1, \mathcal{A}_1, \mathbb{P}_1, \boldsymbol{\alpha}_1, \boldsymbol{\beta}_1)$  and  $m_2 = (\mathcal{N}_2, \mathcal{V}_2, \mathcal{A}_2, \mathbb{P}_2, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_2)$  are  $\mathcal{V}$ -compatible PDGs, then

$$m_1 + m_2 := \left( \mathcal{N}_1 \cup \mathcal{N}_2, \mathcal{V}_1 \cup \mathcal{V}_2, \mathcal{A}_1 \sqcup \mathcal{A}_2, \mathbb{P}_1 \sqcup \mathbb{P}_2, \boldsymbol{\alpha}_1 \sqcup \boldsymbol{\alpha}_2, \boldsymbol{\beta}_1 \sqcup \boldsymbol{\beta}_2, \right),$$

where, for  $f, g : X \rightarrow Y$ , the function  $f \sqcup g : X \sqcup X \rightarrow Y$  is given by

$$f \sqcup g := \begin{cases} \text{inl}(x) \mapsto f(x) \\ \text{inr}(x) \mapsto g(x) \end{cases}. \quad \square$$

### 3.3 The Semantics of PDGs

obs

add section references!

Although the meaning of an individual cpd is clear, we have not yet given PDGs a “global” semantics. We discuss three related approaches to doing so. The first is the simplest: we associate with a PDG the set of distributions that are consistent with it. This set will be empty if the PDG is inconsistent. The second approach associates a PDG with a scoring function, indicating the fit of an arbitrary distribution  $\mu$ , and can be thought of as a *weighted* set of distributions (Halpern and Leung 2015). This approach allows us to distinguish inconsistent PDGs, while the first approach does not. The third approach chooses the distributions with the best score, typically associating with a PDG a unique distribution.

#### 3.3.1 PDGs As Sets Of Distributions

We have been thinking of a PDG as a collection of constraints on distributions, specified by matching cpds. From this perspective, it is natural to consider the set of all distributions that are consistent with the constraints. with  $\mathbb{P}$

**Definition 3.4.** If  $m$  is a PDG (weighted or unweighted) with arcs  $\mathcal{A}$  and cpds  $\mathbb{P}$ , let  $\{m\}$  be the set of distributions over the variables in  $m$  whose conditional marginals are exactly those given by  $\mathbb{P}$ . That is,  $\mu \in \{m\}$  iff, for all arcs  $a \in \mathcal{A}$  from  $X$  to  $Y$ ,  $x \in \mathcal{V}(X)$ , and  $y \in \mathcal{V}(Y)$ , we have that  $\mu(Y=y | X=x) = \mathbb{P}_a(Y=y | X=x)$ . Formally, define

$$\{m\} := \{\mu \in \Delta^{\mathcal{V}\mathcal{X}} \mid \forall X \xrightarrow{a} Y \in \mathcal{A}. \mu(Y|X) = \mathbb{P}_a(Y|X)\}.$$

We say that  $m$  is *inconsistent* if  $\{m\} = \emptyset$ , and *consistent* otherwise.  $\square$

Note that  $\{\mathcal{m}\}$  does not depend on the weights  $\alpha$  or  $\beta$ . Furthermore, it reflects only the quantitative probabilistic information in the PDG; it does not say anything about independencies, for example. Developing an analogous definition for a *qualitative* information in a PDG is far more subtle, and is the subject of Chapter 5.

This semantics is useful, and we will make frequent use of it, but it has one other extremely important shortcoming: viewed through the lens of this semantics, all inconsistent PDGs are equivalent, no matter how small the inconsistency. In the next subsection, we develop another semantics, which, among other things, gives us a modification of this semantics that does not have this problem.

### 3.3.2 PDGs As Scoring Functions over Joint Distributions

We now present the full PDG semantics. Specifically, we will interpret a PDG  $\mathcal{m}$  not as a (convex) set of distributions, but rather as a (convex) *scoring function* on distributions. Given  $\mu \in \Delta\mathcal{VX}$ , the scoring function for  $\mathcal{m}$  returns a real-valued score, indicating how well (or, rather, how poorly)  $\mu$  matches the information in  $\mathcal{m}$ . Distributions with the lowest (best) scores are those that most closely match the cpds in  $\mathcal{m}$ , and contain the fewest unspecified correlations.

These two aspects of the scoring function correspond to the two kinds of information present in a PDG: “structural” information, in the (hyper)graph  $\mathcal{A}$  and weights  $\alpha$ , and “observational” data, in the cpds  $\mathbb{P}$  and weights  $\beta$ . PDG semantics are based on two scoring functions that quantify discrepancy between each type of information and a distribution  $\mu \in \Delta\mathcal{VX}$  over its variables.

**Quantitative/Observational Scoring.** We need a function that measures “distance” between  $\mu$  and the cpds of  $\mathcal{M}$ . The key desideratum for such a function is that it must assign higher scores to distributions  $\mu$  when  $\mathcal{M}$  requires larger changes to be consistent with  $\mu$ . In principle, there are many functions with this property. Yet, because it is a discrepancy between beliefs ( $\mathcal{M}$ ) and reality ( $\mu$ ) there is an especially natural way of measuring its magnitude: relative entropy.

Recall (from [Section 2.6.2](#)) that relative entropy  $D(\mu \parallel p) = \mathbb{E}_\mu[\log \frac{\mu}{p}]$  measures divergence of a belief  $p$  with respect to reality  $\mu$ . It can be viewed, for example, as the overhead (in extra bits per sample) of using representations optimized for  $p$ , when in fact samples are distributed according to  $\mu$  ([MacKay 2003](#)). This suggests scoring distributions with a weighted sum of relative entropies of the appropriate types.

**Definition 3.5.** The *observational incompatibility* of  $\mu(\mathcal{X})$  with  $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathbb{P}, \boldsymbol{\alpha}, \boldsymbol{\beta})$  is given by the weighted sum of relative entropies:

$$OInc_m(\mu) := \sum_{S \xrightarrow{a} T \in \mathcal{A}} \beta_a D\left(\mu(T, S) \parallel \mathbb{P}_a(T|S) \mu(S)\right). \quad \square \quad (3.1)$$

As we have argued, not all inconsistencies are equally problematic. Discrepancies with low confidence cpds are not as problematic as discrepancies with trusted, high-confidence ones. The confidences  $\boldsymbol{\beta}$  operationalize this relative importance. So overall,  $OInc_m$  measures the excess cost of using codes optimized for the cpds of  $\mathcal{M}$  (weighted by their confidences  $\boldsymbol{\beta}$ ), when reality is distributed according to  $\mu$ . =P

In (3.1), we have one “belief” distribution per hyperarc  $X \xrightarrow{a} Y \in \mathcal{A}$ , namely, the one induced by drawing  $x \sim \mu(X)$  and applying the cpd  $\mathbb{P}_a(Y|X=x)$ . At the same time, it also makes sense to model things at a finer level. A case could also

be made that  $\mathbb{P}_a$  describes not one context-dependent belief, but rather a separate belief  $\mathbb{P}_a(Y|x)$  each  $x \in \mathcal{V}X$ . This suggests a different approach: for each  $X \xrightarrow{a} Y$  and  $x \in \mathcal{V}X$ , start by measuring the relative entropy from  $\mathbb{P}_a(Y|x)$  to  $\mu(Y|X=x)$ . We still want to combine these into a single number, and not all of these beliefs have equal importance—those that correspond to the likeliest values of  $X$  should have the most weight—which suggests taking an expectation of these quantities with respect to the marginal distribution  $\mu(X)$ . After taking a weighted sum over arcs as before, we would get the quantity

$$OInc_m(\mu) = \sum_{X \xrightarrow{a} Y \in \mathcal{A}} \beta_a \cdot \mathbb{E}_{x \sim \mu(X)} \left[ D\left(\mu(Y | X=x) \parallel \mathbb{P}_a(Y | x)\right) \right]. \quad (3.2)$$

It is not hard to show that our two formulas for  $OInc$ , (3.1) and (3.2), are in fact equal—reflecting a(nother) virtue of relative entropy. (It would not have been the case, for instance, had we taken  $D(q \parallel p)$  to be the Euclidean distance between  $q$  and  $p$  as vectors.) Over the course of this thesis, we will discover a great many more. Still, any measure of discrepancy  $D$  would have sufficed to get an analogue of the next result, which makes precise that the (observational) scoring function semantics of a PDG generalizes the set-of-distributions semantics from the previous section (3.3.1).

**Proposition 3.1.** *If  $m$  is a PDG with  $\beta > 0$ , then  $\mu \in \{m\}$  iff  $OInc_m(\mu) = 0$ .*

[ link to proof ]

It is worth pausing here to reflect on the meaning of  $\beta$ , which is not a collection of probabilities, but of confidences in a different sense (which we develop at full generality in Chapter 11). A choice of  $\beta_a = 0$  means that the cpd  $\mathbb{P}_a$  is effectively ignored, in the sense that such a PDG is equivalent to one in which  $a$  is attached to a different cpd  $q \neq \mathbb{P}_a$  (or missing altogether, modulo the qualitative effect of removing the arc  $a$ ). This is why Proposition 3.1 requires  $\beta$  to be nonzero.

At the the upper extreme, a large (or even infinite) value of  $\beta_a$  indicates high (or absolute) confidence in  $\mathbb{P}_a$ . By default, we assume an intermediate value of confidence  $\beta = 1$ , which is just a convenient choice of units—what's important are the magnitudes of  $\beta$  relative to one another.

Here's an concrete way to think about it. In your head, fix a reference person—someone who is worth listening to a little bit, but far from infallible—and call that degree of trust “ $\beta = 1$ ”. Let's call this person Bob. To determine your degree of confidence in information ask yourself: how many times would I have to hear this independently from **a person like Bob, before I got to this level of trust?**  
**plural**

The answer to that question is the value of  $\beta$ . Thus,  $\beta = 2$  on an arc  $X \rightarrow Y$  is (quantitatively) indistinguishable from two distinct arcs from  $X \rightarrow Y$ .

**Qualitative/Structural Scoring.** So far, we have said nothing about the semantics of the qualitative information in a PDG. For traditional graphical models—that is, Bayesian Networks (BNs) and Markov Random Fields (MRFs)—the qualitative information is a conjunction of independencies implied by the graphical structure. But what are the independencies implied by a directed hypergraph? Or even an ordinary directed graph with cycles? Despite decades of interest in this simple question, no clear consensus has emerged. As we shall see in [Chapter 5](#), there is a good reason why: qualitative information in an arbitrary directed (hyper)graph is not just about independence; in general, it can be far more subtle. Indeed, for the qualitative information, we have found it easier to motivate the scoring function directly.

Intuitively, each arc  $X \xrightarrow{a} Y$  represents an independent randomized mechanism by which  $X$  leads to  $Y$ . Or, perhaps slightly more precisely, that  $Y$  can be com-

puted from  $X$  alone (and independent random noise). So, given a (hyper)graph  $\mathcal{A}$  whose nodes correspond to variables, and distribution  $\mu(\mathcal{X})$  over (the values of) those variables, contrast the number of bits of randomness needed to do each of the following (in expectation):

- (a) directly specify the values of all variables  $\mathcal{X}$ , and
- (b) separately specify, for each arc  $X \xrightarrow{a} Y$ , the value of  $Y$  given the value of  $X$ .

As one might have expected, we measure the amount of randomness using entropy. In a sense, (b) is the total amount of randomness we need in order to generate components of an outcome of  $\mu$  along the structure of the hypergraph  $\mathcal{A}$ , while (a) is the amount of randomness needed to generate an outcome of  $\mu = \mu(\mathcal{X}|\emptyset)$  along its own simple dependency structure ( $\emptyset \rightarrow \mathcal{X}$ ). Clearly, the two quantities must be equal in order for  $\mathcal{A}$  to be a complete and accurate causal explanation for the origin of  $\mu$ .

#### the distribution

What if they differ? If  $(b) > (a)$ , then there are correlations in  $\mu$  that allow for a more compact representation than  $\mathcal{A}$  represents—that is,  $\mu$  is *information deficient* for the structure  $\mathcal{A}$ . The larger the difference, the larger the overhead of using the structure  $\mathcal{A}$  to compute probabilities, compared to what is necessary, and so the poorer the structural fit of  $\mu$  to  $\mathcal{A}$ . Notice how this parallels the overhead of erroneous quantitative beliefs discussed above. If  $(a) > (b)$ , on the other hand (e.g., if  $\mathcal{A}$  is empty), then  $\mathcal{A}$  must be an incomplete qualitative picture—because to explain everything in  $\mu$ , at least  $H(\mu)$  bits are required. Although an incomplete picture may not be a good thing, it fits data better than a complete one. So, in both cases, the fit is worst when (b) is large and (a) is small. This observation motivates our qualitative scoring function.

**Definition 3.6.** The *structural (information) deficiency* of distribution  $\mu(\mathcal{X})$ , with respect to a directed (hyper)graph  $(\mathcal{X}, \mathcal{A})$  is given by:

say: weighted

$$SDef_{\mathcal{A}}(\mu) := \sum_{X \xrightarrow{a} Y \in \mathcal{A}} \alpha_a H_{\mu}(Y | X) - H_{\mu}(\mathcal{X}). \quad (3.3)$$

For a PDG  $m$ , we take  $SDef_m := SDef_{\mathcal{A}^m}$  to be shorthand for the information deficiency with respect to the structure of  $m$ .  $\square$

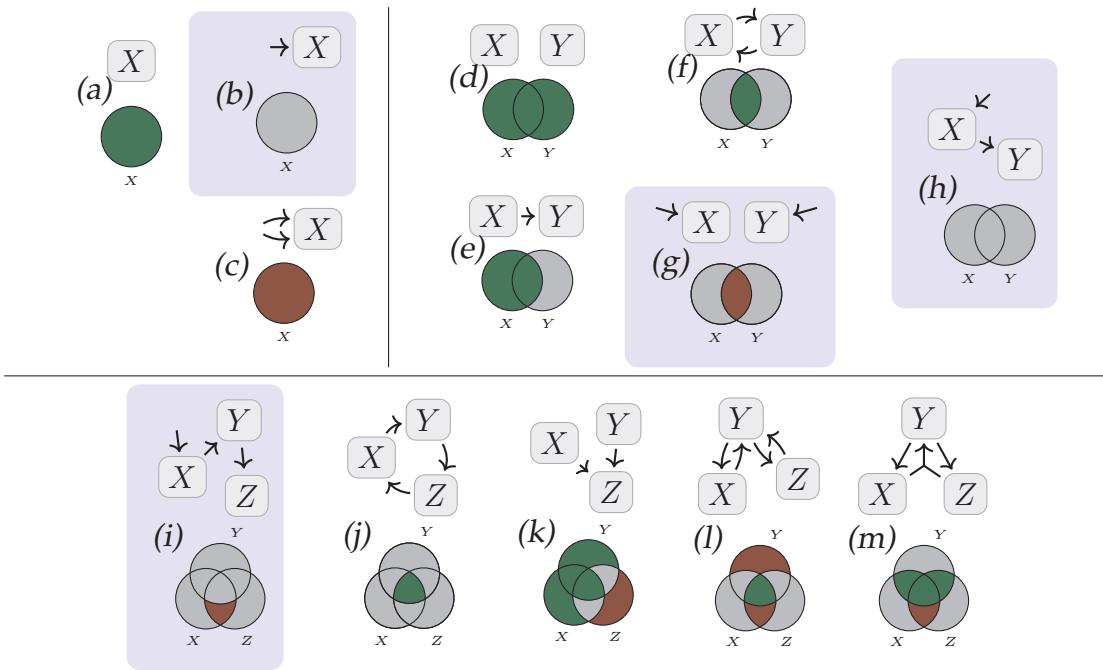
Talk about weights a bit here, or otherwise link to where I do

We illustrate  $SDef$  with some simple examples. First, when  $\mathcal{A}$  has no arcs, then  $SDef_{\mathcal{A}}(\mu) = -H(\mu)$ . Since smaller numbers represent a better fit,  $SDef_m$  tells us to maximize entropy—which is precisely what the literature tells us to do, at least, to break ties between distributions consistent with our observations in the absence of no structural or causal information (Jaynes 1957). Next, suppose  $\mathcal{A} = [X] \rightarrow [Y]$ . In general, specifying  $X$  and  $Y$  together takes the same amount of information as first specifying  $X$ , and then using that value to specify  $Y$  (recall the chain rule, from Section 2.6.1). So  $\mathcal{A}$  is incomplete, and missing precisely an explanation for the origin of  $X$ . By the same reasoning as before, the best bet is to maximize entropy (but now just of  $X$ ); indeed,  $SDef_{\mathcal{A}} = -H_{\mu}(X)$ . If  $\mathcal{A}$  has sufficiently many parallel arcs from  $X$  to  $Y$  and  $H_{\mu}(Y | X) > 0$  (so that  $Y$  is not totally determined by  $X$ ) then we have  $SDef_{\mathcal{A}}(\mu) > 0$ , because the redundant arcs add no information, but there is still a cost to specifying them. In this case,  $SDef_{\mathcal{A}}$  prefers distributions that make  $Y$  a deterministic function of  $X$  (while still maximizing the entropy of  $X$ ). Finally, if  $\mathcal{A} = [X] \rightsquigarrow [Y]$ , then  $SDef_{\mathcal{A}}(\mu) = -I_{\mu}(X; Y)$  encourages mutual information between  $X$  and  $Y$  (and so it is optimal for  $X$  and  $Y$  to be functions of each other).

add ref to maxent section!

add link to Ch5

Since  $SDef_{\mathcal{A}}$  is a linear combination of (conditional) entropies, it is also a linear function of  $\mu$ 's information profile vector. This means  $SDef_{\mathcal{A}}$  is a dual vector in information profile space, and thus can be illustrated with the same diagram



*Figure 3.4:* Illustrations of the structural deficiency  $SDef_{\mathcal{A}}$  with their associated hypergraphs  $\mathcal{A}$ . Each subfigure depicts a hypergraph  $\mathcal{A}$  (above), and a dual vector  $v_{\mathcal{A}}$  to the information profile (below), which represents  $SDef_{\mathcal{A}}$  in the sense that  $SDef_{\mathcal{A}}(\mu) = v_{\mathcal{A}} \cdot I_{\mu}$ . Each circle represents the information in a variable, just as in [Figure 2.2](#). Green stands for  $-1$  (indicating that information in this component makes for a better fit to  $\mathcal{A}$ ), while red stands for  $+1$  (indicating that information in this component makes for a worse fit to  $\mathcal{A}$ ), and gray stands for  $0$  (which is neutral).

as used to show information profiles. A number of small examples have been visualized this way in [Figure 3.4](#). We now make a few observations.

- In the definition of  $SDef$ , each arc  $X \rightarrow Y$  is compiled to a “cost”  $H(Y|X)$  for uncertainty in  $Y$  given  $X$ . One can see this visually in [Figure 3.4](#) in the form of a red crescent that’s added to the information profile as we move from [3.4d](#) to [3.4e](#) to [3.4f](#). In the unconditional case, Subfigures [3.4a](#), [3.4b](#), and [3.4c](#) show how adding arcs increases the incentive for determinism.
- Some hypergraphs (see [Figures 3.4b](#) and [3.4h](#)) are *indiscriminate*, in the sense that every distribution gets the same score. Specifically, a score of zero, because a point mass  $\delta$  always has  $SDef_{\mathcal{A}}(\delta) = 0$ : for such a distribution, no

randomness is required to generate a joint sample over all variables, and nor is any randomness required to generate samples along  $\mathcal{A}$ .

- As we will soon see,  $SDef$  can indicate independencies and conditional independencies, illustrated respectively in Subfigures 3.4g and 3.4i.
- For more complex structures, the structural information deficiency  $SDef$  can represent more than independence and dependence. The cyclic structures in Examples 5.3 and 5.4, correspond to the structural deficiencies pictured in Subfigures 3.4f and 3.4j, respectively, which are functions that encourage shared information between the three variables. add link to Chapter 5
- The structures in blue boxes correspond to Bayesian Network (BN) structures, and their corresponding  $SDef$ 's quantify discrepancy with a set of (conditional) independencies. Conversely, only those structural deficiencies can be expressed with a BN.

**The Combined Scoring Function.**  $OInc_m(\mu)$  and  $SDef_m(\mu)$  give us two measures of compatibility between  $m$  and a distribution  $\mu$ . The right way to combine them into a single score, as we must to handle cases in which the observational structural information conflict with one another, depends on the importance of structure relative to observation. This could be captured by a trade-off parameter  $\hat{\gamma} \in [0, 1]$  that controls the convex combination  $(1 - \hat{\gamma}) \cdot OInc + \hat{\gamma} \cdot SDef$ . However, for several reasons (discussed below), our primary definition will instead be a rescaled variant with a different parameterization. Using  $\gamma := \hat{\gamma}/(1 - \hat{\gamma}) \in [0, \infty]$ ,

define the overall scoring function:

$$\begin{aligned}
\llbracket m \rrbracket_\gamma(\mu) &:= OInc_m(\mu) + \gamma SDef_m(\mu) \\
&= \frac{1}{1 - \hat{\gamma}} \left( (1 - \hat{\gamma}) OInc_m(\mu) + \hat{\gamma} SDef_m(\mu) \right) \\
&= \mathbb{E}_\mu \left[ \sum_{S \xrightarrow{a} T \in \mathcal{A}} \log \frac{\mu(T|S)^{\beta_a} - \gamma \alpha_a}{\mathbb{P}_a(T|S)^{\beta_a}} \right] - \gamma H(\mu).
\end{aligned} \tag{3.4}$$

While  $\gamma$  and  $\hat{\gamma}$  do indeed parameterize a trade-off between observational and structural information, the two quantities are not really symmetric. For instance, the structural information such as independencies could in principle be captured quantitatively (e.g., with a distribution happens to have those independencies), but not vice versa. At a lower level,  $OInc$  is unbounded and non-negative, while  $SDef$  is bounded and can be negative. Because of the inherent asymmetry, there is no inherent reason to prefer a symmetrical formalism.

In addition, there are benefits to using  $\gamma$  and the scoring function as we have defined it. The first is to draw an analogy with thermodynamics, that is common in some sub-communities that study graphical models. A basic principle of thermodynamics is that macro-states (i.e., distributions) with higher entropy are more stable at higher temperatures, while macro-states with lower entropy are more stable at lower temperatures. In the case of no qualitative information (i.e.,  $\alpha = 0$ ), the scoring function (3.4) is a direct analogue of the Gibbs free energy  $G = U - T H$ , where  $T = \gamma$  is temperature and  $H$  is entropy. Furthermore, when  $\beta = \gamma \alpha$ , we will see soon see that (3.4) is the free energy of a factor graph with weights  $\beta$  and temperature  $\gamma$  (Theorem 3.7). Alternatively,  $\gamma$  can be thought of as one final degree of confidence, in the same scale as the values of  $\beta$ , in the qualitative information overall.

For these reasons, and also to simplify the math, we have chosen to use the

asymmetric variant in (3.4) that uses the parameter  $\gamma \in [0, \infty]$  to control the strength of  $SDef$ , rather than the symmetric trade-off parameter  $\hat{\gamma} \in [0, 1]$ .

### 3.3.3 PDGs As Unique Distributions

Finally, we provide an interpretation of a PDG as a probability distribution. Before we provide this semantics, we stress that this distribution does *not* capture all of the important information in the PDG—for example, a PDG can represent inconsistent knowledge states. Still, by giving a distribution, we enable comparisons with other graphical models, and show that PDGs are a surprisingly flexible tool for specifying distributions. The idea is to select the distributions with the best score. We thus define

$$\llbracket m \rrbracket_{\gamma}^* = \arg \min_{\mu \in \Delta V \mathcal{X}} \llbracket m \rrbracket_{\gamma}(\mu). \quad (3.5)$$

In general,  $\llbracket m \rrbracket_{\gamma}^*$  does not yield a unique distribution. But if  $\gamma$  is sufficiently small, then it does:

**Proposition 3.2.** *If  $0 < \gamma \leq \min_{a \in A} \frac{\beta_a^m}{\alpha_a^m}$ , then  $\llbracket m \rrbracket_{\gamma}^*$  is a singleton.*

[ link to proof ]

This is because the scoring function  $\llbracket m \rrbracket_{\gamma}$  is strictly convex for these (small) values of  $\gamma$ . In much of this dissertation, we will be particularly interested in the case where  $\gamma$  is small, which means emphasizing the accuracy of the probability

Bring back up  
Max Ent?

distribution as a description of probabilistic information, over the graphical structure of the PDG. This motivates us to consider what happens as  $\gamma$  goes to 0. If  $S_{\gamma}$  is a set of probability distributions for all  $\gamma > 0$ , we define  $\lim_{\gamma \rightarrow 0} S_{\gamma}$  to consist of all distributions  $\mu$  such that there is a sequence  $(\gamma_i, \mu_i)_{i \in \mathbb{N}}$  with  $\gamma_i \rightarrow 0$  and  $\mu_i \rightarrow \mu$  such that  $\mu_i \in S_{\gamma_i}$  for all  $i$ . It can be further shown that

**Theorem 3.3.** For all proper PDGs (such as when  $\beta > 0$ ),  $\lim_{\gamma \rightarrow 0} \llbracket m \rrbracket_\gamma^*$  is a singleton.

Let  $\llbracket m \rrbracket_{0+}^*$  be the unique element of  $\lim_{\gamma \rightarrow 0} \llbracket m \rrbracket_\gamma^*$ . This limiting semantics intuitively represents the “quantitative” or “observational” extreme, in which the quantitative probabilities and their confidences  $(\mathbb{P}, \beta)$  dominate, and the qualitative information  $(\mathcal{A}, \alpha)$  is used only to break ties. This makes the semantics  $\llbracket - \rrbracket_{0+}^*$  a direct analogue of (and, indeed, a generalization of) the principle of maximum entropy. This semantics has an important property:

**Proposition 3.4.**  $\llbracket m \rrbracket_{0+}^* \in \llbracket m \rrbracket_0^*$ , so if  $m$  is consistent, then  $\llbracket m \rrbracket_{0+}^* \in \{\!\{ m \}\!}$ .

[ link to proof ]

Thus,  $\llbracket m \rrbracket_{0+}^*$  is the unique distribution among those consistent with the cpds of a consistent PDG, that has minimal structural deficiency.

### 3.3.4 Further Semantics: PDGs as Degrees of Inconsistency

#### and as Transformations on Distributions

The three semantics given in the previous section are the ones presented in the original PDG paper ([Richardson and Halpern 2021](#)). However, there are other semantics that will play an important role later on. Far and away the most important semantics, which we have talked about at length but still not explicitly defined, is a PDG’s (*degree of*) *inconsistency*: the smallest possible incompatibility of  $m$  with any distribution. More precisely, for each  $\gamma$ , we can define

$$\langle\!\langle m \rangle\!\rangle_\gamma := \inf_{\mu \in \Delta \mathcal{VX}} \llbracket m \rrbracket_\gamma(\mu) \quad \in \bar{\mathbb{R}}.$$

Thus, it interprets a PDG as a single number. This quantity may at first seem less important than the others, but we will see in [Chapter 6](#) that it is a “universal loss

function”, and then in [Chapter 9](#) that all of the other semantics can be derived from it.

In [Chapter 12](#), we will see how PDGs also naturally arise from a second perspective, where they have a semantics in terms of belief updates, so that we can interpret a PDG  $\mathcal{M}$  as a function  $[\mathcal{M}]^\top : \Delta \mathcal{V}\mathcal{X} \rightarrow \Delta \mathcal{V}\mathcal{X}$  that updates distributions. Both of these alternate semantics turn out to be equivalent to the scoring function semantics developed in [Section 3.3.2](#).

### 3.4 Relationships to Other Graphical Models

In this section, we relate PDGs to two of the most popular graphical models: BNs and factor graphs. PDGs are strictly more general than BNs, and emulate factor graphs for the particular value  $\gamma = 1$ . More precisely, the distribution specified by a BN  $\mathcal{B}$  is the unique one that minimizes both  $OInc_{\mathcal{B}}$  and  $SDef_{\mathcal{B}}$  (and hence every positive linear combination of the two), while the distribution specified by a factor graph  $\Phi$  uniquely minimizes the sum  $OInc_\Phi + SDef_\Phi$  in which the observational and structural information are weighted equally. Now for the details.

#### 3.4.1 Bayesian Networks

[Construction 3.2](#) can be generalized to convert arbitrary Bayesian Networks into PDGs. Given a BN  $\mathcal{B}$  and a positive confidence  $\beta_X$  for the cpd of each variable  $X$  of  $\mathcal{B}$ , let  $\mathcal{M}_{\mathcal{B}, \beta}$  be the PDG comprising the cpds of  $\mathcal{B}$  in this way.

**Theorem 3.5.** *If  $\mathcal{B}$  is a Bayesian network and  $\Pr_{\mathcal{B}}$  is the distribution it specifies, then*

[ link to proof ]

for all  $\gamma > 0$  and all vectors  $\beta$  such that  $\beta_a > 0$  for all arcs  $a$ ,  $[\mathbf{m}_{\mathcal{B}, \beta}]^*_\gamma = \{\text{Pr}_{\mathcal{B}}\}$ , and thus  $[\mathbf{m}_{\mathcal{B}, \beta}]^* = \text{Pr}_{\mathcal{B}}$ .

[Theorem 3.5](#) is quite robust to parameter choices: it holds for every weight vector  $\beta$  and all  $\gamma > 0$ . However, it does lean heavily on our assumption that  $\alpha = \mathbf{1}$ , intuitively because that is what encodes the BN's independencies.

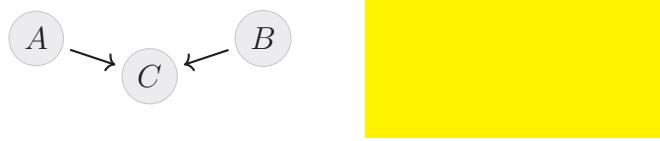
An intuitive proof sketch explains why: a distribution  $\mu$  minimizes  $OInc_{\mathcal{B}}$  iff it has the right cpds (no matter the value of  $\beta$ ), and (as we show in the appendix) it minimizes  $SDef_{\mathcal{B}}$  iff it has the appropriate independencies. We know that the BN's distribution  $\text{Pr}_{\mathcal{B}}$  is the unique distribution with both of these properties. So, no matter how the two quantities are weighted (i.e., for all  $\gamma > 0$ ), it is the unique optimal distribution.

**BNs and Maximum Entropy.** One welcome side effect of this result is to shine light on an old question about the role of maximum entropy in directed graphical models. In [Section 2.6.1](#), we mentioned that BNs are among the few standard ways of specifying probabilistic information that are not maximum-entropy distributions subject to the observational constraints (the specified cpds). To get any result of the kind, prior work has resorted to building up the distribution “in causal order”, applying the principle of maximum entropy many times ([Williamson 2001](#)). This makes some sense, but why should it be necessary?

First, we note that maximum entropy does not always miss the mark: if the BN happens to have the structure of a Markov chain, or more generally be absent of joint dependence, then it is not hard to show that the maximum entropy distribution is in fact the correct one. Here's a broader but more subtle class of BNs for which maximum entropy produces the right answer: if for every node  $X$

with multiple parents, the entropy of conditional probability  $\mathbb{P}_X(X \mid \text{Pa}(X)=\mathbf{z}) \in \Delta \mathcal{V}X$  does not depend on  $\mathbf{z}$ . To see why this would help, it is helpful to see an example of a BN whose distribution does not maximize entropy subject to the observational constraints.

**Example 3.5.** Consider the BN



where  $A$  and  $B$  are binary, and  $C$  can take  $2^k$  values, including  $c_0$ . For the probability tables, suppose that both  $A$  and  $B$  get uniform unconditional probabilities ( $1/2$  apiece), and set  $C$ 's cpt to be

$$\mathbb{P}_C(C \mid A, B) = \begin{bmatrix} \Delta C \\ \text{Unif}(C) \\ \delta_{c_0}(C) \\ \delta_{c_0(C)} \\ \text{Unif}(C) \end{bmatrix} \begin{array}{ll} & a, b \\ & \neg a, b \\ & a, \neg b \\ & \neg a, \neg b \end{array}$$

where  $\delta_{c_0}$  places all probability on  $c_0$ . It has zero entropy, while the uniform distribution on  $C$  has  $k$  bits of entropy, and  $A$  and  $B$  both have one bit of entropy. The semantics of a BN require that  $A$  and  $B$  are independent. Thus, the BN's distribution has  $2 + k/2$  bits of entropy (one bit for each of  $A$  and  $B$  since they are independent, plus an expected  $k/2$  bits from getting the uniform distribution on  $C$  with probability  $1/2$ ). However, a distribution in which  $A$  and  $B$  are correlated so as to always be equal has  $1 + k$  bits (one bit from  $A$  and  $B$ , and  $k$  bits from  $C|A, B$ ). For  $k > 2$ , the latter distribution has larger entropy. Therefore, the maximum entropy distribution consistent with the tables does not encode the independence assumption that a BN does.  $\triangle$

The key observation is that the cpds (because they encode *conditional* information)

tion), can carry different amounts of entropy depending on the realized values of the variables. But intuitively, the principle of maximum entropy was trying to maximize entropy *beyond* the constraints—not within them. Adjusting the principle of maximum entropy so as not to account for the entropy implied by the given cpds leads us directly back to the formula for *SDef*; indeed, this is how we initially discovered it. We argue that the principle of *minimizing structural deficiency* (*SDef*) subject to observational constraints (i.e., the  $0^+$  semantics of a PDG) is the appropriate generalization of the principle of maximum entropy, when there is conditional (and especially causal) information involved. [Theorem 3.5](#) is a strong argument in favor of this; we will see others examples later on.

### 3.4.2 Factor Graphs

Factor graphs ([Kschischang et al. 2001](#)) are another class of graphical model, which are often thought of as generalizing BNs.<sup>2</sup> In this section, we explore the relationship between PDGs and factor graphs.

**Definition 3.7.** A *factor graph*  $\Phi$  is a set of random variables  $\mathcal{X}$  and an indexed collection of *factors*  $\{\phi_J : \mathcal{V}(X_J) \rightarrow \mathbb{R}_{\geq 0}\}_{J \in \mathcal{J}}$ , where  $\mathcal{J}$  is an arbitrary index set, and each  $X_J \subseteq \mathcal{X}$ . In other words, each factor  $\phi_J$  is associated with a subset  $X_J \subseteq \mathcal{X}$  of variables, and maps joint settings of  $X_J$  to non-negative real numbers. The factor graph  $\Phi$  specifies a distribution

$$\Pr_{\Phi}(\vec{x}) := \frac{1}{Z_{\Phi}} \prod_{J \in \mathcal{J}} \phi_J(\vec{x}_J),$$

where  $\vec{x} \in \mathcal{V}\mathcal{X}$  is a joint setting of all of the variables,  $\vec{x}_J$  is the restriction of

---

<sup>2</sup>This claim is true at a quantitative level, but it is worth noting that factor graphs do not capture the same independencies unless augmented with extra information ([Frey 2012](#)), and lack some properties that makes BNs useful in causality.

$\vec{x}$  to only the variables  $X_J$ , and  $Z_\Phi$  is the constant required to normalize the distribution.  $\square$

To view a BN as a factor graph, one essentially views each cpd as a factor. From this perspective, a PDG also seems to specify a collection of factors. How do PDG semantics as we have already defined them compare to this “multiply-and-renormalize” semantics of a factor graph? To answer this, we start by making the translation precise.

**Definition 3.8** (unweighted PDG to factor graph). Given an unweighted PDG  $\mathbb{P} = \{\mathbb{P}_a(T_a | S_a)\}_{a \in \mathcal{A}}$  define the associated FG  $\Phi_n$  as follows: take  $\mathcal{J} := \mathcal{A}$ , and for each  $Z \xrightarrow{a} Y \in \mathcal{J}$ , let  $X_a := \{Z, Y\}$  with factor  $\phi_a(z, y) := \mathbb{P}_a(y | z)$ .  $\square$

not  
anymore

It turns out we can also do the reverse. Using essentially **the same idea as in Construction 3.2**, we can encode a factor graph as an assertion about the unconditional probability distribution over the variables associated to each factor.

**Definition 3.9** (factor graph to unweighted PDG). For a FG  $\Phi = \{\phi_J : \mathbf{X}_J \rightarrow \mathbb{R}\}_{J \in \mathcal{J}}$  over variables  $\mathcal{X}$ , let  $\mathbb{P}_\Phi$  be the unweighted PDG over the same variables  $\mathcal{X}$ , with hypergraph  $\mathcal{A} = \{\emptyset \rightarrow \mathbf{X}_J\}_{J \in \mathcal{J}}$ , such that each arc  $J$  is associated with the renormalized distribution  $\mathbb{P}_J(\mathbf{X}_J) : \propto \phi_J(\mathbf{X}_J) = \phi_J(\mathbf{X}_J) / \sum_{\mathbf{x} \in \mathbf{X}_J} \phi_J(\mathbf{x})$ . The process is illustrated in [Figure 3.5](#).  $\square$

PDGs are directed graphs, while factors graphs are undirected. The map from PDGs to factor graphs thus loses this structure, as shown in [Figure 3.5](#). The resulting PDG is also typically inconsistent. The structural changes are even more significant if we convert it to a *strict* PDG that does not have hyperedges, as done in the original paper ([Richardson and Halpern 2021](#)) and illustrated in

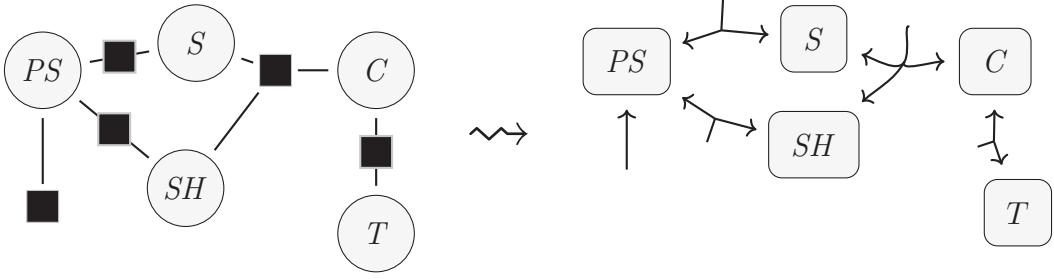


Figure 3.5: Conversion of the PDG in Example 3.2 to a factor graph according to Definition 3.8 (left), and from that factor graph back to a PDG by Definition 3.9 (right). The factors are associated with the unconditional distribution obtained by normalizing the appropriate factor.

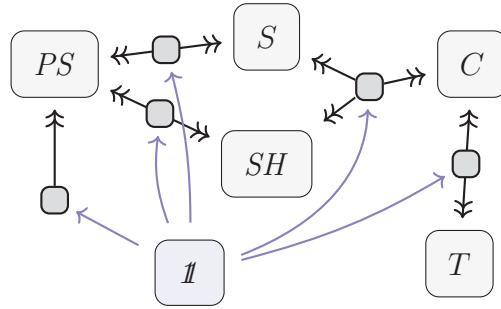


Figure 3.6: The conversion from a factor graph to a strict PDG, as done in the original paper (Richardson and Halpern 2021). In this case, for each factor  $\phi_J$  we introduce a new variable  $X_J$  (displayed as a smaller darker rectangle), whose values are joint settings of the variables connected to it, and also an arc  $1 \rightarrow X_J$  (shown in blue), to which we associate the unconditional distribution given by normalizing  $\phi_J$ .

Figure 3.6. Nevertheless, when  $\gamma = 1$ , so that the quantitative and qualitative terms are weighted equally, the semantics coincide.

**Theorem 3.6.** (a)  $\llbracket \mathcal{N}_\Phi \rrbracket_1^* = \{\Pr_\Phi\}$  for all factor graphs  $\Phi$ .<sup>3</sup>

[ link to proof ]

(b)  $\llbracket \mathcal{N} \rrbracket_1^* = \Pr_{\Phi_n}$  for all unweighted PDGs  $\mathcal{N}$ .

The correspondence hinges on the fact that we take  $\gamma = 1$ , so that *OInc* and *SDef* are weighted equally. Because  $\gamma$  is not part of the data of a PDG but rather a parameter of the semantics, this means the  $\gamma = 1$  semantics of the PDG  $\mathcal{N}_\Phi$

---

<sup>3</sup>Recall that we identify the unweighted PDG  $(\mathcal{A}, \mathbf{p})$  with the weighted PDG  $(\mathcal{A}, \mathbb{P}, \mathbf{1}, \mathbf{1})$ .

exactly captures the traditional semantics of  $\Phi$ . However, the fact that the reverse correspondence requires  $\gamma = 1$  suggests that factor graphs are less flexible than unweighted PDGs: factor graphs only describe one particular aspect of their semantics. In particular, it rules out the maximum entropy semantics.

What about PDGs for which  $\beta \neq 1$ ? There is also a standard notion of weighted factor graph, but to relate them to PDGs with weights, we must stray from this chapter's convention of taking  $\alpha = 1$ .

### 3.4.3 Factored Exponential Families

A *weighted factor graph* (WFG)  $\Psi$  is a pair  $(\Phi, \theta)$  consisting of a factor graph  $\Phi$  together with a vector of non-negative weights  $\{\theta_J\}_{J \in \mathcal{J}}$ .  $\Psi$  specifies a canonical scoring function

$$VFE_{\Psi}(\mu) := \mathbb{E}_{\vec{x} \sim \mu} \left[ \sum_{J \in \mathcal{J}} \theta_J \log \frac{1}{\phi_J(\vec{x}_J)} \right] - H(\mu), \quad (3.6)$$

called the *variational Gibbs free energy* (Mezard and Montanari 2009).  $VFE_{\Psi}$  is uniquely minimized by the distribution  $\Pr_{\Psi}(\vec{x}) = \frac{1}{Z_{\Psi}} \prod_{J \in \mathcal{J}} \phi_J(\vec{x}_J)^{\theta_J}$ , which matches the unweighted case when every  $\theta_J = 1$ . The mapping  $\theta \mapsto \Pr_{(\Phi, \theta)}$  is known as  $\Phi$ 's *exponential family*, and is central to the analysis and development of many algorithms for undirected graphical models (Wainwright et al. 2008). Indeed, this is an alternate and in principle equally expressive approach to modeling with factor graphs: instead of choosing the factors  $\Phi$  based on data and assuming weight 1 on all factors, it is possible to instead choose  $\Phi$  beforehand to be a suitably expressive basis, and then use data to inform the choice of *weights*  $\theta$ .

PDGs can capture the full exponential family of a factor graph, but it requires value of  $\alpha$  other than 1. It turns out that the key is to ensure that the ratio  $\alpha_a/\beta_a$  is

constant across arcs  $a$ . We therefore define a family of translations, parameterized by the ratio of  $\alpha_a$  to  $\beta_a$ .

**Definition 3.10** (WFG to PDG). Given a WFG  $\Psi = (\Phi, \theta)$ , and positive number  $k$ , we define the corresponding PDG  $\mathcal{M}_{\Psi,k} = (\mathbf{n}_\Phi, \alpha_\theta, \beta_\theta)$  by taking  $\beta_J = k\theta_J$  and  $\alpha_J = \theta_J$  for the arc  $\mathbb{I} \rightarrow X_J$ .  $\square$

We now extend Definitions 3.8 and 3.9 to (weighted) PDGs and WFGs. In translating a PDG to a WFG, there will necessarily be some loss of information: PDGs have two sets, while WFGs have only one. Here we throw out  $\alpha$  and keep  $\beta$ , though in its role here as a left inverse of Definition 3.10, either choice would suffice.

**Definition 3.11** (PDG to WFG). Given a (weighted) PDG  $\mathcal{M} = (\mathbf{n}, \beta)$ , we take its corresponding WFG to be  $\Psi_{\mathcal{M}} := (\Phi_{\mathbf{n}}, \beta)$ ; that is,  $\theta_a := \beta_a$  for all arcs  $a$ .  $\square$

We now show that we can capture the entire exponential family of a factor graph, and even its associated free energy, but only for  $\gamma$  equal to the constant  $k$  used in the translation.

**Theorem 3.7.** For all WFGs  $\Psi = (\Phi, \theta)$  and all  $\gamma > 0$ , we have that  $VFE_\Psi = \frac{1}{\gamma} \llbracket \mathcal{M}_{\Psi,\gamma} \rrbracket_\gamma + C$  for some constant  $C$ , so  $\text{Pr}_\Psi$  is the unique element of  $\llbracket \mathcal{M}_{\Psi,\gamma} \rrbracket_\gamma^*$ . [link to proof]

In particular, for  $k=1$ , so that  $\theta$  is used for both the functions  $\alpha$  and  $\beta$  of the resulting PDG, Theorem 3.7 strictly generalizes Theorem 3.6.

**Corollary 3.7.1.** For all WFGs  $(\Phi, \theta)$ , we have that  $\text{Pr}_{(\Phi, \theta)} = \llbracket (\mathbf{n}_\Phi, \theta, \theta) \rrbracket_1^*$ .

Conversely, as long as the ratio of  $\alpha_a$  to  $\beta_a$  is constant, the reverse translation

also preserves semantics. Recall that  $(\mathbb{P}, \alpha, \beta)$  is the PDG obtained by augmenting the unweighted PDG  $\mathbb{P}$  with weights over every arc.

**Theorem 3.8.** *For all unweighted PDGs  $(\mathcal{A}, \mathbb{P})$ , non-negative vectors  $\mathbf{v}$  of shape  $\mathcal{A}$ , and all  $\gamma > 0$ , we have that  $\llbracket (\mathbb{P}, \mathbf{v}, \gamma \mathbf{v}) \rrbracket_\gamma = \gamma \cdot VFE_{(\Phi_{\mathbb{P}}, \mathbf{v})}$ ; consequently,  $\llbracket (\mathbb{P}, \mathbf{v}, \gamma \mathbf{v}) \rrbracket_\gamma^* = \{\Pr_{(\Phi_{\mathbb{P}}, \mathbf{v})}\}$ . In other words, if  $\mathcal{M}$  is a PDG with  $\beta \propto \alpha$ , and  $\gamma$  is the constant of proportionality, then  $\llbracket \mathcal{M} \rrbracket_\gamma^*$  is a singleton containing the distribution of the factor graph  $\Phi_{\mathbb{P}}$  weighted by  $\alpha$ .*

[ link to proof ]

In particular, when  $\alpha = \beta$ , then  $\llbracket \mathcal{M} \rrbracket_1$  is the semantics of the factor graph that regards the cpds  $\mathbb{P}$  as factors with these weights. The key step in proving Theorems 3.7 and 3.8 (and in the proofs of a number of other results) involves rewriting the scoring function as follows.

**Proposition 3.9.** *For all PDGs  $\mathcal{M}$ ,  $\llbracket \mathcal{M} \rrbracket_\gamma(\mu) =$*

$$\mathbb{E}_\mu \left[ \sum_{X \xrightarrow{a} Y \in \mathcal{A}} \underbrace{\left( \beta_a \log \frac{1}{\mathbb{P}_a(Y|X)} + (\alpha_a \gamma - \beta_a) \log \frac{1}{\mu(Y|X)} \right)}_{\text{log likelihood / cross entropy}} - \underbrace{\gamma \log \frac{1}{\mu(\mathcal{X})}}_{\text{global regularization}} \right]. \quad (3.7)$$

[ link to proof ]

For a fixed  $\gamma$ , the first and last terms of (3.7) are equal to a scaled version of the free energy,  $\gamma VFE_\Phi$ , if we set  $\phi_J := \mathbb{P}_a$  and  $\theta_J := \beta_a/\gamma$ . If, in addition,  $\beta_a = \alpha_a \gamma$  for all arcs  $a$ , then the local regularization term disappears, giving us the desired correspondence. Equation (3.7) also makes it clear that taking  $\beta_a = \alpha_a \gamma$  for all arcs  $a$  is essentially necessary to get the result of Theorem 3.6. Of course, fixed  $\gamma$  precludes taking the limit as  $\gamma$  goes to 0, so Proposition 3.4 does apply to factor graphs. This is reflected in some strange behavior in factor graphs trying to capture the same phenomena as PDGs, as the following example shows.

**Example 3.6.** Consider the PDG  $\mathcal{M}$  containing just one variable  $X$ , and two arcs. (Such a PDG can arise if we get different information about the probability of  $X$  from two different sources; this is a situation we certainly want to be able to capture!) Suppose further that  $p$  and  $q$  are both associated with the same distribution on  $X$ . For definiteness, suppose that  $VX = \{x_1, x_2\}$ , and that the distribution associated with both arcs is  $\mu_{.7}$ , which ascribes probability .7 to  $x_1$ . Then, as we would hope  $[\mathcal{M}]_{0^+}^* = \{\mu_{.7}\}$ ; after all, both sources agree on the information. However, it can be shown that  $\Pr_{\Psi_m} = \mu_{.85}$ , so  $[\mathcal{M}]_1^* = \{\mu_{.85}\}$ . In this way, factor graphs seem to be *uncalibrated*.

The reason for this can be viewed as an instance of the caveat in combining qualitative information, mentioned in Section 3.2.2. Our default assumption that  $\alpha = 1$  makes sense for each individual arc, but having two independent randomized mechanisms by which  $X$  is determined intuitively means  $X$  ought to be a constant (a point we will expand on in Section 5.2). Arguably, this is not what was intended in combining the two models. By taking the  $0^+$  semantics, the problem with the qualitative information disappears. Alternatively, we could have merged the qualitative picture so that there is a total of  $\alpha_p + \alpha_q = 1$  between the two arcs. In that case, we would have had no conflict between the two types of information, and indeed  $[\mathcal{M}']_\gamma^* = \{\mu_{.7}\}$  for all values of  $\gamma \geq 0$ .  $\triangle$

could be swept under the rug

Although both  $\theta$  and  $\beta$  are measures of confidence, the way that the Gibbs free energy varies with  $\theta$  is quite different from the way that the score of a PDG varies with  $\beta$ . The scoring function that we use for PDGs can be viewed as extending  $VFE_{\Phi,\theta}$  by including the local regularization term. As  $\gamma$  approaches zero, the importance of the global regularization terms decreases relative to that of the local regularization term, so the PDG scoring function becomes quite different

from Gibbs free energy.

### 3.5 Discussion

We have introduced PDGs, a powerful tool for representing beliefs in the language of probabilities and confidences. They have a number of advantages over other probabilistic graphical models. Most notably:

- PDGs allow us to capture inconsistent beliefs, including conflicting information from multiple sources with varying degrees of reliability. Moreover, they provide a natural way of measuring the degree of this inconsistency.
- PDGs are much more modular than other representations; for example, we can combine information from two sources by simply taking the union of two PDGs, and it is easy to add new information (arcs) and features (nodes) without affecting (the meaning of) previously-received information.
- They allow for a clean separation between quantitative information (the cpds  $\mathbb{P}$  and weights  $\beta$ ) and the qualitative information (carried by the graph structure  $\mathcal{A}$  and the weights  $\alpha$ ); this separation is captured semantically by the terms  $OInc$  and  $SDef$  in our scoring function.
- PDGs have (several) natural semantics; one of them allows us to pick out a unique distribution. Using this distribution, PDGs can capture BNs and factor graphs. In the latter case, a simple parameter shift in the corresponding PDG eliminates arguably problematic behavior of a factor graph.

So far, we have only scratched the surface of what can be done with PDGs. Two major issues that need to be tackled are inference and dynamics. How

should we query a PDG for probabilistic information? How should we modify a PDG in light of new information or to make it more consistent? These issues turn out to be closely related, as we will see in [Chapter 9](#).

For now, we content ourselves with a simple example: conditioning can be understood in terms of resolving inconsistencies in a PDG. To condition on an observation  $Y = y$ , given a situation described by a PDG  $\mathcal{M}$ , we can add an arc from  $\mathbb{1}$  to  $Y$  in  $\mathcal{M}$ , annotated with the cpd that gives probability 1 to  $y$ , to get the (likely inconsistent) PDG  $\mathcal{M}^{+(Y=y)}$ . In the special case where  $\mathcal{M}$  represents or is equivalent to a joint distribution,  $[\mathcal{M}^{+(Y=y)}]^*$  turns out to be the result of conditioning that distribution on the event  $Y = y$ . This account of conditioning generalizes without modification to give Jeffrey's Rule ([Jeffrey 1968](#)), a more general approach to belief updating. This will be made precise in [Section 4.1](#).

Properly modeling inconsistency also turns out to be a recurrent theme in machine learning, as we will see in [Chapter 6](#). A variational autoencoder ([Kingma and Welling 2014](#)), for instance, is essentially three cpds: a prior  $p(Z)$ , a decoder  $p(X | Z)$ , and an encoder  $q(Z | X)$ . Because two cpds target  $Z$  (and the cpds are inconsistent until fully trained), this situation can be represented by PDGs but not by other graphical models.

^directed

In the coming chapters, we will explore these connections, and the broader theory of probabilistic modeling that is enabled by the PDG representation.

## APPENDICES FOR CHAPTER 3

### 3.A Proofs

For brevity, when the appropriate variables are clear from context, we write  $\mu(x, y)$  in place of  $\mu(X = x, Y = y)$ ,  $\mu(x | y)$  in place of  $\mu(X = x | Y = y)$ , and so forth.

#### 3.A.1 Properties of the Scoring Semantics

In this section, we prove the properties of scoring functions that we mentioned in the main text, Propositions 3.1, 3.2, and 3.4. We repeat the statements for the reader's convenience.

**Proposition 3.1.** *If  $\mathcal{M}$  is a PDG with  $\beta > 0$ , then  $\mu \in \{\mathcal{M}\}$  iff  $OInc_{\mathcal{M}}(\mu) = 0$ .*

*Proof.* By taking  $\gamma = 0$ , the score is just  $OInc$ . By definition, a distribution  $\mu \in \{\mathcal{M}\}$  satisfies all the constraints, so  $\mu(Y = \cdot | X = x) = \mathbb{P}_a(x)$  for all arcs  $X \rightarrow Y \in \mathcal{A}^{\mathcal{M}}$  and  $x$  with  $\mu(X = x) > 0$ . By Gibbs' inequality (MacKay 2003),  $D(\mu(Y|x) \| \mathbb{P}_a(x)) = 0$ . Since this is true for all arcs, we must have  $OInc_{\mathcal{M}}(\mu) = 0$ . Conversely, if  $\mu \notin \{\mathcal{M}\}$ , then it fails to marginalize to the cpd  $\mathbb{P}_a$  on some arc  $a$ , and so again by Gibbs inequality,  $D(\mu(Y|x) \| \mathbb{P}_a(x)) > 0$ . As relative entropy is non-negative, the sum of these terms over all arcs must be positive as well, (because we have assumed  $\beta > 0$ ) and so  $OInc_{\mathcal{M}}(\mu) \neq 0$ .  $\square$

Before proving the remaining results, we prove a lemma that will be useful in other contexts as well.

The next proposition gives us a useful representation of  $\llbracket \mathbf{m} \rrbracket_\gamma$ , and letting us decompose into three more interpretable pieces.

**Proposition 3.9.** *For all PDGs  $\mathbf{m}$ ,  $\llbracket \mathbf{m} \rrbracket_\gamma(\mu) =$*

$$\mathbb{E}_\mu \left[ \sum_{X \xrightarrow{a} Y \in \mathcal{A}} \underbrace{\left( \beta_a \log \frac{1}{\mathbb{P}_a(Y|X)} + (\alpha_a \gamma - \beta_a) \log \frac{1}{\mu(Y|X)} \right)}_{\text{log likelihood / cross entropy}} - \underbrace{\gamma \log \frac{1}{\mu(\mathcal{X})}}_{\text{global regularization}} \right]. \quad (3.7)$$

*Proof.* We use the more general formulation of *SDef* given in [Section 3.4.3](#), in which each arc  $a$ 's conditional information is weighted by  $\alpha_a$ .

$$\begin{aligned} \llbracket \mathbf{m} \rrbracket_\gamma(\mu) &:= OInc_m(\mu) + \gamma SDef_m(\mu) \\ &= \left[ \sum_{X \xrightarrow{a} Y} \beta_a \mathbb{E}_{x \sim \mu_X} \mathbf{D}\left(\mu(Y|X=x) \parallel \mathbb{P}_a(x)\right) \right] + \gamma \left[ \sum_{X \xrightarrow{a} Y} \alpha_a H_\mu(Y | X) - H(\mu) \right] \\ &= \sum_{X \xrightarrow{a} Y} \mathbb{E}_{x \sim \mu_X} \left[ \beta_a \mathbf{D}\left(\mu(Y | x) \parallel \mathbb{P}_a(Y | x)\right) + \gamma \alpha_a H(Y | X=x) \right] - \gamma H(\mu) \\ &= \sum_{X \xrightarrow{a} Y} \mathbb{E}_{x \sim \mu_X} \left[ \beta_a \left( \sum_{y \in V(Y)} \mu(y | x) \log \frac{\mu(y | x)}{\mathbb{P}_a(y | x)} \right) \right. \\ &\quad \left. + \alpha_a \gamma \left( \sum_{y \in V(Y)} \mu(y | x) \log \frac{1}{\mu(y | x)} \right) \right] - \gamma H(\mu) \\ &= \sum_{X \xrightarrow{a} Y} \mathbb{E}_{x \sim \mu_X} \left[ \sum_{y \in V(Y)} \mu(y | x) \left( \beta_a \log \frac{\mu(y | x)}{\mathbb{P}_a(y | x)} + \alpha_a \gamma \log \frac{1}{\mu(y | x)} \right) \right] - \gamma H(\mu) \\ &= \sum_{X \xrightarrow{a} Y} \mathbb{E}_{x \sim \mu_X} \left[ \mathbb{E}_{y \sim \mu(Y|X=x)} \left( \beta_a \log \frac{\mu(y | x)}{\mathbb{P}_a(y | x)} + \alpha_a \gamma \log \frac{1}{\mu(y | x)} \right) \right] - \gamma \sum_{\mathbf{w} \in \mathcal{V}\mathcal{X}} \mu(\mathbf{w}) \log \frac{1}{\mu(\mathbf{w})} \\ &= \sum_{X \xrightarrow{a} Y} \mathbb{E}_{x, y \sim \mu_{XY}} \left[ \beta_a \log \frac{\mu(y | x)}{\mathbb{P}_a(y | x)} + \alpha_a \gamma \log \frac{1}{\mu(y | x)} \right] - \gamma \mathbb{E}_{\mathbf{w} \sim \mu} \left[ \log \frac{1}{\mu(\mathbf{w})} \right] \\ &= \mathbb{E}_{\mathbf{w} \sim \mu} \left\{ \sum_{X \xrightarrow{a} Y} \left[ \beta_a \log \frac{1}{\mathbb{P}_a(y|x)} - \beta_a \log \frac{1}{\mu(y|x)} + \alpha_a \gamma \log \frac{1}{\mu(y|x)} \right] - \gamma \log \frac{1}{\mu(\mathbf{w})} \right\} \\ &= \mathbb{E}_{\mathbf{w} \sim \mu} \left\{ \sum_{X \xrightarrow{a} Y} \left[ \beta_a \log \frac{1}{\mathbb{P}_a(y | x)} + (\alpha_a \gamma - \beta_a) \log \frac{1}{\mu(y | x)} \right] - \gamma \log \frac{1}{\mu(\mathbf{w})} \right\}. \end{aligned}$$

□

We can now prove Proposition 3.2.

**Proposition 3.2.** *If  $0 < \gamma \leq \min_{a \in \mathcal{A}} \frac{\beta_a^m}{\alpha_a^m}$ , then  $\llbracket m \rrbracket_\gamma^*$  is a singleton.*

*Proof.* It suffices to show that  $\llbracket m \rrbracket_\gamma$  is a strictly convex function of  $\mu$ , since every strictly convex function has a unique minimum. Note that

$$\begin{aligned} \llbracket m \rrbracket_\gamma(\mu) &= \mathbb{E}_{\mathbf{w} \sim \mu} \left\{ \sum_{X \xrightarrow{a} Y} \left[ \beta_a \log \frac{1}{\mathbb{P}_a(y|x)} + (\alpha_a \gamma - \beta_a) \log \frac{1}{\mu(y|x)} \right] - \gamma \log \frac{1}{\mu(\mathbf{w})} \right\} \\ &= \mathbb{E}_{\mathbf{w} \sim \mu} \left\{ \sum_{X \xrightarrow{a} Y} \left[ \gamma \alpha_a \log \frac{1}{\mathbb{P}_a(y|x)} + (\beta_a - \alpha_a \gamma) \log \frac{1}{\mathbb{P}_a(y|x)} \right. \right. \\ &\quad \left. \left. - (\beta_a - \alpha_a \gamma) \log \frac{1}{\mu(y|x)} \right] - \gamma \log \frac{1}{\mu(\mathbf{w})} \right\} \\ &= \mathbb{E}_{\mathbf{w} \sim \mu} \left\{ \sum_{X \xrightarrow{a} Y} \left[ \gamma \alpha_a \log \frac{1}{\mathbb{P}_a(y|x)} + (\beta_a - \alpha_a \gamma) \log \frac{\mu(y|x)}{\mathbb{P}_a(y|x)} \right] - \gamma \log \frac{1}{\mu(\mathbf{w})} \right\} \\ &= \sum_{X \xrightarrow{a} Y} \left[ \gamma \alpha_a \mathbb{E}_{(x,y) \sim \mu_{XY}} \log \frac{1}{\mathbb{P}_a(y|x)} + (\beta_a - \alpha_a \gamma) \mathbb{E}_{x \sim \mu_X} D(\mu(Y|x) \parallel \mathbb{P}_a(x)) \right] - \gamma H(\mu). \end{aligned}$$

The remainder of the argument analyzes the convexity of the three terms of this final formula. The first term,  $\mathbb{E}_{x,y \sim \mu_{XY}} [-\log \mathbb{P}_a(y|x)]$  is linear in  $\mu$  (since  $\mathbb{P}_a(y|x)$  does not depend on  $\mu$ ), and hence convex. As for the second term, it is well-known that KL divergence is convex, in the sense that

$$\begin{aligned} D(\lambda q_1 + (1-\lambda)q_2 \parallel \lambda p_1 + (1-\lambda)p_2) \\ \leq \lambda D(q_1 \parallel p_1) + (1-\lambda)D(q_2 \parallel p_2). \end{aligned}$$

Instantiate this for distributions over values of  $Y$ , setting  $p_1 = p_2 := \mathbb{P}_a(x)$ , and  $q_1 := \mu_1(Y|X=x)$  and  $q_2 := \mu_2(Y|X=x)$ , to get:

$$\begin{aligned} D(\lambda \mu_1(Y|x) + (1-\lambda) \mu_2(Y|x) \parallel \mathbb{P}_a(x)) \\ \leq \lambda D(\mu_1(Y|x) \parallel \mathbb{P}_a(x)) + (1-\lambda) D(\mu_2(Y|x) \parallel \mathbb{P}_a(x)). \end{aligned}$$

So  $D(\mu(Y \mid x) \parallel \mathbb{P}_a(Y \mid x))$  is convex as a function of  $\mu$ . As convex combinations of convex functions are convex, the second term,  $\mathbb{E}_{x \sim \mu(X)} D(\mu(Y \mid x) \parallel \mathbb{P}_a(x))$ , is convex. Finally, negative entropy (the third term) is well known to be strictly convex.

Any non-negative linear combinations of the three terms is convex, and if this combination applies a positive coefficient to the (strictly convex) negative entropy, it must be strictly convex. Therefore, as long as  $\beta_a \geq \gamma$  for all arcs  $a \in \mathcal{A}^m$ ,  $\llbracket \mathbf{m} \rrbracket_\gamma$  is strictly convex. The result follows.  $\square$

We next prove [Theorem 3.3](#). The first step is provided by the following lemma, which shows that any limiting optimal distributions as  $\gamma \rightarrow 0$  must also be optimal distributions for  $\gamma = 0$ .

**Lemma 3.10.**  $\lim_{\gamma \rightarrow 0} \llbracket \mathbf{m} \rrbracket_\gamma^* \subseteq \llbracket \mathbf{m} \rrbracket_0^*$ .

*Proof.* Since  $SDef_m$  is a finite weighted sum of entropies and conditional entropies over the variables  $\mathcal{N}^m$ , which have finite support, it is bounded. Thus, there exist bounds  $k$  and  $K$  depending only on  $\mathcal{N}^m$  and  $\mathcal{V}^m$ , such that  $k \leq SDef_m(\mu) \leq K$  for all  $\mu$ . Since  $\llbracket \mathbf{m} \rrbracket_\gamma = OInc_m + \gamma SDef_m$ , it follows that, for all  $\mu \in \Delta \mathcal{V}\mathcal{X}$ , we have

$$OInc_m(\mu) + \gamma k \leq \llbracket \mathbf{m} \rrbracket_\gamma(\mu) \leq OInc_m(\mu) + \gamma K.$$

For a fixed  $\gamma$ , since this inequality holds for all  $\mu$ , and both  $OInc$  and  $SDef$  are bounded below, it must be the case that

$$\min_{\mu \in \Delta \mathcal{V}\mathcal{X}} [OInc_m(\mu) + \gamma k] \leq \min_{\mu \in \Delta \mathcal{V}\mathcal{X}} \llbracket \mathbf{m} \rrbracket_\gamma(\mu) \leq \min_{\mu \in \Delta \mathcal{V}\mathcal{X}} [OInc_m(\mu) + \gamma K],$$

even though the distributions that minimize each expression will in general be different. Let  $OInc(\mathbf{m}) = \min_{\mu} OInc_m(\mu)$ . Since  $\Delta \mathcal{V}\mathcal{X}$  is compact, the minimum

of the middle term is achieved. Therefore, for  $\mu_\gamma \in \llbracket \mathbf{m} \rrbracket_\gamma^*(\mu)$  that minimizes it, we have

$$OInc(\mathbf{m}) + \gamma k \leq \llbracket \mathbf{m} \rrbracket_\gamma(\mu_\gamma) \leq OInc(\mathbf{m}) + \gamma K$$

for all  $\gamma \geq 0$ . Now taking the limit as  $\gamma \rightarrow 0$  from above, we get that  $OInc(\mathbf{m}) = \llbracket \mathbf{m} \rrbracket_0(\mu^*)$ . Thus,  $\mu^* \in \llbracket \mathbf{m} \rrbracket_0^*$ , as desired.  $\square$

We now apply Lemma 3.10 to show that the limit as  $\gamma \rightarrow 0$  is unique, as stated in Theorem 3.3.

**Theorem 3.3.** *For all proper PDGs (such as when  $\beta > 0$ ),  $\lim_{\gamma \rightarrow 0} \llbracket \mathbf{m} \rrbracket_\gamma^*$  is a singleton.*

*Proof.* First we show that  $\lim_{\gamma \rightarrow 0} \llbracket \mathbf{m} \rrbracket_\gamma^*$  cannot be empty. Let  $(\gamma_n) = \gamma_1, \gamma_2, \dots$  be a sequence of positive reals converging to zero. For all  $n$ , choose some  $\mu_n \in \llbracket \mathbf{m} \rrbracket_{\gamma_n}^*$ . Because  $\Delta V\mathcal{X}$  is a compact metric space, it is sequentially compact, and so, by the Bolzano–Weierstrass Theorem, the sequence  $(\mu_n)$  has at least one accumulation point, say  $\nu$ . By our definition of the limit,  $\nu \in \lim_{\gamma \rightarrow 0} \llbracket \mathbf{m} \rrbracket_\gamma^*$ , as witnessed by the sequence  $(\gamma_n, \mu_n)_n$ . It follows that  $\lim_{\gamma \rightarrow 0} \llbracket \mathbf{m} \rrbracket_\gamma^* \neq \emptyset$ .

Now, choose  $\nu_1, \nu_2 \in \lim_{\gamma \rightarrow 0} \llbracket \mathbf{m} \rrbracket_\gamma^*$ . Thus, there are subsequences  $(\mu_i)$  and  $(\mu_j)$  of  $(\mu_n)$  converging to  $\nu_1$  and  $\nu_2$ , respectively. By Lemma 3.10,  $\nu_1, \nu_2 \in \llbracket \mathbf{m} \rrbracket_0^*$ , so  $OInc_m(\nu_1) = OInc_m(\nu_2)$ . Because  $(\mu_{j_n}) \rightarrow \nu_1$ ,  $(\mu_{k_n}) \rightarrow \nu_2$ , and  $SDef_m$  is continuous on  $\Delta V\mathcal{X}$ , we conclude that  $(SDef_m(\mu_i)) \rightarrow SDef_m(\nu_1)$  and  $(SDef_m(\mu_j)) \rightarrow SDef_m(\nu_2)$ .

Suppose that  $SDef_m(\nu_1) \neq SDef_m(\nu_2)$ . Without loss of generality, suppose that  $SDef_m(\nu_1) > SDef_m(\nu_2)$ . Since  $(SDef_m(\mu_i)) \rightarrow SDef_m(\nu_1)$ , there exists some  $i^* \in \mathbb{N}$  such that for all  $i > i^*$ ,  $SDef_m(\mu_i) > SDef_m(\nu_2)$ . But then for all  $\gamma$  and

$i > i^*$ , we have

$$[\![\mathbf{m}]\!]_{\gamma}(\mu_i) = OInc(\mu_i) + \gamma SDef_m(\mu_i) > OInc(\nu_2) + \gamma SDef_m(\nu_2) = [\![\mathbf{m}]\!]_{\gamma}(\nu_2),$$

contradicting the assumption that  $\mu_i$  minimizes  $[\![\mathbf{m}]\!]_{\gamma}$ . We thus conclude that we cannot have  $SDef_m(\nu_1) > SDef_m(\nu_2)$ . By the same argument, we also cannot have  $SDef_m(\nu_1) < SDef_m(\nu_2)$ , so  $SDef_m(\nu_1) = SDef_m(\nu_2)$ .

Now, suppose that  $\nu_1$  and  $\nu_2$  distinct. Since  $[\![\mathbf{m}]\!]_{\gamma}$  is strictly convex for  $\gamma > 0$ , among the possible convex combinations of  $\nu_1$  and  $\nu_2$ , the distribution  $\nu_3 = \lambda\nu_1 + (1 - \lambda)\nu_2$  that minimizes  $[\![\mathbf{m}]\!]_{\gamma}$  must lie strictly between  $\nu_1$  and  $\nu_2$ . Because  $OInc$  itself is convex and  $OInc_m(\nu_1) = OInc_m(\nu_2) =: v$ , we must have  $OInc_m(\nu_3) \leq v$ . But since  $\nu_1, \nu_2 \in [\![\mathbf{m}]\!]_0^*$  minimize  $OInc$ , we must have  $OInc_m(\nu_3) \geq v$ . Thus,  $OInc_m(\nu_3) = v$ . Now, because, for all  $\gamma > 0$ ,

$$[\![\mathbf{m}]\!]_{\gamma}(\nu_3) = v + \gamma SDef_m(\nu_3) < v + \gamma SDef_m(\nu_1) = [\![\mathbf{m}]\!]_{\gamma}(\nu_1),$$

it must be the case that  $SDef_m(\nu_3) < SDef_m(\nu_1)$ .

We can now get a contradiction by applying the same argument as that used to show that  $SDef_m(\nu_1) = SDef_m(\nu_2)$ . Because  $(\mu_i) \rightarrow \nu_1$ , there exists some  $i^*$  such that for all  $i > i^*$ , we have  $SDef_m(\mu_i) > SDef_m(\nu_3)$ . Thus, for all  $i > i^*$  and all  $\gamma > 0$ ,

$$[\![\mathbf{m}]\!]_{\gamma}(\mu_i) = OInc(\mu_i) + \gamma SDef_m(\mu_i) > OInc(\nu_3) + \gamma SDef_m(\nu_3) = [\![\mathbf{m}]\!]_{\gamma}(\nu_3),$$

again contradicting the assumption that  $\mu_i$  minimizes  $[\![\mathbf{m}]\!]_{\gamma}$ . Thus, our supposition that  $\nu_1$  was distinct from  $\nu_2$  cannot hold, and so  $\lim_{\gamma \rightarrow 0} [\![\mathbf{m}]\!]_{\gamma}^*$  must be a singleton, as desired.  $\square$

Finally, Proposition 3.4 is a simple corollary of Lemma 3.10 and Theorem 3.3, as we now show.

**Proposition 3.4.**  $\llbracket m \rrbracket_{0^+}^* \in \llbracket m \rrbracket_0^*$ , so if  $m$  is consistent, then  $\llbracket m \rrbracket_{0^+}^* \in \{m\}$ .

*Proof.* By Theorem 3.3,  $\lim_{\gamma \rightarrow 0} \llbracket m \rrbracket_\gamma^*$  is a singleton. As in the body of the chapter, we refer to its unique element by  $\llbracket m \rrbracket_{0^+}^*$ . Lemma 3.10 therefore immediately gives us  $\llbracket m \rrbracket_{0^+}^* \in \llbracket m \rrbracket_0^*$ . If  $m$  is consistent, then by Proposition 3.1,  $OInc(m) = 0$ , so  $\llbracket m \rrbracket_0(\llbracket m \rrbracket_{0^+}^*) = 0$ , and thus  $\llbracket m \rrbracket_{0^+}^* \in \{m\}$ .  $\square$

### 3.A.2 Bayesian Networks as PDGs

In this section, we prove Theorem 3.5, starting with the details of how to transform a BN into a PDG. The transformation to a PDG with a directed hypergraph structure is immediate. Here, we give the formal details of how to compile it to a strict PDG, with only ordinary directed edges.

**Definition 3.12** (BN to strict PDG). Recall that a (quantitative) Bayesian Network  $(G, \mathbf{P})$  consists of two parts: its qualitative graphical structure  $G$ , described by a dag, and quantitative data  $\mathbf{P}$  which consists of a cpd  $P_i(X_i \mid \text{Pa}(X_i))$  for each variable  $X_i$ . Given a Bayesian network  $\mathcal{B}$  over variables  $X_1, \dots, X_n$ , we construct the corresponding *strict* unweighted PDG  $m_{\mathcal{B}}$  as follows: we take  $\mathcal{N} := \{X_1, \dots, X_n\} \cup \{\text{Pa}(X_1), \dots, \text{Pa}(X_n)\}$ . That is, the variables of  $m_{\mathcal{B}}$  consist of all the variables in  $\mathcal{B}$  together with a variable corresponding to the parents of  $X_i$ . These additional variables are used to emulate the hyperarcs. The values  $\mathcal{V}X_i$  for the original variables  $X_i$  are unchanged, and we set  $\mathcal{V}(\text{Pa}(X_i)) := \prod_{Y \in \text{Pa}(X_i)} \mathcal{V}(Y)$ . (If  $\text{Pa}(X_i) = \emptyset$ , so that  $X_i$  has no parents, then we then we identify  $\text{Pa}(X_i)$  with  $\mathbb{1}$  and take  $\mathcal{V}(\text{Pa}(X_i)) = \{\star\}$ , as discussed in Section 2.2.).

We take the set of arcs  $\mathcal{A} := \{(\text{Pa}(X_i) \rightarrow X_i) : i = 1, \dots, n\} \cup \{(\text{Pa}(X_i) \rightarrow Y) :$

$Y \in \text{Pa}(X_i)\}$  to be the set of arcs to a variable  $X_i$  from its parents, together with an arc from  $\text{Pa}(X_i)$  to each of the elements of  $\text{Pa}(X_i)$ , for  $i = 1, \dots, n$ . For the cpds, we set  $\mathbb{P}_{(\text{Pa}(X_i) \rightarrow X_i)}$  to be the cpd associated with  $X_i$  in  $\mathcal{B}$ . Finally, we give the standard projection  $X_j \in \text{Pa}(X_i)$  defined in [Construction 3.2](#) to the remaining arcs. So, if  $X_j \in \text{Pa}(X_i)$ , we set

$$\mathbb{P}_{(\text{Pa}(X_i) \rightarrow X_j)}(\dots, x_j, \dots) = \delta_{x_j};$$

that is,  $\mathbb{P}_{(\text{Pa}(X_i), X_j)}^{m_{\mathcal{B}}}$  is the cpd on  $X_j$  that, given a setting  $(\dots, x_j, \dots)$  of  $\text{Pa}(X_i)$ , yields the distribution that puts all mass on  $x_j$ .  $\square$

Consider a BN  $\mathcal{B}$  with variables  $\mathcal{X}X = \{X_1, \dots, X_n\}$ , and let  $\beta > 0$  be a positive vector over  $\mathcal{X}$ . Let  $(\mathcal{N}, \mathcal{V}, \mathcal{A}, \mathbb{P}) := m_{\mathcal{B}}$  be the corresponding strict (unweighted) PDG.  $\mathcal{N}$  contains  $\mathcal{X}$ , but also contains variables of the form  $\text{Pa}(X_i)$ , so it is a strict superset. However, there is a unique way to determine all of these other variables from  $\mathcal{X}$ , in a way that is compatible with the extra structural cpds in  $m_{\mathcal{B}}$ . Specifically, given a joint setting  $\omega \in \mathcal{V}\mathcal{X}$ , the variable  $\text{Pa}(X_i)$  must take on the restriction  $\omega[\text{Pa}(X_i)]$  of  $\omega$  to those variables. Let  $f : \mathcal{V}\mathcal{X} \rightarrow \mathcal{V}\mathcal{N}$  denote this function. Thus, it makes sense to identify distributions  $\mu \in \Delta\mathcal{V}\mathcal{N}$  of the form  $\mu(\mathcal{X})\delta f(\mathcal{N} \mid \mathcal{X})$ , and their marginals  $\mu(\mathcal{X}) \in \Delta\mathcal{V}\mathcal{X}$ . It is easy to see that any extended distribution  $\nu(\mathcal{N})$  that is not of this form will have infinite score, and so will not be relevant for the scoring function.

**Theorem 3.5.** *If  $\mathcal{B}$  is a Bayesian network and  $\Pr_{\mathcal{B}}$  is the distribution it specifies, then for all  $\gamma > 0$  and all vectors  $\beta$  such that  $\beta_a > 0$  for all arcs  $a$ ,  $[m_{\mathcal{B}, \beta}]_{\gamma}^* = \{\Pr_{\mathcal{B}}\}$ , and thus  $[m_{\mathcal{B}, \beta}]^* = \Pr_{\mathcal{B}}$ .*

*Proof.* For the cpd  $p(X_i \mid \text{Pa}(X_i))$  associated to a node  $X_i$  in  $\mathcal{B}$ , we have that  $\Pr_{\mathcal{B}}(X_i \mid \text{Pa}(X_i)) = p(X_i \mid \text{Pa}(X_i))$ . For all nodes  $X_i$  in  $\mathcal{B}$  and  $X_j \in \text{Pa}(X_i)$ , by

construction,  $\text{Pr}_{\mathcal{B}}$ , when viewed as a distribution on  $\mathcal{N}$ , is also with the cpd on the arc from  $\text{Pa}(X_i)$  to  $X_j$ . Thus,  $\text{Pr}_{\mathcal{B}}$  is consistent with all the cpds in  $\mathcal{M}_{\mathcal{B},\beta}$ ; so  $OInc_{\mathcal{B},\beta}(\text{Pr}_{\mathcal{B}}) = 0$ .

We next want to show that  $SDef_{\mathcal{B},\beta}(\mu) \geq 0$  for all distributions  $\mu$ . To do this, we first need some definitions. Let  $\rho$  be a permutation of  $1, \dots, n$ . Define an order  $\prec_\rho$  by taking  $j \prec_\rho i$  if  $j$  precedes  $i$  in the permutation; that is, if  $\rho^{-1}(j) < \rho^{-1}(i)$ . Say that a permutation is *compatible with  $\mathcal{B}$*  if  $X_j \in \text{Pa}(X_i)$  implies  $j \prec_\rho i$ . There is at least one permutation compatible with  $\mathcal{B}$ , since the graph underlying  $\mathcal{B}$  is acyclic.

Consider an arbitrary distribution  $\mu$  over the variables in  $\mathcal{X}$  (which we also view as a distribution over the variables in  $\mathcal{N}$ , as discussed above). Recall from [Definition 3.12](#) that the cpd on the arc in  $\mathcal{M}_{\mathcal{B},\beta}$  from  $\text{Pa}(X_i)$  to  $X_i$  is just the cpd associated with  $X_i$  in  $\mathcal{B}$ , while the cpd on the arc in  $\mathcal{M}_{\mathcal{B},\beta}$  from  $\text{Pa}(X_i)$  to  $X_j \in \text{Pa}(X_i)$  consists only of deterministic distributions (i.e., ones that put probability 1 on one element), which all have entropy 0. Thus,

$$\sum_{X \xrightarrow{\alpha} Y \in \mathcal{A}^{\mathcal{M}_{\mathcal{B}}}} H_\mu(Y | X) = \sum_{i=1}^n H_\mu(X_i | \text{Pa}(X_i)). \quad (3.8)$$

Given a permutation  $\rho$ , let  $\mathbf{X}_{\prec_\rho i} = \{X_j : j \prec_\rho i\}$ . Observe that

$$\begin{aligned} SDef_{\mathcal{B},\beta}(\mu) &= \left[ \sum_{X \xrightarrow{\alpha} Y \in \mathcal{A}^{\mathcal{M}_{\mathcal{B}}}} H_\mu(Y | X) \right] - H(\mu) \\ &= \sum_{i=1}^n H_\mu(X_i | \text{Pa}(X_i)) - \sum_{i=1}^n H_\mu(X_i | \mathbf{X}_{\prec_\rho i}) \quad [\text{by Fact 2.2 and (3.8)}] \\ &= \sum_{i=1}^n \left[ H_\mu(X_i | \text{Pa}(X_i)) - H_\mu(X_i | \mathbf{X}_{\prec_\rho i}) \right] \\ &= \sum_{i=1}^n I_\mu \left( X_i ; \mathbf{X}_{\prec_\rho i} \setminus \text{Pa}(X_i) \mid \text{Pa}(X_i) \right). \end{aligned}$$

[by ??]

Using ??, it now follows that, for all distributions  $\mu$ ,  $SDef_{m_B}(\mu) \geq 0$ . Furthermore, for all  $\mu$  and permutations  $\rho$ ,

$$SDef_{m_B}(\mu) = 0 \quad \text{iff} \quad \forall i. X_i \perp\!\!\!\perp_{\mu} \mathbf{X}_{\prec_{\rho} i}. \quad (3.9)$$

Since the left-hand side of (3.9) is independent of  $\rho$ , it follows that  $X_i$  is independent of  $\mathbf{X}_{\prec_{\rho} i}$  for some permutation  $\rho$  iff  $X_i$  is independent of  $\mathbf{X}_{\prec_{\rho} i}$  for every permutation  $\rho$ . Since there is a permutation compatible with  $\mathcal{B}$ , we get that  $SDef_{m_{\mathcal{B}, \beta}}(\Pr_{\mathcal{B}}) = 0$ . We have now shown that that  $SDef_{m_{\mathcal{B}, \beta}}$  and  $OInc$  are non-negative functions of  $\mu$ , and both are zero at  $\Pr_{0\mathcal{B}}$ . Thus, for all  $\gamma \geq 0$  and all vectors  $\beta$ , we have that  $\llbracket m_{\mathcal{B}, \beta} \rrbracket_{\gamma}(\Pr_{\mathcal{B}}) \leq \llbracket m_{\mathcal{B}, \beta} \rrbracket_{\gamma}(\mu)$  for all distributions  $\mu$ . We complete the proof by showing that if  $\mu \neq \Pr_{\mathcal{B}}$ , then  $\llbracket m_{\mathcal{B}, \beta} \rrbracket_{\gamma}(\mu) > 0$  for  $\gamma > 0$ .

So suppose that  $\mu \neq \Pr_{\mathcal{B}}$ . Then  $\mu$  must also match each cpd of  $\mathcal{B}$ , for otherwise  $OInc_{m_{\mathcal{B}, \beta}}(\mu) > 0$ , and we are done. Because  $\Pr_{\mathcal{B}}$  is the *unique* distribution that matches the both the cpds and independencies of  $\mathcal{B}$ ,  $\mu$  must not have all of the independencies of  $\mathcal{B}$ . Thus, some variable  $X_i$ ,  $X_i$  is not independent of some nondescendant  $X_j$  in  $\mathcal{B}$  with respect to  $\mu$ . There must be some permutation  $\rho$  of the variables in  $\mathcal{X}$  compatible with  $\mathcal{B}$  such that  $X_j \prec_{\rho} X_i$  (e.g., we can start with  $X_j$  and its ancestors, and then add the remaining variables appropriately). Thus, it is not the case that  $X_i$  is independent of  $X_{\prec_{\rho} i}$ , so by (3.9),  $SDef_{m_B}(\mu) > 0$ . This completes the proof.  $\square$

### 3.A.3 Factor Graph Proofs

Theorems 3.6(a) and (b) are immediate corollaries of their more general counterparts, Theorems 3.7 and 3.8, which we now prove.

**Theorem 3.8.** *For all unweighted PDGs  $(\mathcal{A}, \mathbb{P})$ , non-negative vectors  $\mathbf{v}$  of shape  $\mathcal{A}$ , and all  $\gamma > 0$ , we have that  $\llbracket(\mathbb{P}, \mathbf{v}, \gamma\mathbf{v})\rrbracket_\gamma = \gamma \cdot VFE_{(\Phi_{\mathbb{P}}, \mathbf{v})}$ ; consequently,  $\llbracket(\mathbb{P}, \mathbf{v}, \gamma\mathbf{v})\rrbracket_\gamma^* = \{\Pr_{(\Phi_{\mathbb{P}}, \mathbf{v})}\}$ . In other words, if  $\mathbf{m}$  is a PDG with  $\boldsymbol{\beta} \propto \boldsymbol{\alpha}$ , and  $\gamma$  is the constant of proportionality, then  $\llbracket\mathbf{m}\rrbracket_\gamma^*$  is a singleton containing the distribution of the factor graph  $\Phi_{\mathbb{P}}$  weighted by  $\boldsymbol{\alpha}$ .*

*Proof.* Let  $\mathbf{m} := (\mathcal{N}, \mathbf{v}, \gamma\mathbf{v})$  be the PDG in question. Explicitly,  $\alpha_a^m = v_a$  and  $\beta_a^m = \gamma v_a$  for all  $a \in \mathcal{A}$ . By Proposition 3.9,

$$\llbracket\mathbf{m}\rrbracket_\gamma(\mu) = \mathbb{E}_{\mathbf{w} \sim \mu} \left\{ \sum_{X \xrightarrow{a} Y} \left[ \beta_a \log \frac{1}{\mathbb{P}_a(y | x)} + (\alpha_a \gamma - \beta_a) \log \frac{1}{\mu(y | x)} \right] - \gamma \log \frac{1}{\mu(\mathbf{w})} \right\}.$$

Let  $\{\phi_a\}_{L \in \mathcal{A}} := \Phi_{\mathcal{N}}$  denote the factors of the factor graph associated with  $\mathbf{m}$ . Because we have  $\alpha_a \gamma = \beta_a$ , the middle term cancels, leaving us with

$$\begin{aligned} \llbracket\mathbf{m}\rrbracket_\gamma(\mu) &= \mathbb{E}_{\mathbf{w} \sim \mu} \left\{ \sum_{X \xrightarrow{a} Y} \left[ \beta_a \log \frac{1}{\mathbb{P}_a(y | x)} \right] - \gamma \log \frac{1}{\mu(\mathbf{w})} \right\} \\ &= \mathbb{E}_{\mathbf{w} \sim \mu} \left\{ \sum_{X \xrightarrow{a} Y} \left[ \gamma v_a \log \frac{1}{\phi(x, y)} \right] - \gamma \log \frac{1}{\mu(\mathbf{w})} \right\} \quad [\text{as } \beta_a = v_a \gamma] \\ &= \gamma \mathbb{E}_{\mathbf{w} \sim \mu} \left\{ \sum_{X \xrightarrow{a} Y} \left[ v_a \log \frac{1}{\phi(x, y)} \right] - \log \frac{1}{\mu(\mathbf{w})} \right\} \\ &= \gamma VFE_{(\Phi_{\mathcal{N}}, \mathbf{v})}. \end{aligned}$$

It immediately follows that the associated factor graph has  $\llbracket\mathbf{m}\rrbracket_\gamma^* = \{\Pr_{\Phi(\mathbf{m})}\}$ , because the free energy is clearly a constant plus the KL divergence from its associated probability distribution.  $\square$

**Theorem 3.7.** *For all WFGs  $\Psi = (\Phi, \theta)$  and all  $\gamma > 0$ , we have that  $VFE_\Psi = 1/\gamma \llbracket\mathbf{m}_{\Psi, \gamma}\rrbracket_\gamma + C$  for some constant  $C$ , so  $\Pr_\Psi$  is the unique element of  $\llbracket\mathbf{m}_{\Psi, \gamma}\rrbracket_\gamma^*$ .*

*Proof.* In  $\mathcal{M}_{\Psi,\gamma}$ , there is an arc  $1 \rightarrow X_J$  for every  $J \in \mathcal{J}$ , and also arcs  $X_J \twoheadrightarrow X_j$  for each  $X_j \in X_J$ . Because the cpds attached to the latter arcs are deterministic, a distribution  $\mu$  that is not consistent with one of the arcs, say  $X_J \twoheadrightarrow X_j$ , has  $OInc_m(\mu) = \infty$ . This is a property of relative entropy: if there exist  $j^* \in \mathcal{V}(X_j)$  and  $\mathbf{z}^* \in \mathcal{V}(J)$  such that  $\mathbf{z}_j^* \neq j^*$  and  $\mu$  places positive probability on their co-occurrence (i.e.,  $\mu(j^*, \mathbf{z}^*) > 0$ ), then we would have

$$\begin{aligned}\mathbb{E}_{\mathbf{z} \sim \mu_J} D\left(\mu(X_j \mid X_J = \mathbf{z}) \parallel \mathbb{1}[X_j = \mathbf{z}_j]\right) &= \sum_{\substack{\mathbf{z} \in \mathcal{V}(X_J), \\ \iota \in \mathcal{V}(X_j)}} \mu(\mathbf{z}, \iota) \log \frac{\mu(\iota \mid \mathbf{z})}{\mathbb{1}[\mathbf{z}_j = \iota]} \\ &\geq \mu(\mathbf{z}^*, j^*) \log \frac{\mu(j^* \mid \mathbf{z})}{\mathbb{1}[\mathbf{z}_j^* = j^*]} = \infty.\end{aligned}$$

Consequently, a distribution  $\mu$  that does not satisfy the the projections has  $\llbracket \mathcal{M}_{\Psi,\gamma} \rrbracket_\gamma(\mu) = \infty$  for every  $\gamma$ . Thus, a distribution that has a finite score must match the constraints, so we can identify such a distribution with its restriction to the original variables of  $\Phi$ . Moreover, for all distributions  $\mu$  with finite score and projections  $X_J \twoheadrightarrow X_j$ , the conditional entropy  $H(X_j \mid X_J) = -\mathbb{E}_\mu \log(\mu(x_j \mid x_J))$  and divergence from the constraints are both zero. Therefore the per-arc terms for both  $SDef_m$  and  $OInc_m$  can be safely ignored for the projections. Let  $\mathbb{P}_J$  be the normalized distribution  $\frac{1}{Z_J} \phi_J$  over  $X_J$ , where  $Z_J = \sum_{x_J} \phi_J(x_J)$  is the appropriate normalization constant. By [Definition 3.10](#), we have  $\mathcal{M}_{\Psi,\gamma} = (\mathcal{N}_\Phi, \theta, \gamma\theta)$ , so

by Proposition 3.9,

$$\begin{aligned}
\llbracket \mathbf{m}_{\Psi,\gamma} \rrbracket_\gamma(\mu) &= \mathbb{E}_{\mathbf{x} \sim \mu} \left\{ \sum_{J \in \mathcal{J}} \left[ \beta_J \log \frac{1}{\mathbb{P}_J(x_J)} + (\alpha_J \gamma - \beta_J) \log \frac{1}{\mu(x_J)} \right] - \gamma \log \frac{1}{\mu(\mathbf{x})} \right\} \\
&= \mathbb{E}_{\mathbf{x} \sim \mu} \left\{ \sum_{J \in \mathcal{J}} \left[ (\gamma \theta_J) \log \frac{1}{\mathbb{P}_J(x_J)} + (\theta_J \gamma - \gamma \theta_J) \log \frac{1}{\mu(x_J)} \right] - \gamma \log \frac{1}{\mu(\mathbf{x})} \right\} \\
&= \mathbb{E}_{\mathbf{x} \sim \mu} \left\{ \sum_{J \in \mathcal{J}} \left[ \gamma \theta_J \log \frac{1}{\mathbb{P}_J(x_J)} \right] - \gamma \log \frac{1}{\mu(\mathbf{x})} \right\} \\
&= \gamma \cdot \mathbb{E}_{\mathbf{x} \sim \mu} \left\{ \sum_{J \in \mathcal{J}} \theta_J \log \frac{Z_J}{\phi_J(x_J)} - \log \frac{1}{\mu(\mathbf{x})} \right\} \\
&= \gamma \cdot \mathbb{E}_{\mathbf{x} \sim \mu} \left\{ \sum_{J \in \mathcal{J}} \theta_J \left[ \log \frac{1}{\phi_J(x_J)} + \log Z_J \right] - \log \frac{1}{\mu(\mathbf{x})} \right\} \\
&= \gamma \cdot \mathbb{E}_{\mathbf{x} \sim \mu} \left\{ \sum_{J \in \mathcal{J}} \theta_J \log \frac{1}{\phi_J(x_J)} - \log \frac{1}{\mu(\mathbf{x})} \right\} + \sum_{J \in \mathcal{J}} \theta_J \log Z_J \\
&= \gamma VFE_\Psi + k \log \prod_J Z_J,
\end{aligned}$$

which differs from  $VFE_\Psi$  by the value  $\sum_J \theta_J \log Z_J$ , which is constant in  $\mu$ .  $\square$

## CHAPTER 4

### REPRESENTING THINGS WITH PDGS

PDGs are extremely expressive. We saw in [Sections 3.4.1 to 3.4.3](#) that PDGs can capture graphical models such as Bayesian Networks and Factor graphs—but this is only the beginning. In this chapter, we will see how a wide variety of other kinds of epistemic information can be captured with PDGs.

Probability is the dominant way that computer scientists and micro-economists think about epistemic state. This is due to standard betting arguments suggesting that any sufficiently rational agent (e.g., one resistant to dutch books) must act as if it had probabilistic beliefs ([Vineberg 2022](#); [Savage 1954](#)). This leads us to make a simple but conceptually important observation: a joint probability distribution is itself a special case of a PDG ([Section 4.1](#)).

In [Section 4.2](#), we develop a library of more intricate tools, which we call *widgets*, that can capture various modeling tools and fragments of epistemic information. Doing so will enable us to deliver on the promises made in [Section 3.2.1](#), of being able to use the current PDG formalism to express seeming generalizations of PDGs.

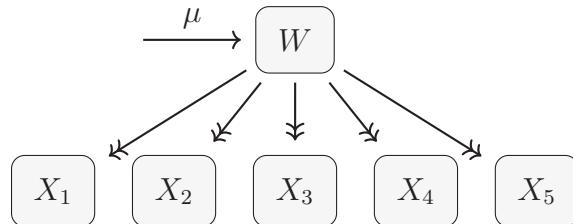
Probability, despite its dominance, does have shortcomings, and is not the only tool we have for representing knowledge and uncertainty ([Halpern 2017](#), §2). In [Section 4.3](#), we will use the tools developed in [Section 4.2](#) to show that PDGs capture some of these other well-established representations of uncertainty, most notably sets of probability measures, known as credal sets ([Walley 1991](#)) ([Section 4.3.1](#)) and Dempster-Shafer Belief functions ([Shafer 1976](#)) ([Section 4.3.2](#)). We also give a quick overview of other representations captured by PDGs that

will play larger roles later on: cluster (pseudo)marginals, causal models, and neural networks ([Sections 4.3.3 to 4.3.5](#)).

## 4.1 Probabilities and Random Variables

We start with an obvious construction: a joint distribution can be viewed as a very special case of a PDG. Let  $\mathcal{X}$  be a set of variables, and recall that  $\mathcal{V}\mathcal{X}$  is the set of all joint settings of the variables  $\mathcal{X}$ . A joint distribution  $\mu \in \Delta\mathcal{V}\mathcal{X}$  can be implicitly regarded as a PDG that has a single hyperarc  $\emptyset \rightarrow \mathcal{X}$ . We attach  $\mu$  as the (c)pd, and give it weights  $\alpha = 0$  (at least, if the distribution is intended to represent purely observational information), and  $\beta = 1$  (default confidence).

With this simple construction in mind, let's revisit the point made [Section 2.3.1](#) now that we can talk about PDGs. To a probability theorist, these joint distributions  $\mu$  may seem to be of a very special form, because they are over product spaces. In probability theory, the setup is typically instead that one has a (measurable) set  $\Omega$  of outcomes, and then random variables are in fact (*measurable*) functions  $X : \Omega \rightarrow \mathcal{V}\mathcal{X}$ . Observe that this too is an immediate special case of a PDG in which  $W$  is a variable with possible values  $\mathcal{V}W := \Omega$ :



(In fact, this PDG happens to also be a BN, if one isn't worried about calling  $W$ , itself typically a product of variables, a variable.) It is easy to see that these PDGs are consistent, and represent precisely the distribution  $\mu$  and the set of random

variables, no matter what  $\beta > 0$  we select.

Of course, the semantics of PDGs developed in Section 3.3 make heavy use of the usual definition of a joint distribution  $\mu \in \Delta \mathcal{V}\mathcal{X}$ , and so it would be circular to implicitly convert joint distributions to PDGs before developing the results of the previous chapter—but now that PDG semantics are on solid ground, we may freely regard a joint distribution as a PDG. What makes PDGs special is their capability to do this while simultaneously representing other things.

**Updating with PDGs combination.** We can regard a probability distribution  $\mu \in \Delta \Omega$  as a PDG  $\xrightarrow{\mu} [W]$  with  $\mathcal{V}W = \Omega$ . At the same time, we can view an observed event  $U : \Omega \rightarrow \{0, 1\}$  as a PDG  $[W] \xrightarrow{U} [U?] \xleftarrow{\delta_1}$  with  $\mathcal{V}(U?) = \{0, 1\}$ . What happens when we combine them, to form a new PDG  $\mathcal{M} := \mu + U?$  Because the conditional distribution  $\mu|U$  is the distribution that minimizes relative entropy from  $\mu$ , it follows that

- The unique element of  $[\mathcal{M}]_0^*$  is  $\mu|U$
- The observational inconsistency  $\langle\!\langle \mathcal{M} \rangle\!\rangle_0 = I_\mu[U] = -\log \mu(U)$  is the log probability of  $\mu$  (Proposition 6.2)

More generally, if our observation had been not an event (i.e., a deterministic distribution over a binary variable), but some other high-confidence distribution over a variable  $p(X)$ , we would have found that the resulting optimal distribution was the result of applying Jeffrey’s rule (Jeffrey 1968) to update  $\mu$  with  $p$ . This an immediate restatement of a standard result (Halpern 2017, pg. 109).

## 4.2 Widgets

PDGs may be expressive, but they are structured objects with clear and specific specific syntax. After some brief reflection, one might even find the syntax unnecessarily restrictive. Recall: in specifying the data for an arc  $X \rightarrow Y$ , one must specify a complete probability distribution  $p(Y|x)$  over the values of  $Y$  for *every* value  $x \in \mathcal{V}X$ . This appears to be a serious limitation. For instance, if  $X$  and  $Y$  are binary variables, one might want to annotate an arc  $X \rightarrow Y$  with data indicating that  $X \implies Y$  (i.e., if  $X = 1$  then  $Y = 1$ ). There are no problems in supplying a probability over  $Y$  when  $X=1$ , but unfortunately we also seem to be on the hook to provide a distribution over  $Y$  when  $X=0$ . To take another example, what if we do not want to supply the full probabilistic information, but only whether or not it is possible that  $Y=y$  given  $X=x$ ? Farther afield, what if we want to model soft constraints, or distances, or couplings between marginal distributions?

At first glance, it appears that modeling any of these concepts might require introducing generalizations of the PDG formalism. Yet it turns out that each of these concepts can be compiled to a small PDG that exactly captures it, which we call a *widget*.

### 4.2.1 Relations and Constraints

A *constraint* on a set  $\Omega$  of possible worlds is a subset  $C \subseteq \Omega$  of satisfying values, or equivalently, a function  $C : \Omega \rightarrow \{\text{true}, \text{false}\}$  assigning to each  $\omega \in \Omega$  a Boolean indicating whether or not it is allowed. A probability distribution  $\mu \in \Delta\Omega$  over a finite set  $\Omega$  encodes in particular a constraint on the possible

values of  $\Omega$  that can be observed—namely, that they must be among  $\text{Supp } \mu = \{\omega \in \Omega : \mu(\omega) > 0\} \subseteq \Omega$ . But is it possible to encode a constraint (in a PDG) without adding any additional probabilistic information? In a factor graph, the answer is yes: simply multiply by a factor that encodes the constraint  $C$ —a factor that happens to be equivalent to the uniform distribution over  $C$ . Unfortunately, this construction only works because factors lose their probabilistic interpretations in context. The same approach does not have the effect we are looking for in a PDG, because beyond articulating the constraint, it also involves specifying a belief that all elements of  $C$  were equally likely—information we do not have and would like to avoid assuming. In this section, we will develop a widget that will enable PDGs to represent precisely this information and nothing more.

Recall that, in a PDG, the sample space  $\Omega$  consists of joint settings of variables; a constraint on joint settings of variables is known more familiarly as a *relation* (see ??). We have already seen how PDGs can encode certain relations, such as in compiling a directed hypergraph to a graph (Construction 3.2), where we managed to enforce a constraint on the triple of variables  $(A, B, A \times B)$  ensuring that the value of  $A \times B$  is always the pair consisting of the value of  $A$  and the value of  $B$ . But this is a very special kind of relation—an equality relationship between variables that already appear to have a clear structural relationship. Can we encode arbitrary relational constraints with a PDG? It turns out we can.

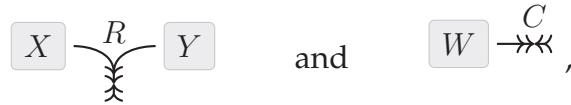
Consider a relation  $R = R(A_1, \dots, A_n)$  between the (values of) variables  $A_1, \dots, A_n$ . We associate  $R$  with a PDG

$$m_R := \begin{array}{c} A_1 \quad \cdots \quad A_n \\ \swarrow \quad \downarrow \quad \searrow \\ R \downarrow \\ \xrightarrow{\tau} T \end{array},$$

where  $T$  is a variable that intuitively represents “truth”; it technically can take

on two possible values:  $\mathcal{V}T := \{t, f\}$ , but it always takes on  $t$ . Note that the arc labeled “ $R$ ” is attached to the obvious way of regarding  $R$  as a conditional probability distribution, i.e., the cpd  $\delta R(T | A_1, \dots, A_n)$ .

We simplify notation by implicitly converting  $R$  to the PDG  $m_R$  whenever it would be helpful to regard it as one, thereby identifying  $R$  and  $m_R$ . Furthermore, we also adopt the graphical notation that leaves  $T$  implicit, so that the binary relation  $R(X, Y)$  and the constraint  $C \subseteq \mathcal{V}W$  can be depicted graphically as



respectively, in PDG notation. Intuitively, the variable  $T$  is uninteresting because it always takes the value  $t$ , so we may as well suppress it. So, visually, all we have done is to shrink the variable  $T$  to a single point, and suppress the label  $t$  on the event  $T=t$ ; the result is a pair of double-headed arcs joined head-to-head. Just as we have left  $T$  implicit in the graphical notation, we will want to identify a distribution  $\mu(\mathcal{X})$  with the “extended” distribution  $\mu(\mathcal{X})\delta(T=t)$ —except in the following theorem statement, where we treat it carefully.

While it is not surprising that this construction encodes a relation as far as PDG semantics are concerned, it is still worth verifying.

**Proposition 4.1.** *If  $R = R(A_1, \dots, A_n) = R(\mathbf{X})$  is a relational constraint on the variables  $\mathbf{X} = \{A_1, \dots, A_n\}$ , then, writing  $R$  for both the PDG  $m_R$  and the event  $R \subseteq \mathcal{V}\mathbf{X}$ , we have:*

[ link to proof ]

1.  $\{\{R\}\} = \{\mu(\mathbf{X}, T) : \text{Supp } \mu \subseteq R \times \{t\}\} = \{\mu(\mathbf{X}, T) : \mu(R) = \mu(T=t) = 1\}.$
2. Moreover, for all PDGs  $\mathbf{m}$  with variables  $\mathcal{X} \supseteq \mathbf{X}$ ,

- (a)  $\mu \in \{\mathcal{m} + R\}$  iff  $\mu(\mathcal{X}) \in \{\mathcal{m}\}$  and  $\mu(R) = 1$  (and also  $\mu(T=t) = 1$ ).
- (b) provided  $\langle\!\langle \mathcal{m} \rangle\!\rangle_\gamma < \infty$ ,  $\llbracket \mathcal{m} + R \rrbracket_\gamma^*$  consists entirely of  $\mu$  satisfying  $\mu(R) = 1$ .
- (c) for all  $\gamma \geq 0$ ,  $\langle\!\langle \mathcal{m} + R \rangle\!\rangle_\gamma \geq \langle\!\langle \mathcal{m} \rangle\!\rangle_\gamma$  with equality if and only if there exists some  $\mu \in \llbracket \mathcal{m} \rrbracket_\gamma^*$  for which  $\mu(R) = 1$ .

#### 4.2.2 “Soft” Constraints and Barriers

We have seen how “hard constraints”  $VX \rightarrow \{0, 1\}$  can be added to a PDG. But what about “soft” constraints, i.e., the gray area between zero and one? This is very different from specifying a probability over  $X$ ; we could assign every  $X$  a value of 0.99, or 0.01; there is no reason that the assignments should sum to 1. In light of the construction in the previous section, we now have an obvious candidate for how to do this: simply do precisely what we did for a hard constraint, except use a cpd  $c(T \mid X)$  that is not deterministic in place of  $\delta R(T \mid X)$ .

Given a set  $\mathbf{X}$  of variables, and a function  $s : V\mathbf{X} \rightarrow [0, 1]$ , let  $\mathcal{m}_s$  denote the PDG with variables  $\mathbf{X} \cup \{T\}$  and two cpds: the event  $T=t$  (as before), and the cpd  $s(T \mid X)$  with  $s(T=t \mid X=x) = s(x)$ . Semantically, the PDG  $\mathcal{m}_s$  has almost all of the properties we have seen in the case of hard constraints; the following is a strengthening of [Proposition 4.1](#).

**Proposition 4.2.** *Given a function  $s : V\mathbf{X} \rightarrow [0, 1]$ . we can regard  $s$  as a cpd  $s(T \mid X)$  with  $s(T=t \mid X=x) = s(x)$ . We then have the following analogues of the results for hard constraints:*

1.  $\{\mathcal{m}_s\} = \{\mu(\mathbf{X}, T) : \mu(s(X) = 1) = 1 = \mu(T=t)\}$ .

2. Moreover, for all PDGs  $\mathbf{m}$  with variables  $\mathcal{X} \supseteq \{X\}$ ,

- (a)  $\mu \in \{\mathbf{m} + s(X)\}$  iff  $\mu(\mathcal{X}) \in \{\mathbf{m}\}$  and  $\mu(s(X)) = 1$  (and  $\mu(T=t) = 1$ ).
- (b) provided  $\langle\!\langle \mathbf{m} \rangle\!\rangle_\gamma < \infty$ , the set  $\llbracket \mathbf{m} + \mathbf{m}_s \rrbracket_\gamma^*$  consists entirely of  $\mu$  satisfying  $\mu(s(X) > 0) = 1$ .
- (c) for all  $\gamma \geq 0$ ,  $\langle\!\langle \mathbf{m} + \mathbf{m}_s \rangle\!\rangle_\gamma \geq \langle\!\langle \mathbf{m} \rangle\!\rangle_\gamma$  with equality if and only if there exists some  $\mu \in \llbracket \mathbf{m} \rrbracket_\gamma^*$  for which  $\mu(s(X) = 1) = 1$ .

Despite the similarity so far, there is a big difference between hard and soft constraints: values of  $x$  that have intermediate satisfaction scores  $s(x) \in (0, 1)$ , while not fully consistent, are still possible. Indeed, they even be likely in the optimal distribution, when the context is inconsistent enough. To get a deeper understanding, it is helpful to look at the scoring function of  $\mathbf{m}_s$  in the context of another (arbitrary) PDG:

$$\llbracket \mathbf{m}_s + \mathbf{m} \rrbracket_\gamma(\mu) = \llbracket \mathbf{m} \rrbracket_\gamma(\mu) - \mathbb{E}_{\mu}[\log s(X)].$$

The key is the final term; adding  $\mathbf{m}_s$  to  $\mathbf{m}$  has had the effect of augmenting the scoring function  $\llbracket \mathbf{m} \rrbracket_\gamma$  with something called a *log barrier function*. Log barrier functions play an important role in solving constrained optimization problems with interior point methods. In fact, a similar function will ultimately be the key to inference in PDGs themselves, in [Chapter 8](#).

### 4.2.3 Couplings and Wasserstein Metrics

A *coupling* between distributions  $p(X)$  and  $q(Y)$  is a joint distribution over  $X$  and  $Y$  whose marginal on  $X$  is  $p$  and whose marginal on  $Y$  is  $q$ . Couplings are an important tool for probabilistic reasoning, and are the basis of optimal transport

theory (Santambrogio 2015), a wide range of programming logics (Kaminski et al. 2020). Formally, the set of couplings between  $p$  and  $q$  is defined as

$$\Pi(p(X), q(Y)) := \left\{ \mu \in \Delta \mathcal{V}(X, Y) : \quad \mu(X) = p, \mu(Y) = q \right\};$$

Observe that this is exactly the set of distributions consistent with a PDG containing  $p$  and  $q$ , i.e.,

$$\Pi(p, q) = \left\{ \begin{array}{c} p \downarrow \quad q \downarrow \\ \boxed{X} \quad \boxed{Y} \end{array} \right\}.$$

The couplings themselves are not all that a PDG can capture. Suppose we have a distance metric  $d$  on a space  $X$ . The *Wasserstein distance* (also called earth-mover's distance or Kantorovich metric) between  $p, q \in \Delta X$ , given by

$$W_1(p, q) := \inf_{\mu \in \Pi(p, q)} \mathbb{E}_{\mu} [d(X, Y)],$$

is a foundational quantity in optimal transport (). This definition effectively takes  $p(X)$  and  $q(X)$  with high confidence, by constraining to  $\mu \in \Pi(p, q)$ . But, in order to represent this as a PDG, we need to represent the  $d$  in probabilistic terms. A distance is not a probability, but we can convert it to a soft constraint and write it in probabilistic terms, as in Section 4.2.2. Accordingly, we have:

**Proposition 4.3.** *Suppose that  $p(X)$  and  $q(Y)$  are probability distributions. Given a distance measure  $d : \mathcal{V}(X, Y) \rightarrow [0, \infty]$ , let  $\hat{d}(x, y) := \exp(-d(x, y))$  be a similarity measure. Interpreting  $\hat{d}$  as a soft constraint, we have:*

$$\langle\!\langle m_{\hat{d}} + p! + q! \rangle\!\rangle = \left\langle\!\left\langle \begin{array}{c} p! \downarrow \quad q! \downarrow \\ \boxed{X} \quad \boxed{Y} \\ \hat{d} \\ \xrightarrow{\text{t}} \boxed{T} \end{array} \right\rangle\!\right\rangle = \inf_{\mu \in \Pi(p, q)} \mathbb{E}_{\mu} [d(X, Y)] = W_1(p, q),$$

where the exclamation points indicate high confidence ( $\beta = \infty$ ).

*Proof.* The high confidence specification of  $p$  and  $q$  constrains the optimal distributions to  $\Pi(p, q)$ , and any distribution generates its expected distance, as a consequence of the results of Section 4.2.2. □

eqn #

#### 4.2.4 Incomplete CPDs and Individual (Conditional) Probabilities

A cpd between discrete variables can be represented by a stochastic matrix (i.e., a matrix whose rows sum to one). It turns out that it is possible to use the machinery of PDGs to, effectively, give only some of the values in that matrix. In PDG notation, we abbreviate the widget that implements  $p(Y|X)$  with some missing values by adding an empty circle at the tail, i.e., drawing  $X \circ \rightarrow Y$ .

this lets us capture implication

We now present an important extreme case of that construction: how, for any  $p \in [0, 1]$ , we can construct a PDG that represents the belief that  $\Pr(Y=y|X=x) = p$ , but say nothing about how the probability splits between other values of  $y$ , and also says nothing about the probability of  $Y$  if  $X \neq x$ .

First, we introduce two new auxiliary variables. The first variable, which we might like to call “ $Y=y$ ”, but mostly refer to as  $Y_y$  to prevent confusion with the synonymous event, is a binary variable, with  $\mathcal{V}(Y_y) = \{y, \neg y\}$ , and takes the value  $y$  if  $Y = y$ , and  $\neg y$  if  $Y \neq y$ . The second variable, which we would like to call “ $X=x||Y=y$ ”, but instead mostly refer to as  $X_x Y_y$  to prevent notational confusion, can take three values:  $\mathcal{V}(X_x Y_y) := \{x, y, \neg y\}$ . The value  $x$  is meant to correspond exactly to the event  $X=x$ , much like before, so that  $X_x Y_y = x$  if and only if  $X = x$ . The values  $y$  and  $\neg y$  also correspond to their respective events, but more loosely; the variable  $X_x Y_y$  only takes one of these values when  $X \neq x$ . Note that both variables can be determined from  $X$  and  $Y$  (although we will

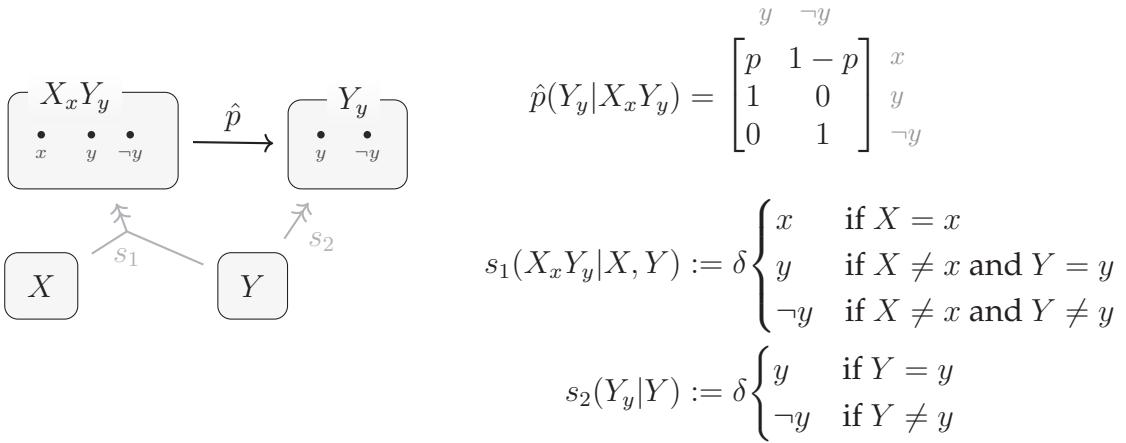


Figure 4.1: A widget PDG for capturing a single conditional probability: a statement of the form  $\Pr(Y=y | X=x) = p$ , for  $p \in [0, 1]$ .

need to enforce this with additional arcs), and therefore there is a unique way to extend a distribution over  $X$  and  $Y$  to also include the variables  $Y_y$  and  $X_xY_y$ .

With these definitions in place, there is now an obvious way to add an arc from the variable  $(X_xY_y)$  to the variable  $Y_y$ , together with a cpd asserting that  $\Pr(Y=y|X=x) = p$ . This cpd is written as a stochastic matrix  $\hat{p}$  defined on the right of Figure 4.1. The PDG we have just constructed is illustrated on the left of Figure 4.1. In addition to  $\hat{p}$  and the new variables, this PDG includes the structural constraints  $s_1$  and  $s_2$  needed to define the variables  $X_xY_y$  and  $Y_y$  in terms of  $X$  and  $Y$ ; they are deterministic functions, drawn in double-headed gray arrows.

So, when we add  $\Pr(Y = y|X = x) = p$  to a PDG  $m$ , what we really mean is: first convert construct a widget as above, and add that structure (i.e., the new variables  $X_xY_y$  and  $Y_y$ , their definitions  $s_1$  and  $s_2$ , and the cpd  $\hat{p}$ ) to  $m$ .

In what sense does this “work”? The first order of business is to prove that it behaves as we should expect, semantically. This means showing that the widget

contains precisely the information that  $\Pr(Y=y|X=x) = p$ , and nothing more. We now explore two ways of making this precise with the semantics of PDGs.

**Proposition 4.4.** *If  $\mathcal{M}$  is a PDG, then  $\mu \in \{\mathcal{M}\}$  with  $\mu(Y=y|X=x) = p$  if and only if  $\mu$  extends (via  $s_1, s_2$ ) to  $\bar{\mu} \in \{\mathcal{M} + \Pr(Y=y|X=x) = p\}$ .*

*Proof.* ( $\implies$ ) Suppose  $\mu \in \{\mathcal{M}\}$  has  $\mu(y|x) = p$ , and let  $\bar{\mu}$  be its extenson via the functions  $s_1$  and  $s_2$  to the variables  $X_x Y_y$  and  $Y_y$ .  $s_1$  ensures that  $X_x Y_y = x$  precisely when  $X = x$ . By assumption, the probability (according to  $\mu$ ) that  $Y=y$  when  $X=x$  is  $p$ . But  $s_2$  ensures that  $Y=y$  precisely when  $Y_y=y$ . Thus,  $\bar{\mu}(Y_y=y | X_x Y_y=x) = \mu(Y=y | X=x) = p$ . Thus the first row of  $\hat{p}$  When  $X_x Y_y \neq x$ , on the other hand, then  $s_1$  and  $s_2$  ensure that  $X_x Y_y = Y_y = Y$ , thereby satisfying the other rows of the cpd  $\hat{p}$ . Thus,  $\bar{\mu}$  satisfies all cpds of  $\mathcal{M}$ , in addition to  $s_1, s_2$ , and  $\hat{p}$ , and therefore  $\bar{\mu} \in \{\mathcal{M} + \Pr(Y=y|X=x) = p\}$ .

( $\impliedby$ ) Choose some  $\bar{\mu} \in \{\mathcal{M} + \Pr(Y=y|X=x) = p\}$ , and let  $\mu$  be its marginal on the variables of  $\mathcal{M}$ . By the logic above,  $\bar{\mu}(Y_y=y | X_x Y_y=x) = \mu(Y=y | X=x)$ . Since  $\bar{\mu}$  satisfies the cpd  $\hat{p}$ , we know that  $\bar{\mu}(Y_y=y | X_x Y_y=x) = \mu(Y=y | X=x) = p$  as desired. Since  $\bar{\mu}$  also satisfies all the cpds of  $\mathcal{M}$ , we also have  $\mu \in \{\mathcal{M}\}$ .  $\square$

We have shown that the effect of our widget on a PDG's set-of-distributions semantics is precisely to restrict to distributions  $\mu$  in which  $\mu(y|x) = p$ . But this result is vacuous for inconsistent PDGs, whose set-of-distributions semantics is empty. We now give a result of similar character for the inconsistency semantics of a PDG, which shows our construction behaves appropriately for all PDGs.

**Proposition 4.5.** *Suppose  $\mathcal{M}$  is a PDG with variables  $\mathcal{X}$  and  $\beta \geq 0$ . Then, for all*

$X, Y \subseteq \mathcal{X}$ ,  $x \in \mathcal{V}X$ ,  $y \in \mathcal{V}Y$ ,  $p \in [0, 1]$  and  $\gamma \geq 0$ , we have that:

$$\langle\!\langle m + \Pr(Y=y|X=x) = p \rangle\!\rangle_\gamma \geq \langle\!\langle m \rangle\!\rangle_\gamma,$$

with equality if and only if there exists  $\mu \in \llbracket m \rrbracket_\gamma^*$  such that  $\mu(Y=y|X=x) = p$ . (Note that this condition is trivially satisfied when  $\mu(X=x) = 0$ .)

*Proof.* The inequality is immediate; it is an instance of monotonicity of inconsistency Lemma 6.1, which we will discuss in depth in ?? Intuitively: believing more cannot make you any less inconsistent. We now prove that equality holds iff there is a minimizer with the appropriate conditional probability.

( $\Leftarrow$ ). Suppose there is some  $\mu \in \llbracket m \rrbracket_\gamma^*$  with  $\mu(Y=y|X=x) = p$ . Because  $\mu \in \llbracket m \rrbracket_\gamma^*$ , we know that  $\llbracket m \rrbracket_\gamma(\mu) = \langle\!\langle m \rangle\!\rangle$ . Let  $\hat{\mu}$  be the extension of  $\mu$  to the new variables “ $X=x|Y=y$ ” and “ $Y=y$ ”, whose values are functions of  $X$  and  $Y$  according to  $s_1$  and  $s_2$ . Then,

$$\begin{aligned} & \langle\!\langle m + \Pr(Y=y|X=x) = p \rangle\!\rangle_\gamma \\ & \leq \llbracket m + \Pr(Y=y|X=x) = p \rrbracket_\gamma(\hat{\mu}) \\ & = \llbracket m \rrbracket_\gamma(\mu) + \mathbb{E}_\mu \left[ \log \frac{\hat{\mu}(Y_y|X_xY_y)}{\hat{p}(Y_y|X_xY_y)} \right] \\ & = \llbracket m \rrbracket_\gamma(\mu) + \mu(X=x, Y=y) \log \frac{\mu(Y=y|X=x)}{p} \\ & \quad + \mu(X=x, Y \neq y) \log \frac{\mu(Y \neq y|X=x)}{1-p} \\ & = \llbracket m \rrbracket_\gamma(\mu) + \mu(X=x, Y=y) \log(1) + \mu(X=x, Y \neq y) \log(1) \\ & = \llbracket m \rrbracket_\gamma(\mu) = \langle\!\langle m \rangle\!\rangle_\gamma. \end{aligned}$$

The equality between the third and fourth lines is perhaps the trickiest to see, but follows because for joint settings in which  $X \neq x$ , one can easily see that  $\hat{\mu}(Y_y|X_xY_y)$  equals 1 with probability 1, as does  $\hat{p}(Y_y|X_xY_y)$ . So, after dividing

one by the other and taking a logarithm, these cases contribute nothing to the expectation. What remains are the two possibilities where  $X=x$ , which are shown in the second line. To complete this direction of the proof, it suffices to observe that we already knew the inequality held in the opposite direction (by monotonicity), so the two terms are equal.

( $\implies$ ). Suppose the two inconsistencies are equal, i.e.,

$$\langle\!\langle \mathbf{m} + \Pr(Y=y|X=x) = p \rangle\!\rangle_\gamma = \langle\!\langle \mathbf{m} \rangle\!\rangle_\gamma.$$

This time, choose  $\hat{\mu} \in [\![\mathbf{m} + \Pr(Y=y|X=x) = p]\!]_\gamma^*$ , and define  $\mu$  to be its marginal on the variables of  $\mathbf{m}$  (which contains the same information as  $\hat{\mu}$  itself). Let  $q := \mu(Y=y|X=x)$ . Then,

$$\begin{aligned} \langle\!\langle \mathbf{m} \rangle\!\rangle_\gamma &= \langle\!\langle \mathbf{m} + \Pr(Y=y|X=x) = p \rangle\!\rangle_\gamma \\ &= [\![\mathbf{m} + \Pr(Y=y|X=x) = p]\!]_\gamma(\hat{\mu}) \\ &= [\![\mathbf{m}]\!]_\gamma(\mu) + \mu(X=x, Y=y) \log \frac{\mu(Y=y|X=x)}{p} \\ &\quad + \mu(X=x, Y \neq y) \log \frac{\mu(Y \neq y|X=x)}{1-p} \\ &= [\![\mathbf{m}]\!]_\gamma(\mu) + \mu(X=x) \left[ q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p} \right] \\ &= [\![\mathbf{m}]\!]_\gamma(\mu) + \mu(X=x) D(q \parallel p) \\ &\geq \langle\!\langle \mathbf{m} \rangle\!\rangle_\gamma + \mu(X=x) D(q \parallel p) \end{aligned}$$

Therefore  $0 \geq \mu(X=x) D(q \parallel p)$ . But relative entropy is non-negative, by Gibbs inequality. This shows  $\mu(X=x) D(q \parallel p) = 0$ . So either  $\mu(X=x)$ , or  $p = \mu(Y=y|X=x)$ , and the first case is just a special case of the second one. In addition, the algebra above shows that  $\mu \in [\![\mathbf{m}]\!]_\gamma^*$ , as its score is  $\langle\!\langle \mathbf{m} \rangle\!\rangle_\gamma$ . Thus, we have found  $\mu \in [\![\mathbf{m}]\!]_\gamma^*$  such that  $\mu(Y=y|X=x) = p$ , completing the proof.  $\square$

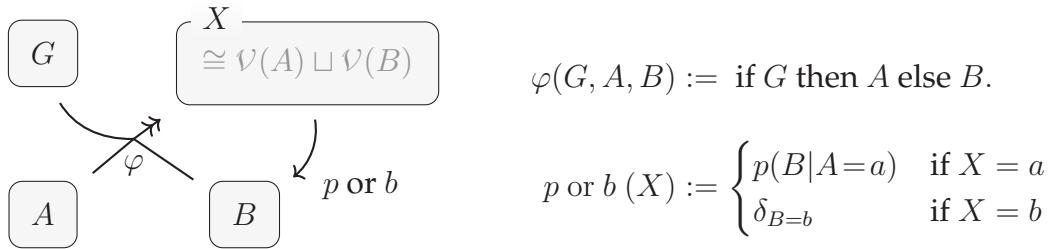
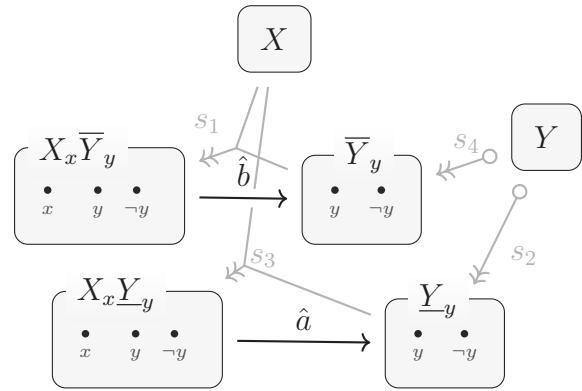


Figure 4.2: A widget implementing a CPD whose presence is conditioned on a guard variable, i.e.,  $p(B|A, G=1)$

**“Conditional Conditional” Probabilities.** We now turn to another, closely related modeling task. What if we wanted to specify a full cpd  $p(B|A)$ , but with the stipulation that it was only meaningful when a third “guard variable”  $G$  takes a certain value (say,  $G=1$ )? This can be a useful thing to model. For instance, it exactly describes a common practice for reducing errors in machine learning, by identifying domain shift: train a classifier (whose output is  $G$ ) to detect if the original predictor  $p(B|A)$  is relevant to the current context, and only apply  $p$  if it is (i.e.,  $G=1$ ). A second, more abstract use for this primitive is to enrich PDGs with another notion of confidence. By adding a guard variable  $G_a$  for each  $a \in \mathcal{A}$ , and asserting that each  $G_a = 1$  occurs with probability  $1 - \epsilon_a$  (for some  $\epsilon_a \in [0, 1]$ ), it becomes possible to articulate a different and more standard kind of probability-based confidence in the reliability of each cpd. (Observe that the special case in which each  $\epsilon_a = 0$  is the definition of a PDG we already have.) We discuss the difference between approaches like this and the meaning of our primitive  $\beta$  in [Chapters 11](#) and [12](#).

Technically speaking, the specification of  $p(B|A)$  under the condition that  $G = 1$  is really no different from a cpd  $p(B|A, G=1)$ . Given that we have already proved that PDGs can represent arbitrary conditional probabilities, it is entirely expected that PDGs can represent this too. A widget for it is illustrated in [Figure 4.2](#).



$$\hat{b}(\bar{Y}_y | X_x \bar{Y}_y) = \hat{a}(Y_y | X_x Y_y) = \begin{bmatrix} y & \neg y \\ a & 1-a \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{array}{c} x \\ y \\ \neg y \end{array},$$

$$s_2(Y_y | Y) := \begin{cases} \delta_y & \text{if } Y = y \\ (\text{null}) & \text{if } Y \neq y \end{cases} \quad \text{and} \quad s_4(\bar{Y}_y | Y) := \begin{cases} (\text{null}) & \text{if } Y = y \\ \delta_{\neg y} & \text{if } Y \neq y \end{cases}$$

where `(null)` indicates no information (see [Section 4.2.4](#)),

$$\text{and } s_1(X_x \bar{Y}_y | X, \bar{Y}_y) = s_3(X_x Y_y | X, Y_y) = \delta \begin{cases} x & \text{if } X = x \\ y & \text{if } X \neq x \text{ and } Y_y = y \\ \neg y & \text{if } X \neq x \text{ and } Y \neq y \end{cases} \quad \text{as in Figure 4.1.}$$

*Figure 4.3: A widget PDG for capturing a conditional probability range, a statement of the form  $\Pr(Y=y | X=x) \in [a, b]$ .*

#### 4.2.5 Probability Ranges

Probability requires the modeler to assign a specific number to every event  $U \subseteq \Omega$ . This can be restrictive; what if you simply don't know, and so are not prepared to assign a probability to  $U$ ? One heavily studied line of work aimed at addressing this issue allows the modeler to instead specify a range of values rather than a single one; thus one can say "the probability of  $U$  lies between  $\frac{1}{3}$  and  $\frac{1}{2}$ ", for example. It is easy to see that specifying an actual probability distribution is the special case where the upper and lower bounds are equal, but in general,

this approach yields a far richer theory ([Walley 1991](#)).

At first, it appears that allowing a modeler to specify beliefs this way would require us to define a generalized PDG in which the modeler specifies not a cpd  $p(Y|X)$ , but rather upper and lower bounds on it. But yet again, it turns out to be possible with the definition we already have.

There are multiple ways to do this. One idea to specify  $\Pr(Y=y \mid X=x) \in [a, b]$  with a PDG is to add an “overapproximation” variable  $\bar{Y}_y$  that is guaranteed to take on the value  $y$  whenever  $Y$  does (but may also take on the value  $y$  other times), as well as an “underapproximation” variable  $\underline{Y}_y$  that can only take on the value  $y$  when  $Y$  does (but even then, may not). Once these are defined (e.g., with constraints), it is not too difficult to encode the upper and lower probabilities as cpds that target these new variables. A (perhaps clunky) way of fully formalizing this is fleshed out in [Figure 4.3](#).

### 4.3 Other Representations of Knowledge and Uncertainty

Probability is the dominant way of talking about epistemic uncertainty in economics and computer science, but it is far from the only one ([Halpern 2017](#)). Probability—at least, in the form of a joint distribution  $\mu$ —has some shortcomings. One of the biggest is its inability to represent ignorance well (a point we will return to in [Chapter 11](#)). This has lead many to develop generalizations of probability that are ([Shafer and Shenoy 1990](#); [Walley 1991](#)). It turns out that several of the most celebrated generalizations of probabilities can be viewed as special cases of PDGs.

### 4.3.1 Convex Sets of Probabilities

We have seen that  $\{\mathcal{m}\}$  represents a convex set of probabilities. But does this work the other way around? What if we already have a convex set  $\mathcal{P}$  of distributions over  $\Omega$  that we would like to model with a PDG? There is actually a very straightforward way of doing this. To simplify matters, suppose that  $\mathcal{P}$  is a *polytope*, meaning that it has finitely many extreme points, which we call vertices. Let  $V$  be a **random** variable whose possible values  $\mathcal{V}$  are vertices of  $\mathcal{P}$ , and define a conditional probability distribution  $p(\Omega \mid V=v) := v$ , which typechecks because a vertex  $v$  is a distribution over  $\Omega$ .

$$\mathcal{m}_{\mathcal{P}} := \boxed{V} \xrightarrow{p} \boxed{W}$$

**Proposition 4.6.**  $\mu \in \mathcal{P}$  iff it can be extended to some  $\bar{\mu} \in \{\mathcal{m}_{\mathcal{P}}\}$ . That is,  $\mathcal{P} = \{\mu(W) : \mu \in \{\mathcal{m}_{\mathcal{P}}\}\}$ .

Interestingly, this approach also resolves some confusion about maximum entropy. Let's look at an example. [Return to Example 2.? !!](#)

**Example 4.1.** Suppose we are about to flip a coin; but we know that it is either fair or double-headed, but do not know which. One way of modeling this is might be to take  $\Omega = \{H, T\}$  to be the possible outcomes of the coin flip, and represent our belief with the set of distributions  $\mathcal{P} = \{\mu_{.5}, \delta_H\}$ , where  $\mu_{.5}$  assigns probability 1/2 to both  $H$  and  $T$ , and  $\delta_H$  assigns probability 1 to  $H$ . From a set of distributions consistent with observations, the principle of maximum entropy tells us to select the one with the highest entropy—which is a fair coin. This seems quite extreme; the possibility of the double headed coin has made no difference to the maximum-entropy distribution!

Part of the problem, arguably, is that we haven't actually modeled the whole situation—we've left out an important part of the picture: our knowledge about the bias of the coin, and how that affects the outcome. This information is very easy to add to a PDG:

$$\begin{array}{c} \text{Bias} \\ \longrightarrow \\ \text{Coin} \end{array} \quad p = \begin{bmatrix} H & T \\ .5 & .5 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} F \\ H^2 \end{bmatrix}$$

Now, in the maximum entropy distribution, the coin is twice as likely to be fair as it is to be one-sided, and overall it has a  $2/3$  chance of landing heads. This is a more reasonable summary of our belief state—we do not truly believe that the coin is fair. Furthermore, if we select  $\alpha = 1$  for this edge, then the distribution with minimal structural deficiency (the analogue of maximum-entropy when there is causal information present) prescribes the distribution that says the coin has a  $3/4$  chance of landing heads, which is in the very middle of our range.  $\triangle$

It is well-known that different representations can change the result of maximizing entropy. But some representations better capture a scenario more naturally than others, and PDGs facilitate the choice of representations that interact well with entropy. This is because, generally speaking, PDGs force you to make your modeling choices in an explicit, uniform way (a point we will emphasize repeatedly in [Chapter 6](#)). In [Example 4.1](#), the PDG representation lead us to a space which, upon applying the principle of maximum entropy (or better yet, of minimum information deficiency) gives an answer that appears eminantly more reasonable than the simplest model we found without the PDG. Indeed, it appears impossible to obtain a representation that leads 50/50 maximum entropy presumption when using a PDG. On its own [Example 4.1](#) is far from is

just one (more) piece of supporting evidence, adding it to a pile that includes [Theorem 3.5](#).

The construction in this section is a generic way of leaving out information in a probabilistic model. At a technical level, it amounts to exploiting the fact that a cpd can be viewed as imposing a barycentric coordinate system for a convex set of distributions (as laid out at the end of [Section 2.3.1](#)). In the next section, we investigate another representation of uncertainty with more structure, which can be captured with PDGs in a rather different way.

### 4.3.2 Belief and Plausibility Functions

We now move on to another representation of uncertainty, which generalizes the notion of a probability distribution over a (for simplicity, finite) set  $\Omega$ , called a *belief function* ([Shafer and Shenoy 1990](#)). Like a probability measure, a belief function  $Bel$  assigns a degree of belief in  $[0, 1]$  to subsets  $U \subseteq \Omega$ . Belief functions must satisfy certain axioms ensuring that  $Bel(U) + Bel(\bar{U}) \leq 1$ , and thus  $Plaus(U) := 1 - Bel(\bar{U}) \geq Bel(U)$ . It can be shown that a probability distribution is the special case when these two relationships hold with equality, so that  $Bel = Plaus$ .

Belief functions admit an alternate representation in terms of a *mass function*  $m : 2^\Omega \rightarrow [0, 1]$ , which yields belief and plausibility functions according to

$$Bel_m(U) := \sum_{V \subseteq U} m(V) \quad \text{and} \quad Plaus_m(U) := \sum_{\substack{V \subseteq \Omega \\ V \cap \bar{U} \neq \emptyset}} m(V).$$

Moreover, the correspondence is unique. That is, there are 1-1 correspondence between belief functions  $Bel$ , plausibility functions  $Plaus$ , and mass functions  $m$

([Halpern 2017](#), Thm 2.6.3). The only requirements on  $m$  are that:

$$m(\emptyset) = 0 \quad \text{and} \quad \sum_{A \subseteq \Omega} m(A) = 1.$$

So, in other words,  $m$  is a probability over non-empty subsets  $V \subseteq \Omega$ . There is also a natural relation between values of  $V$  (i.e., subsets of  $\Omega$ ) and values of  $W$  (i.e., elements of  $\Omega$ ): containment ( $\ni$ ). Both  $m$  and  $\ni$  can be modeled with a PDG. What happens if we put them together? Intuitively, this PDG describes a situation in which a subset  $V \subseteq \Omega$  is drawn according to  $m$ , and then  $\omega \in V$  is picked non-deterministically.

**Definition 4.1.** If  $m$  is the mass function representing the belief function  $Bel$  and the plausibility function  $Plaus$ , then we associate all of these objects with the same PDG,

$$m_m, m_{Bel}, m_{Plaus} := \xrightarrow{m} V \ni W,$$

where  $W$  is a variable taking values in  $\mathcal{V}W := \Omega$ , and  $V$  is a variable whose possible values  $\mathcal{V}(V) := 2^\Omega$  are subsets of  $\Omega$ .  $\square$

Move Mm  
and M\_{\ni}  
into defn +  
add  
intermediate  
eqn over  
yellow bit

For those who are not yet fluent with the constraint notation,  $m_m$  is just the PDG containing the mass function  $m$  in the form of a distribution over subsets of  $\Omega$ , as well as a constraint that the actual world  $W$  is a member of whatever subset is chosen by  $m$ .

**Theorem 4.7.** If  $m$  is the mass function the belief function  $Bel$  and the plausibility function  $Plaus$ , and  $M := m_{Bel} = m_{Plaus}$ , then:

- (a) A distribution  $\mu \in \Delta\Omega$  is the marginal of an extended distribution  $\bar{\mu} \in \{M\} \subseteq \Delta(2^\Omega \times \Omega)$  if and only if  $Bel(U) \leq \mu(U) \leq Plaus(U)$  for all  $U \subseteq \Omega$ .

[ link to proof ]

(b) For all  $U \subseteq \Omega$ ,  $\left\langle \xrightarrow{m} V \dashv \exists W \xrightarrow{\in U} \right\rangle = -\log \text{Plaus}(U)$ . That is, the inconsistency of simultaneously having belief function  $m_{Bel}$  and also that the event  $U$  occurs, is the analogue of the information content but for plausibility.

This theorem shows that PDGs of this form coincide semantically with Dempster-Shafer belief functions. Part (a) states that the set-of-distribution semantics is precisely the set-of-distribution semantics for  $Bel$ . Indeed, if we follow the conventions of [Halpern \(2017\)](#), Theorem 2.6.1) and define  $\mathcal{P}_{Bel} := \{\mu \in \Delta\Omega : Bel(U) \leq \mu(U) \leq \text{Plaus}(U) \text{ for all } U \subseteq \Omega\}$ ,<sup>1</sup> then we have shown that  $\mu$  is compatible with  $m_{Bel}$  if and only if is an element of  $\mathcal{P}_{Bel}$ .

Meanwhile, part (b) states that something quite intuitive, but might benefit from being unpacked somewhat. The PDG on the left-hand side is just  $m_{Bel}$  plus a constraint representing the event  $U$ . How inconsistent is it to have belief function  $Bel$  and also believe  $U$  occurs with certainty? It's the analogue of the information content  $I_\mu[U] = -\log \mu(U)$ , but with plausibility instead of probability. Indeed, if  $\text{Plaus}(U) = 1$ , then there is no inconsistency to also believing  $U$  occurs, and if  $\text{Plaus}(U) = 0$ , then it's so implausible so as to be infinitely inconsistent.

Now for the proof of the theorem.

*Proof. (a)* The forwards direction is easy. Fix some  $\bar{\mu} \in \{m\}$ , let  $\mu \in \Delta\Omega$  be its marginal on  $W$ . Select an arbitrary  $U \subseteq \Omega$ . Keep in mind that  $\bar{\mu}$  is a joint distribution over pairs  $(A, \omega)$  that satisfy  $\omega \in A \subseteq \Omega$ , whose marginal on  $A$  is distributed according to  $m$ . For every such pair  $(A, \omega) \in \text{Supp } \bar{\mu}$ , if  $A \subseteq U$  then

---

<sup>1</sup>For observant readers who followed the reference: it is not hard to see that the upper and lower bounds are equivalent; [Halpern \(2017\)](#) uses only the lower bound, but we give both for symmetry.

clearly  $\omega \in U$ . Thus,  $Bel(U) = \sum_{A \subseteq U} m(A) \leq \sum_{\omega \in U} \mu(W=\omega) = \mu(U)$ . Similarly, if  $\omega \in U$ , then it must be that  $A \cap U \neq \emptyset$ , so  $\mu(U) \leq \sum_{A: A \cap U \neq \emptyset} m(A) = Plaus(U)$ .

The reverse direction is far more subtle than it appears; here we present a relatively compact proof of it based on Hall's Marriage Theorem ([Hall 1935](#)). Suppose that  $\mu \in \Delta\Omega$  satisfies  $Bel(U) \leq \mu(U) \leq Plaus(U)$  for all  $U \subseteq \Omega$ . Assume for simplicity that the possible outputs of  $m$  and  $\mu$  are rational numbers, and let  $N$  be their common denominator. We now construct a bipartite graph  $G = (L, R, \mathcal{E})$  with  $|L| = |R| = N$  vertices in each part. The left part  $L = \sqcup_{\omega \in \Omega} L_\omega$  is partitioned to be in bijection with the elements of  $\Omega$ , with  $|L_\omega| = \mu(\omega) \cdot N$ . Symmetrically,  $R = \sqcup_{A \in \Omega} R_A$ , where each  $R_A$  consists of  $|R_A| = m(A) \cdot N$  vertices. For notational convenience, for  $u \in L$ , let  $\omega_u$  be the element of  $\Omega$  corresponding to the partition containing  $u$ ; symmetrically, let  $A_v$  be the subset associated with the partition of  $v \in R$ . Finally, let  $\mathcal{E} := \{(u, v) \in L \times R : \omega_u \in A_v\}$ . Observe that a perfect matching  $E \subseteq \mathcal{E}$  in the graph  $G$  amounts to a coupling between the marginal distributions  $\mu \in \Delta\Omega$  and  $m \in \Delta(2^\Omega)$ , supported only on  $(\omega, A)$  such that  $\omega \in A$ . To be fully precise, that coupling would be a joint distribution  $\bar{\mu}_E(\omega, A) := \frac{1}{N} |\{(u, v) \in E : \omega_u \in A_v\}|$  that is both an element of  $\{\mathcal{M}_{Bel}\}$  and an extension of  $\mu$ . What remains is only to show that there exists such a perfect matching, for which we turn to Hall's Marriage criterion.

Hall's Marriage Theorem states that there exists a perfect matching for  $G$  if and only if,  $|T| \leq |\partial_G(T)|$  for all  $T \subseteq L$ , where  $\partial_G(T) = \{v \in R : \exists u \in T. \omega_u \in A_v\}$  is the set of vertices connected to  $T$  by an edge. We now prove that this is the case. Choose any subset  $T \subseteq L$  of vertices, and define  $U_T := \{\omega \in \Omega : T \cap L_\omega \neq \emptyset\}$  to be the elements of  $\Omega$  represented by some vertex in  $T$ . On one hand,

$$|T| \leq |\bigcup_{u \in T} L_{\omega_u}| = \sum_{\substack{\omega \in \Omega \\ L_\omega \cap T \neq \emptyset}} = \mu(U_T).$$

On the other hand,

$$\begin{aligned} |\partial_G(T)| &= |\{v \in R : \exists u \in T. \omega_u \in A_v\}| \\ &= \sum_{\substack{A \subseteq \Omega \\ \exists u \in T. \omega_u \in A}} m(A) = \sum_{\substack{A \subseteq \Omega \\ U_T \cap A = \emptyset}} m(A) = \text{Plaus}(A). \end{aligned}$$

But recall that, by assumption,  $\mu(U_T) \leq \text{Plaus}(U_T)$ . This allows us to chain together the previous seven (in)equalities, thereby proving that  $|T| \leq |\partial_G(T)|$  as desired. By Hall's criterion, this means there must exist a perfect matching, and we have already seen this means  $\mu$  can be extended to some  $\bar{\mu} \in \{\mathcal{M}_{\text{Bel}}\}$ , as desired.

**(b)** Let  $\mathcal{P} := \{\bar{\mu} \in \Delta(2^\Omega \times \Omega) : \bar{\mu}(W \in U) = \bar{\mu}(W \in V) = 1\}$  be the set of all distributions satisfying the hard constraints of the PDG in question. Let  $\mathcal{A}_U := \{A \subseteq \Omega : A \cap U \neq \emptyset\}$  be the collection of subsets of  $\Omega$  with non-empty intersection with  $U$ . It is not too difficult to see that the marginal projection of  $\mathcal{P}$  onto the variable  $V$  is the same as the set  $\Delta\mathcal{A}_U$  of distributions supported on sets with non-empty intersection with  $U$ . With these facts in mind, we calculate

$$\begin{aligned} \left\langle \xrightarrow{m} V \begin{array}{c} \ni \\[-1ex] \exists \\[-1ex] \ni \end{array} W \xrightarrow{\in U} \right\rangle &= \inf_{\bar{\mu} \in \mathcal{P}} D(\mu(V) \parallel m) \\ &= \inf_{\nu \in \Delta\{\mathcal{A}_U : A \cap U \neq \emptyset\}} \mathbb{E}_{A \sim \nu} \left[ \log \frac{\nu(A)}{m(A)} \right], \end{aligned}$$

since the objective on first line does not depend on the marginal on  $W$ , except through the constraint. For those who are not yet used to the notation, recall that the PDG in question is just  $\mathcal{M}_m$  together with a constraint that the value of the variable  $W$  lies in the given set  $U$ . In turn, this inconsistency is equal to

$$\inf_{\nu \in \Delta\mathcal{A}_U} \sum_{A \in \mathcal{A}_U} \nu(A) \log \frac{\nu(A)}{m(A)} \geq \left( \sum_{A \in \mathcal{A}_U} \nu(A) \right) \log \frac{\sum_{A \in \mathcal{A}_U} \nu(A)}{\sum_{A \in \mathcal{A}_U} m(A)},$$

by the log-sum inequality (Cover and Thomas 1991). Moreover, the log-sum inequality states that this holds with equality if and only if there is some constant

$k$  so that  $\nu(A) = km(A)$  for all  $A \in \mathcal{A}_U$ . By definition,  $\text{Supp } \nu \subseteq \mathcal{A}_U$ , and thus  $\sum_{A \in \mathcal{A}_U} \nu(A) = 1$ . Therefore the inconsistency we have been calculating is equal to

$$-\log \sum_{A \in \mathcal{A}_U} m(A) = -\log \text{Plaus}(U). \quad \square$$

### 4.3.3 Pseudomarginals on Cluster Graphs

Many inference algorithms for graphical models make use of a data-structure that tracks various marginal distributions, which may not be consistent (Wainwright et al. 2003, 2008; Koller and Friedman 2009), sometimes called a *pseudomarginal* (Wainwright et al. 2008) or a *cluster graph (with associated data)* (Koller and Friedman 2009); we will call them *cluster marginals*.

Concretely, given a set of variables  $\mathcal{X}$ , a cluster graph is graph, whose node set is a collection  $\mathcal{C}$  of clusters, each  $C \in \mathcal{C}$  of which is a subset of  $\mathcal{X}$ . A cluster marginal  $\mu = \{\mu_C(C)\}_{C \in \mathcal{C}}$  consists of a distribution  $\mu_C \in \Delta^{VC}$  over the variables in each cluster  $C \in \mathcal{C}$ . Because  $\mu$  is just a collection of joint distributions, it is exactly what is needed to supply observational data for a hypergraph  $\mathcal{A} = \{\emptyset \rightarrow C\}_{C \in \mathcal{C}}$ . Let  $m_\mu$  denote this PDG.

A cluster marginal is said to be *calibrated* if, for every edge  $C-D$  in the cluster graph, the distributions  $\mu_C$  and  $\mu_D$  agree the marginals of their common variables, which is a necessary condition for  $m_\mu$  to be consistent. Belief propagation is a way of calibrating these objects, and effectively is a way of reducing local inconsistencies between pairs of clusters. We will return to this in Chapter 7.

When the cluster graph has a special property that makes it a *tree decomposition* (see Section 8.2 or “clique tree” in (Koller and Friedman 2009)), then local

consistency is equivalent to global consistency (Wainwright et al. 2008), and so a cluster marginal  $\mu$  of this kind is calibrated if and only if  $\{m_\mu\} \neq \emptyset$ . We will cover these concepts in detail when we use them to an inference algorithm for PDGs in Chapter 8; Here, we aim only to make an orthogonal point: that these structures themselves are PDGs.

#### 4.3.4 Causal Models

A *structural equations model* (SEM) (Pearl 2009) is a collection of equations that explain how each “endogenous variable” gets its value, using a mechanism that can depend on other variables in the model. In other words, a SEM is a collection of (deterministic) functions, each taking as input joint settings of variables, and outputting a single value. Since deterministic functions are special cases of cpds, there is a natural way to regard a causal model as a PDG, and moreover, the confidence weights do not impact the semantics when the cpd is deterministic. Furthermore, the usual way of adding probability to a causal model is no different from adding a probability distribution to the corresponding PDG. The semantics of the resulting PDG exactly capture the behavior of causal model in the absence of intervention.

We have not yet given an analogue of intervention for a general PDG, although it’s straightforward to encode interventions by altering the functions in the usual way (i.e., replacing the causal equations with constants). We will make all of this precise in Section 5.3. More broadly, Chapter 5 develops a framework that allows us to say something much deeper about the relationship between causal models and a concept called *QIM-compatibility* (the subject of Chapter 5), a notion closely related to (and captured by) the qualitative information in a PDG.

### 4.3.5 Implicit Neural Representations

A more modern way to represent knowledge is implicitly, by means of the parameters of a neural network. But neural networks are often viewed as conditional probability distributions, and hence can be directly included as the quantitative data for an arc of a PDG. So too can the other components of machine learning systems such as data, priors, and observations. Indeed, as we will see in [Chapters 6 and 7](#), the PDG formalism has a great deal to say about modern neural representations as well.

Visually, it is uncommon to use the notation of graphical models to describe the architecture of neural networks; instead, the standard is to represent operations and layers of a neural network as nodes, and their connections as edges—like a circuit diagram. In many cases, the duals of these diagrams, appropriately annotated with the network parameters, are literally PDGs.

⟨ TODO: diagram here would be nice touch! ⟩

## APPENDICES FOR CHAPTER 4

### 4.A Proofs

**Proposition 4.1.** *If  $R = R(A_1, \dots, A_n) = R(\mathbf{X})$  is a relational constraint on the variables  $\mathbf{X} = \{A_1, \dots, A_n\}$ , then, writing  $R$  for both the PDG  $m_R$  and the event  $R \subseteq \mathcal{V}\mathbf{X}$ , we have:*

1.  $\{\{R\}\} = \{\mu(\mathbf{X}, T) : \text{Supp } \mu \subseteq R \times \{t\}\} = \{\mu(\mathbf{X}, T) : \mu(R) = \mu(T=t) = 1\}.$
2. Moreover, for all PDGs  $m$  with variables  $\mathcal{X} \supseteq \mathbf{X}$ ,
  - (a)  $\mu \in \{\{m + R\}\}$  iff  $\mu(\mathcal{X}) \in \{\{m\}\}$  and  $\mu(R) = 1$  (and also  $\mu(T=t) = 1$ ).
  - (b) provided  $\langle\!\langle m \rangle\!\rangle_\gamma < \infty$ ,  $\llbracket m + R \rrbracket_\gamma^*$  consists entirely of  $\mu$  satisfying  $\mu(R) = 1$ .
  - (c) for all  $\gamma \geq 0$ ,  $\langle\!\langle m + R \rangle\!\rangle_\gamma \geq \langle\!\langle m \rangle\!\rangle_\gamma$  with equality if and only if there exists some  $\mu \in \llbracket m \rrbracket_\gamma^*$  for which  $\mu(R) = 1$ .

*Proof.* 1. If  $\mu \in \{\{R\}\}$ , then  $\mu(T=t) = 1$ , so if  $\mu(\mathbf{X}=x) > 0$ , then  $\mu(T=t \mid \mathbf{X}=x) = 1$  for all joint settings  $x \in \mathcal{V}(\mathbf{X})$ . But  $\mu(T=t \mid \mathbf{X}=x) = R(x)$ , and thus  $R(x) = 1$ . Conversely, if  $\mu(\mathbf{X}, T) = \mu(\mathbf{X})\delta(T=t)$  and  $\mu(R) = 1$ , then

$$\mu(T \mid \mathbf{X}) = (\mu(\mathbf{X})\delta(T=t))/\mu(\mathbf{X}) = \delta R(T \mid \mathbf{X}).$$

Thus  $\mu$  is consistent with all conditional probabilities in the widget  $m_R$ , and so  $\mu \in \{\{m_R\}\} = \{\{R\}\}$ .

2. (a) If  $\mu \in \{\{m + R\}\}$ , then by definition  $\mu(T=t) = 1$ . By the same reasoning as in part 1, we also find that  $\mu(R) = 1$ . We also know that  $\mu$  satisfies the cpds of  $m$ , so  $\mu(\mathcal{X}) \in \{\{m\}\}$ . Conversely, if  $\mu(\mathcal{X}) \in \{\{m\}\}$ , and  $\mu(T=t) = \mu(R) = 1$ , then,

again by the logic in part 1,  $\mu(T \mid X) = \delta R$ . Therefore  $\mu$  satisfies all cpds of  $m$ , and all cpds of the widget  $m_R$ , and thus  $\mu \in \{m + R\}$ .

(b) For all  $\mu \in [m]_\gamma^*$ , we have

$$\begin{aligned} \langle m \rangle_\gamma &= [m]_\gamma^*(\mu) \\ &\geq \mathbb{E}_\mu \left[ \log \frac{\mu(T)}{\mathbb{1}[T=t]} + \log \frac{\mu(T \mid X)}{\delta R(T \mid x)} \right] \\ &= \infty \cdot \mathbb{1}[\mu(T=t) = 1] + \infty \cdot \mathbb{E}_{x \sim \mu} [\mathbf{x} \notin R \Rightarrow \mu(T=t \mid x) = 0]. \end{aligned}$$

Now, suppose that  $\mu(R) < 1$ , meaning there is some mass on joint settings  $x \notin R$ . If  $\mu(T=t \mid x) \neq 0$ , then the score is infinite, by the second term. Yet if  $\mu(T=t \mid x) = 0$  and there is some mass on  $x$ , then  $\mu(T=t) < 1$ , and so again the score is infinite, by the first term. Thus, if the score is finite, then  $\mu(R) = 1$ .

(c) The inequality is monotonicity (Lemma 6.1). Suppose there is some  $\mu \in [m]_\gamma^*$  with  $\mu(R) = 1$ . Define  $\bar{\mu}(\mathcal{X}, T) := \mu(\mathcal{X})\delta(T=t)$ . By the same logic as in the proof of part 1 of the proposition,  $\mu$  satisfies all of the cpds of the widget  $m_R$ ; and thus  $[m + R](\bar{\mu}) = [m]_\gamma(\mu)$ . Chaining together some inequalities, we find:

$$\langle m + R \rangle_\gamma \leq [m + R](\bar{\mu}) = [m]_\gamma(\mu) = \langle m \rangle_\gamma \leq \langle m + R \rangle_\gamma,$$

and thus all of these quantities are equal.

Conversely, suppose that  $\langle m + R \rangle_\gamma = \langle m \rangle_\gamma$ . If  $[m] = \infty$ , then  $[m]_\gamma^*$  consists of all distributions, and so the statement is vacuous. On the other hand, select some  $\bar{\mu} \in [m + R]_\gamma^*$ . By applying part (b), we find that  $\bar{\mu}(R) = 1$ . And, because  $\langle m + R \rangle_\gamma = \langle m \rangle_\gamma$  and  $\bar{\mu}$  is an optimal distribution for  $m + R$ , we find that its marginal  $\mu(\mathcal{X}) := \bar{\mu}(\mathcal{X})$  is optimal for  $m$ :

$$[m]_\gamma(\mu(\mathcal{X})) = \langle m + R \rangle_\gamma = \langle m \rangle_\gamma \leq [m]_\gamma(\mu(\mathcal{X}))$$

Thus, we have constructed  $\mu \in [m]_\gamma^*$  satisfying  $\mu(R) = 1$ .  $\square$

## CHAPTER 5

### QUALITATIVE MECAHNISM INDEPENDENCE

In Section 3.3.1, we defined what it meant for a joint distribution  $\mu$  to be *quantitatively* compatible with the information in a PDG—it must match all of the cpds. But conspicuously absent is a qualitative analogue. What should it mean for a distribution to be compatible with the qualitative structure of a PDG, i.e., with a hypergraph?

In this chapter, we define what it means for a joint probability distribution to be compatible with a set of independent causal mechanisms, at a qualitative level—or, more precisely, with a directed hypergraph  $\mathcal{A}$ , i.e., the qualitative structure of a PDG. When  $\mathcal{A}$  represents a qualitative Bayesian network (BN), this notion of *QIM-compatibility* with  $\mathcal{A}$  reduces to satisfying the appropriate conditional independencies. But giving semantics to hypergraphs using QIM-compatibility lets us do much more. For one thing, we can capture functional *dependencies*. For another, we can capture important aspects of causality using compatibility: we can use compatibility to understand cyclic causal graphs, and to demonstrate structural compatibility, we must essentially produce a causal model. Finally, compatibility has deep connections to Shannon information. Applying compatibility to cyclic structures helps to clarify a longstanding conceptual issue in information theory. Compatibility also has a close, but far from obvious, relationship with the original scoring-function semantics for qualitative PDGs, which underlies many of our results.

## 5.1 Introduction

The structure of a probabilistic graphical model encodes a set of conditional independencies among variables. This is useful because it enables a compact description of probability distributions that have those independencies; it also lets us use graphs as a visual language for describing important qualitative properties of a probabilistic world. Yet these kinds of independencies are not the only important qualitative aspects of a probability measure. In this [paper](#), we study a natural generalization of standard graphical model structures that can describe far more than conditional independence.

For example, another qualitative aspect of a probability distribution is that of functional *dependence*, which is also exploited across computer science to enable compact representations and simplify probabilistic analysis. Acyclic causal models, for instance, specify a distribution via a probability over *contexts* (the values of variables whose causes are viewed as outside the model), and a collection of equations (i.e., functional dependencies) ([Pearl 2000](#)). Deep learning is all about learning functional dependencies between variables; to take just one example *normalizing flows* ([Tabak and Vanden-Eijnden 2010](#); [Kobyzev et al. 2021](#)) specify a distribution by composing a fixed distribution over some latent space with a function (i.e., a functional dependence) fit to observational data. Similarly, complexity theorists often regard a probabilistic Turing machine as a deterministic function that takes as input a uniformly random string ([Sipser 2006](#)). Functional dependence and independence are deeply related and interacting notions. For instance, if  $B$  is a function of  $A$  (written  $A \twoheadrightarrow B$ ) and  $A$  is independent of  $C$  (written  $A \perp\!\!\!\perp C$ ), then  $B$  and  $C$  are also independent ( $B \perp\!\!\!\perp C$ ).<sup>1</sup> Moreover,

---

<sup>1</sup>This well-known fact ([Lemma 5.10](#)) is formalized and proved in [Section 5.A](#), where all proofs

dependence can be written in terms of independence:  $Y$  is a function of  $X$  if and only if  $Y$  is conditionally independent of itself given  $X$  (i.e.,  $X \twoheadrightarrow Y$  iff  $Y \perp\!\!\!\perp Y | X$ ). Traditional graph-based languages such as Bayesian Networks (BNs) and Markov Random Fields (MRFs) cannot capture these relationships. Indeed, the graphoid axioms (which describe BNs and MRFs) (Pearl and Paz 1987) and axioms for conditional independence (Naumov and Nicholls 2013), do not even consider statements like  $A \perp\!\!\!\perp A$  to be syntactically valid. Yet such statements are perfectly meaningful, and reflect a deep relationship between independence, dependence, and generalizations of both notions (grounded in information theory, a point we will soon revisit).

This chapter describes a simple yet expressive graphical language for describing qualitative structure such as dependence and independence in probability distributions. The idea behind our approach is to specify the inputs and outputs of a set of *independent mechanisms*. In slightly more detail, by “independent mechanism”, we mean a process by which some (set of) the target variables  $T$  are determined as a (possibly randomized) function of a (set of) source variables  $S$ . So, at a qualitative level, the modeler specifies not a graph, but rather a *directed hypergraph*—which, as we have already seen, is the structure of a PDG.

Although some qualitative aspects of PDGs were characterized using a scoring function (Section 3.3.2), that scoring function does not obviously correspond to an analogue of our set-of-distributions semantics  $\{\!\!\{ - \}\!\!\}$  for quantitative information (Section 3.3.1). Part of the problem is that it is not even clear what the analogue of a BN’s independencies should be for a cyclic graph, let alone for an arbitrary directed hypergraph. In this chapter, we develop precisely such a

---

can be found.

notion. More precisely, we define what it means for a distribution to be *QIM-compatible* (qualitatively independent-mechanism compatible, or just *compatible* when unambiguous) with a directed hypergraph  $\mathcal{A}$ . This definition allows us to use directed hypergraphs as a language for specifying structure in probability distributions, of which the semantics of qualitative BNs are a special case ([Theorem 5.1](#)).

But QIM-compatibility can do much more than represent conditional independencies in acyclic networks. For one thing, it can encode arbitrary functional dependencies ([Theorem 5.2](#)); for another, it gives meaningful semantics to cyclic models. Indeed, compatibility lets us go well beyond capturing dependence and independence. The fact that [Pearl \(2000\)](#) also views causal models as representing independent mechanisms suggests that there might be a connection between causality and QIM-compatibility. In fact, there is.

A *witness* that a distribution  $\mu$  is compatible with a hypergraph  $\mathcal{A}$  is an extended distribution  $\bar{\mu}$  that is nearly equivalent to (and guarantees the existence of) a causal model that explains  $\mu$  with dependency structure  $\mathcal{A}$ . As we shall see, thinking in terms of witnesses and compatibility allows us to tie together causality, dependence, and independence.

Perhaps surprisingly, compatibility also has deep connections with information theory ([Section 5.4](#)). The conditional independencies of a BN can be viewed as a very specific kind of information-theoretic constraint. Our notion of compatibility with a hypergraph  $\mathcal{A}$  turns out to imply a generalization of this constraint (closely related to the qualitative PDG scoring function) that is meaningful for all hypergraphs. Applied to cyclic models, it yields a causally inspired notion of pairwise interaction that clarifies some important misunderstandings

in information theory ([Examples 5.5](#) and [5.6](#)).

Saying that one approach to qualitative graphical modeling has connections to so many different notions is a rather bold claim. We spend the rest of the chapter justifying it.

## 5.2 Qualitative Independent-Mechanism (QIM) Compatibility

In this section, we present the central definition of the chapter: a way of making precise Pearl's notion of "independent mechanisms", used to motivate Bayesian Networks from a causal perspective. [Pearl \(2009, p.22\)](#) states that "*each parent-child relationship in a causal Bayesian network represents a stable and autonomous physical mechanism.*" But, technically speaking, a parent-child relationship only partially describes the mechanism. Instead, the autonomous mechanism that determines the child is really represented by that child's joint relationship with all its parents. So, the qualitative aspect of a mechanism is best represented as a directed *hyperarc* that can have multiple sources.

Recall that a directed hypergraph ([Definition 2.5](#)) consists of a set  $\mathcal{N}$  of nodes and a set  $\mathcal{A}$  of hyperarcs, each  $a \in \mathcal{A}$  of which is associated with a set  $S_a \subseteq \mathcal{N}$  of source nodes and a set  $T_a \subseteq \mathcal{N}$  of target nodes. For the rest of [Chapter 5](#), we will call  $(\mathcal{N}, \mathcal{A}) = \mathcal{A}$  simply a *hypergraph*, since all our hypergraphs will be directed.

As before, we are interested in hypergraphs whose nodes represent variables, so that each  $X \in \mathcal{N}$  will ultimately be associated with a set  $\mathcal{V}X$  of possible values. However, one should not think of  $\mathcal{V}$  as part of the information carried by the hypergraph. (Indeed,  $\mathcal{V}$  and  $(\mathcal{N}, \mathcal{A})$  are separate components of a PDG.) It makes

perfect sense to say that  $X$  and  $Y$  are independent without specifying the possible values of  $X$  and  $Y$ . Of course, when we talk concretely about a distribution  $\mu$  on a set of variables  $\mathcal{X} \cong (\mathcal{N}, \mathcal{V})$ , those variables must have possible values—but the *qualitative* properties of  $\mu$ , such as independence, can be expressed purely in terms of  $\mathcal{N}$ , without reference to  $\mathcal{V}$ .

Intuitively, we expect a joint distribution  $\mu(\mathcal{X})$  to be qualitatively compatible with a set of independent mechanisms (whose structure is given by a hypergraph  $\mathcal{A}$ ) if there is a mechanistic explanation of how each target arises as a function of the variable(s) on which it depends and independent random noise. This is made precise by the following definition.

**Definition 5.1** (QIM-compatibility). Let  $\mathcal{X}$  and  $\mathcal{Y}$  be (possibly identical) sets of variables, and  $\mathcal{A} = \{S_a \xrightarrow{a} T_a\}_{a \in \mathcal{A}}$  be a hypergraph with nodes  $\mathcal{X}$ . We say a distribution  $\mu(\mathcal{Y})$  is *qualitatively independent-mechanism compatible*, or (QIM-)compatible, with  $\mathcal{A}$  (symbolically:  $\mu \models \Diamond \mathcal{A}$ ) iff there exists an extended distribution  $\bar{\mu}(\mathcal{Y} \cup \mathcal{X} \cup \mathcal{U}_\mathcal{A})$  of  $\mu(\mathcal{Y})$  to  $\mathcal{X}$  and to  $\mathcal{U}_\mathcal{A} = \{U_a\}_{a \in \mathcal{A}}$ , an additional set of “noise” variables (one variable per hyperarc) according to which:

- (a) the variables  $\mathcal{Y}$  are distributed according to  $\mu$  (i.e.,  $\bar{\mu}(\mathcal{Y}) = \mu(\mathcal{Y})$ ),
- (b) the variables  $\mathcal{U}_\mathcal{A}$  are mutually independent (i.e.,  $\bar{\mu}(\mathcal{U}_\mathcal{A}) = \prod_{a \in \mathcal{A}} \bar{\mu}(U_a)$ ),
- (c) and the target variable(s)  $T_a$  of each hyperarc  $a \in \mathcal{A}$   
are determined by  $U_a$  and the source variable(s)  $S_a$  (i.e.,  $\forall a \in \mathcal{A}. \bar{\mu} \models (S_a, U_a) \rightarrow\!\!\! \rightarrow T_a$ ).

We call such a distribution  $\bar{\mu}(\mathcal{X} \cup \mathcal{Y} \cup \mathcal{U}_\mathcal{A})$  a *witness* to the fact that  $\mu$  is QIM-compatible with  $\mathcal{A}$ . □

While Definition 5.1 requires the noise variables  $\{U_a\}_{a \in \mathcal{A}}$  to be independent

of one another, note that they need not be independent of any variables in  $\mathcal{X}$ . In particular,  $U_a$  may not be independent of  $S_a$ , and so the situation can diverge from what one would expect from a randomized algorithm, whose randomness  $U$  is assumed to be independent of its input  $S$ . Furthermore, the variables in  $\mathcal{U}$  may not be independent of one another conditional on the value of some  $X \in \mathcal{X}$ .

**Example 5.1.**  $\mu(X, Y)$  is compatible with  $\mathcal{A} = \{\emptyset \xrightarrow{1} \{X\}, \emptyset \xrightarrow{2} \{Y\}\}$  (depicted in PDG notation as  $\rightarrow[X] [Y] \leftarrow$ ) iff  $X$  and  $Y$  are independent, i.e.,  $\mu(X, Y) = \mu(X)\mu(Y)$ . For if  $U_1$  and  $U_2$  are independent and respectively determine  $X$  and  $Y$ , then  $X$  and  $Y$  must also be independent.  $\triangle$

This is a simple illustration of a more general phenomenon: when  $\mathcal{A}$  describes the structure of a Bayesian Network (BN), then QIM-compatibility with  $\mathcal{A}$  coincides with satisfying the independencies of that BN (which are given, equivalently, by the *ordered Markov properties* (Lauritzen et al. 1990), factoring as a product of probability tables, or *d-separation* (Geiger et al. 1990)). To state the general result (Theorem 5.1), we must first clarify how the graphs of standard graphical and causal models give rise to directed hypergraphs.

Suppose that  $G = (V, E)$  is a graph, whose edges may be directed or undirected. Given a vertex  $u \in V$ , write  $\text{Pa}_G(u) := \{v : (v, u) \in E\}$  for the set of vertices that can “influence”  $u$ . There is a natural way to interpret the graph  $G$  as giving rise to a set of mechanisms: one for each variable  $u$ , which determines the value of  $u$  based the values of the variables on which  $u$  can depend. Formally, let  $\mathcal{A}_G := \{ \text{Pa}_G(u) \xrightarrow{u} \{u\} \}_{u \in V}$  be the hypergraph *corresponding* to the graph  $G$ .

**Theorem 5.1.** *If  $G$  is a directed acyclic graph and  $\mathcal{I}(G)$  consists of the independencies of its corresponding Bayesian network, then  $\mu \models \Diamond \mathcal{A}_G$  if and only if  $\mu$  satisfies  $\mathcal{I}(G)$ .*

[ link to proof ]

[Theorem 5.1](#) shows, for hypergraphs that correspond to directed acyclic graphs (dags), our definition of compatibility reduces exactly to the well-understood independencies of BNs. This means that QIM-compatibility, a notion based on the independence of causal mechanisms, and seemingly unrelated to other notions of independence in BNs, gives us a completely different way of characterizing these independencies—one that can be generalized to much larger classes of graphical models, that includes, for example, cyclic variants ([Baier et al. 2022](#)). Moreover, QIM-compatibility can capture properties other than independence. As the following example shows, it can capture determinism.

**Example 5.2.** If  $\mathcal{A} = \{\overset{1}{\rightarrow} X, \overset{2}{\rightarrow} X\}$  consists of just two hyperarcs pointing to a single variable  $X$ , then a distribution  $\mu(X)$  is QIM-compatible with  $\mathcal{A}$  iff  $\mu$  places all mass on a single value  $x \in \mathcal{V}(X)$ .  $\triangle$

Intuitively, if two independent coins always give the same answer (the value of  $X$ ), then neither coin can be random. This simple example shows that we can capture determinism with multiple hyperarcs pointing to the same variable. Such hypergraphs do not correspond to graphs; recall that in a BN, two arrows pointing to  $X$  (e.g.,  $Y \rightarrow X$  and  $Z \rightarrow X$ ) represent a single mechanism by which  $X$  is jointly determined (by  $Y$  and  $Z$ ), rather than two distinct mechanisms. A central thrust of our original argument for PDGs over BNs is their ability to describe two different probabilities describing a single variable, such as  $\Pr(X|Y)$  and  $\Pr(X|Z)$ . The qualitative analogue of that expressiveness is precisely what allows us to capture functional dependence.

Given a hypergraph  $\mathcal{A} = (\mathcal{N}, \mathcal{A})$ ,  $X, Y \subseteq \mathcal{N}$ , and a natural number  $n \geq 0$ , let  $\mathcal{A} \sqcup_{X \rightarrow Y}^{(+n)}$  denote the hypergraph that results from augmenting  $\mathcal{A}$  with  $n$  additional (distinct) hyperarcs from  $X$  to  $Y$ .

- Theorem 5.2.** (a)  $\mu \models X \rightarrow\!\!\!\rightarrow Y \wedge \diamond \mathcal{A}$  if and only if  $\forall n \geq 0. \mu \models \diamond \mathcal{A} \sqcup_{X \rightarrow Y}^{(+n)}$ .
- (b) if  $\mathcal{A} = \mathcal{A}_G$  for a dag  $G$ , then  $\mu \models X \rightarrow\!\!\!\rightarrow Y \wedge \diamond \mathcal{A}$  if and only if  $\mu \models \diamond \mathcal{A} \sqcup_{X \rightarrow Y}^{(+1)}$ .
- (c) if  $\exists a \in \mathcal{A}$  such that  $S_a = \emptyset$  and  $X \in T_a$ , then  $\mu \models X \rightarrow\!\!\!\rightarrow Y \wedge \diamond \mathcal{A}$  iff  $\mu \models \diamond \mathcal{A} \sqcup_{X \rightarrow Y}^{(+2)}$ .

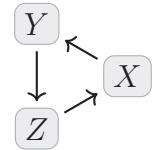
Based on the intuition given after [Example 5.2](#), it may seem unnecessary to ever add more than two parallel hyperarcs to ensure functional dependence in part (a). However, this intuition implicitly assumes that the randomness  $U_1$  and  $U_2$  of the two mechanisms is independent conditional on  $X$ , which may not be the case. See [Section 5.B.1](#) for counterexamples.

Finally, as alluded to above, QIM-compatibility gives meaning to cyclic structures, a topic that we will revisit ~~often~~ in [Sections 5.3](#) and [5.4](#). We start with some simple examples.

**Example 5.3.** Every  $\mu(X, Y)$  is compatible with  $[X] \rightleftarrows [Y]$ , because every distribution is compatible with  $\rightarrow [X] \rightarrow [Y]$ , and a mechanism with no inputs is a special case of one that can depend on  $Y$ .  $\triangle$

The logic above is an instance of an important reasoning principle, which we develop in [Section 10.2.1](#). Although the 2-cycle in [Example 5.3](#) is straightforward, generalizing it even slightly to a 3-cycle raises a not-so-straightforward question.

**Example 5.4.** What  $\mu(X, Y, Z)$  are compatible with the 3-cycle, shown on the right? By the reasoning above, among them must be all distributions consistent with a linear chain  $\rightarrow X \rightarrow Y \rightarrow Z$ . Thus, any distribution in which two variables are conditionally independent given the third is compatible with the 3-cycle. Are there distributions that



are *not* compatible with this hypergraph? It is not obvious. We return to this in [Section 5.4](#).  $\triangle$

Because QIM-compatibility applies to cyclic structures, one might wonder if it also captures the independencies of undirected models. Our definition of  $\mathcal{A}_G$ , as is common, implicitly identifies a undirected edge  $A-B$  with the pair  $\{A \rightarrow B, B \rightarrow A\}$  of directed edges; in this way, it naturally converts even an *undirected* graph  $G$  to a (directed) hypergraph. Compatibility with  $\mathcal{A}_G$ , however, does not coincide with any of the standard Markov properties corresponding to  $G$  ([Koller and Friedman 2009](#)). This may appear to be a flaw in [Definition 5.1](#), but it is unavoidable (see [Section 10.2.1](#)) if we wish to also capture causality, as we do in the next section.

### 5.3 QIM-Compatibility and Causality

Recall that in the definition of QIM-compatibility, each hyperarc represents an independent mechanism. Equations in a causal model are also viewed as representing independent mechanisms. This suggests a possible connection between the two formalisms, which we now explore. We will show that QIM-compatibility with  $\mathcal{A}$  means exactly that a distribution can be generated by a causal model with the corresponding dependency structure ([Section 5.3.1](#)). Moreover, such causal models and QIM-compatibility witnesses are themselves closely related ([Section 5.3.2](#)). In this section, we establish a causal grounding for QIM-compatibility. To do so, we must first review some standard definitions.

**Definition 5.2** ([Pearl \(2009\)](#)). A *structural equations model* (SEM) is a tuple  $M = (\mathcal{U}, \mathcal{V}, \mathcal{F})$ , where

- $\mathcal{U}$  is a set of exogenous variables;
- $\mathcal{V}$  is a set of endogenous variables (disjoint from  $\mathcal{U}$ );
- $\mathcal{F} = \{f_Y\}_{Y \in \mathcal{V}}$  associates to each endogenous variable  $Y$  an equation  $f_Y : \mathcal{V}(\mathcal{U} \cup \mathcal{V} - Y) \rightarrow \mathcal{V}(Y)$  that determines its value as a function of the other variables.

□

In a SEM  $M$ , a variable  $X \in \mathcal{V}$  does not depend on  $Y \in \mathcal{V} \cup \mathcal{U}$  if  $f_X(\dots, y, \dots) = f_X(\dots, y', \dots)$  for all  $y, y' \in \mathcal{V}(Y)$ . Let the parents  $\text{Pa}_M(X)$  of  $X$  be the set of variables on which  $X$  depends.  $M$  is acyclic iff  $\text{Pa}_M(X) \cap \mathcal{V} = \text{Pa}_G(X)$  for some dag  $G$  with vertices  $\mathcal{V}$ . In an acyclic SEM, it is easy to see that a setting of the exogenous variables uniquely determines the values of the endogenous variables (symbolically:  $M \models \mathcal{U} \rightarrow \mathcal{V}$ ). A probabilistic SEM (PSEM)  $\mathcal{M} = (M, P)$  is a SEM, together with a probability  $P$  over the exogenous variables. When  $\mathcal{M} \models \mathcal{U} \rightarrow \mathcal{V}$  (such as when  $M$  is acyclic), the distribution  $P(\mathcal{U})$  extends uniquely to a distribution over  $\mathcal{V}(\mathcal{V} \cup \mathcal{U})$ . A cyclic PSEM, however, may induce more than one such distribution, or none at all. In general, a PSEM  $\mathcal{M}$  induces a (possibly empty) convex set of distributions over  $\mathcal{V}(\mathcal{U} \cup \mathcal{V})$ . This set is defined by two (linear) constraints: the equations  $\mathcal{F}$  must hold with probability 1, and, in the case of a PSEM, the marginal probability over  $\mathcal{U}$  must equal  $P$ . Formally, for a PSEM  $\mathcal{M} = (M, P)$ , this means defining  $\{\mathcal{M}\} :=$

$$\left\{ \nu \in \Delta \mathcal{V}(\mathcal{V} \cup \mathcal{U}) \mid \forall Y \in \mathcal{V}. \nu(f_Y(\mathcal{U}, \mathcal{V} - Y) = Y) = 1, \nu(\mathcal{U}) = P(\mathcal{U}) \right\}$$

and defining  $\{M\}$  for an “ordinary” SEM  $M$  in the same way, except without the constraint involving  $P$ . To unpack the other constraint,  $f_Y(\mathcal{U}, \mathcal{V} - Y)$  is a random variable on the outcome space  $\mathcal{V}(\mathcal{V}, \mathcal{U})$ , and that it has the same value as  $Y$  is an event which, according to the equation  $f_Y$ , must always occur.  $\{M\}$  can be thought of as the set of distributions compatible wth  $M$ , and captures

the behavior of the causal model  $\mathcal{M}$  in the absence of intervention. It is worth noting that  $\{\mathcal{M}\}$  is also precisely the set-of-distribution semantics of  $\mathcal{M}$ , when regarded as a PDG; hence the notation. A joint distribution  $\mu(\mathbf{X})$  over  $\mathbf{X} \subseteq \mathcal{V} \cup \mathcal{U}$  can arise from a (P)SEM  $\mathcal{M}$  iff there is some  $\nu \in \{\mathcal{M}\}$  whose marginal on  $\mathbf{X}$  is  $\mu$ .

We now review the syntax of a language for describing causality. A *basic causal formula* is one of the form  $[\mathbf{Y} \leftarrow \mathbf{y}] \varphi$ , where  $\varphi$  is a Boolean expression over the endogenous variables  $\mathcal{V}$ ,  $\mathbf{Y} \subseteq \mathcal{V}$  is a subset of them, and  $\mathbf{y} \in \mathcal{V}(\mathbf{Y})$ . The language then consists of all Boolean combinations of basic formulas. In a causal model  $M$  and context  $\mathbf{u} \in \mathcal{V}(\mathcal{U})$ , a Boolean expression  $\varphi$  over  $\mathcal{V}$  is true iff it holds for all  $(\mathbf{u}, \mathbf{x}) \in \mathcal{V}(\mathcal{U}, \mathcal{V})$  consistent with the equations of  $M$ . Basic causal formulas are then given semantics by  $(M, \mathbf{u}) \models [\mathbf{Y} \leftarrow \mathbf{y}] \varphi$  iff  $(M_{\mathbf{Y} \leftarrow \mathbf{y}}, \mathbf{u}) \models \varphi$ , where  $M_{\mathbf{Y} \leftarrow \mathbf{y}}$  is the result of changing each  $f_Y$ , for  $Y \in \mathbf{Y}$ , to the constant function  $s \mapsto \mathbf{y}[Y]$ , which returns (on all inputs  $s$ ) the value of  $Y$  in the joint setting  $\mathbf{y}$ . From here, the truth relation can be extended to arbitrary causal formulas by structural induction in the usual way.<sup>2</sup> The dual formula  $\langle \mathbf{Y} \leftarrow \mathbf{y} \rangle \varphi := \neg [\mathbf{Y} \leftarrow \mathbf{y}] \neg \varphi$  is equivalent to  $[\mathbf{Y} \leftarrow \mathbf{y}] \varphi$  in SEMs where each context  $\mathbf{u}$  induces a unique setting of the endogenous variables (Halpern 2000). A PSEM  $\mathcal{M} = (M, P)$  assigns probabilities to causal formulas according to  $\Pr_{\mathcal{M}}(\varphi) := P(\{\mathbf{u} \in \mathcal{V}(\mathcal{U}) : (M, \mathbf{u}) \models \varphi\})$ .

Some authors assume that for each variable  $X$ , there is a special “independent noise” exogenous variable  $U_X$  (often written  $\epsilon_X$  in the literature) on which only the equation  $f_X$  can depend; we call a PSEM  $(M, P)$  *randomized* if it contains such exogenous variables that are mutually independent according to  $P$ , and *fully randomized* if all its exogenous variables are of this form. Randomized PSEMs

---

<sup>2</sup> $M \models \varphi_1 \wedge \varphi_2$  iff  $M \models \varphi_1$  and  $M \models \varphi_2$ ;  $M \models \neg \varphi$  iff  $M \not\models \varphi$ .

are clearly a special class of PSEM $s$ —yet, at the same time, correspond precisely to a “generalization” of PSEM $s$  whose equations need not be deterministic. Indeed, any PSEM can be converted to an equivalent randomized PSEM by extending it with additional dummy variables  $\{U_X\}_{X \in \mathcal{V}}$  that can take only a single value. Thus, we do not lose expressive power by using randomized PSEM $s$ . In fact, *qualitatively*, randomized PSEM $s$  are more expressive: they can encode independence. Based on their definitions, it should come as no surprise that randomized PSEM $s$  and QIM-compatibility are related.

### 5.3.1 The Equivalence Between QIM-Compatibility and Arising from a Randomized PSEM

We are now equipped to formally describe the connection between QIM-compatibility and causality. At a high level, this connection should be unsurprising: witnesses and causal models both relate dependency structures to distributions, but in “opposite directions”. QIM-compatibility starts with distributions and asks what dependency structures they are compatible with. Causal models, on the other hand, are explicit (quantitative) representations of dependency structures that give rise to sets of distributions. We now show that the existence of a causal model coincides with the existence of a witness. We start by showing this for the hypergraphs generated by graphs (like Bayesian networks, except possibly cyclic), which we show correspond to fully randomized causal models ([Proposition 5.3](#)). We then give a natural generalization of a causal model that exactly captures QIM-compatibility with an arbitrary hypergraph ([Proposition 5.4](#)). In both cases, the high-level result is the same:  $\mu \models \mathcal{A}$  iff there is a causal model that “has dependency structure  $\mathcal{A}$ ” that gives

rise to  $\mu$ .

More precisely, we say that a randomized causal model  $\mathcal{M}$  has dependency structure  $\mathcal{A}$  iff there is a 1-1 correspondence between  $a \in \mathcal{A}$  and the equations of  $\mathcal{M}$ , such that the equation  $f_a$  produces a value of  $T_a$  and depends only on  $S_a$  and  $U_a$ . This definition emphasizes the hypergraph; here is a more concrete alternative emphasizing the randomized PSEM:  $\mathcal{M}$  is of dependency structure  $\mathcal{A}$  iff the targets of  $\mathcal{A}$  are disjoint singletons (the elements of  $\mathcal{V}$ ), and  $\text{Pa}_{\mathcal{M}}(Y) \subseteq S_Y \cup \{U_Y\}$  for all  $Y \in \mathcal{V}$ . We start by presenting the result in the case where  $\mathcal{A}$  corresponds to a directed graph.

**Proposition 5.3.** *Given a graph  $G$  and a distribution  $\mu$ ,  $\mu \models \Diamond \mathcal{A}_G$  iff there exists a fully randomized PSEM of dependency structure  $\mathcal{A}_G$  from which  $\mu$  can arise.*

[ link to proof ]

Proposition 5.3 shows that, for those hypergraphs induced by graphs, QIM-compatibility means arising from a fully randomized PSEM of the appropriate dependency structure. Theorem 5.1 makes precise a phenomenon that seems to be almost universally implicitly understood but, to the best of our knowledge, has not been formalized before: every acyclic fully randomized SEM induces a distribution with the independencies of the corresponding Bayesian Network—and, conversely, every distribution with those independencies arises from such a causal model.

It is easy to extend this result to the dependency structures of all randomized PSEMs. But what happens if  $\mathcal{A}$  contains hyperarcs with overlapping targets? Here the correspondence starts to break down for a simple reason: by definition, there is at most one equation per variable in a (P)SEM; thus, no PSEM can have dependency structure  $\mathcal{A}$ . Nevertheless, the correspondence between witnesses

and causal models persists if we simply drop the (traditional) requirement that  $\mathcal{F}$  is indexed by  $\mathcal{V}$ . This leads us to consider a natural generalization of a (randomized) PSEM that has an arbitrary set of equations—not just one per variable.

**Definition 5.3.** Let  $(\mathcal{N}, \mathcal{A})$  be a hypergraph. A *generalized randomized PSEM*  $\mathcal{M} = (\mathcal{X}, \mathcal{U}, \mathcal{F}, P)$  with structure  $\mathcal{A}$  consists of sets of variables  $\mathcal{X}$  and  $\mathcal{U} = \{U_a\}_{a \in \mathcal{A}}$ , together with a set of functions  $\mathcal{F} = \{f_a : \mathcal{V}(S_a) \times \mathcal{V}(U_a) \rightarrow \mathcal{V}(T_a)\}_{a \in \mathcal{A}}$ , and a probability  $P_a$  over each independent noise variable  $U_a$ . The meanings of  $\{\mathcal{M}\}$  and *can arise* are the same as for a PSEM. □

**Proposition 5.4.**  $\mu \models \Diamond \mathcal{A}$  iff there exists a generalized randomized PSEM with structure  $\mathcal{A}$  from which  $\mu$  can arise. link to proof

Generalized randomized PSEMs can capture functional dependencies, and constraints. For instance, an equality (say  $X = Y$ ) can be encoded in a generalized randomized PSEM with a second equation for  $X$ . Indeed, we believe that generalized randomized PSEMs can capture a wide class of constraints, and are closely related to *causal models with constraints* (Beckers et al. 2023), a discussion we defer to future work.

### 5.3.2 Interventions, and the Correspondence

#### Between Witnesses and Causal Models

We have seen that QIM-compatibility with  $\mathcal{A}$  (i.e., the existence of a witness  $\bar{\mu}$ ) coincides exactly with the existence of a causal model  $\mathcal{M}$  from which a distribution can arise. But which witnesses correspond to which causal models? The

answer to this question will be critical to extend the correspondence we have given so that it can deal with interventions. Different causal models may give rise to the same distribution, yet handle interventions differently.

There are two directions of the correspondence. Given a randomized PSEM  $\mathcal{M}$ , distributions arising from it are compatible with its dependency structure, and the corresponding witnesses are exactly the distributions in  $\{\mathcal{M}\}$  (see [Section 5.C](#)). In particular, if  $\mathcal{M}$  is acyclic, there is a unique witness. The converse is more interesting: how can we turn a witness into a causal model?

**Construction 5.4.** Given a witness  $\bar{\mu}(\mathcal{X})$  to compatibility with a hypergraph  $\mathcal{A}$  with disjoint targets, construct a PSEM according to the following (non-deterministic) procedure. Take  $\mathcal{V} := \cup_{a \in \mathcal{A}} T_a$ ,  $\mathcal{U} := \mathcal{U}_{\mathcal{A}} \cup (\mathcal{X} - \mathcal{V})$ , and  $P(\mathcal{U}) := \bar{\mu}(\mathcal{U})$ . For each  $X \in \mathcal{V}$ , there is a unique  $a_X \in \mathcal{A}$  whose targets  $T_{a_X}$  contain  $X$ . Since  $\bar{\mu} \models (U_{a_X}, S_{a_X}) \rightarrow\!\!\! \rightarrow T_{a_X}$  (this is just property (c) in [Definition 5.1](#)),  $X \in T_{a_X}$  must also be a function of  $S_{a_X}$  and  $U_{a_X}$ ; take  $f_X$  to be such a function. More precisely, for each  $u \in \mathcal{V}(U_{a_X})$  and  $s \in \mathcal{V}(S_{a_X})$  for which  $\bar{\mu}(U_{a_X}=u, S_{a_X}=s) > 0$ , there is a unique  $t \in \mathcal{V}(T_{a_X})$  such that  $\bar{\mu}(u, s, t) > 0$ . In this case, set  $f_X(u, s, \dots) := t[X]$ . If  $\bar{\mu}(U_{a_X}=u, S_{a_X}=s) = 0$ ,  $f_X(u, s, \dots)$  can be an arbitrary function of  $u$  and  $s$ . Let  $\text{PSEMs}_{\mathcal{A}}(\bar{\mu})$  denote the set of PSEMs that can result. □

It's clear from [Construction 5.4](#) that  $\text{PSEMs}_{\mathcal{A}}(\bar{\mu})$  is always nonempty, and is a singleton iff  $\bar{\mu}(u, s) > 0$  for all  $(a, u, s) \in \sqcup_{a \in \mathcal{A}} \mathcal{V}(U_a, S_a)$ . A witness with this property exists when  $\mu$  is positive (i.e.,  $\mu(\mathcal{X}=\mathbf{x}) > 0$  for all  $\mathbf{x} \in \mathcal{V}(\mathcal{X})$ ), in which case the construction gives a unique causal model. Conversely, we have seen that an acyclic model  $\mathcal{M}$  gives rise to a unique witness. So, in the simplest cases, models  $\mathcal{M}$  with structure  $\mathcal{A}$  and witnesses  $\bar{\mu}$  to compatibility with  $\mathcal{A}$  are equivalent. But there are two important caveats.

1. A causal model  $\mathcal{M}$  can contain more information than a witness  $\bar{\mu}$  if some events have probability zero. For instance,  $\bar{\mu}$  could be a point mass on a single joint outcome  $\omega$  of all variables that satisfies the equations of  $\mathcal{M}$ . But  $\mathcal{M}$  cannot be reconstructed uniquely from  $\bar{\mu}$  because there may be many causal models for which  $\omega$  is a solution.
2. A witness  $\bar{\mu}$  can contain more information than a causal model  $\mathcal{M}$  if  $\mathcal{M}$  is cyclic. For example, suppose that  $\mathcal{M}$  consists of two variables,  $X$  and  $X'$ , and equations  $f_X(X') = X'$  and  $f_{X'}(X) = X$ . In this case,  $\bar{\mu}$  cannot be reconstructed from  $\mathcal{M}$ , because  $\mathcal{M}$  does not contain information about the distribution of  $X$ .

These two caveats appear to be very different, but they fit together in a surprisingly elegant way.

**Proposition 5.5.** *If  $\bar{\mu}(\mathcal{X}, \mathcal{U}_{\mathcal{A}})$  is a witness for QIM-compatibility with  $\mathcal{A}$  and  $\mathcal{M}$  is a PSEM with dependency structure  $\mathcal{A}$ , then  $\bar{\mu} \in \{\mathcal{M}\}$  if and only if  $\mathcal{M} \in \text{PSEMs}_{\mathcal{A}}(\bar{\mu})$ .*

[ link to proof ]

Equivalently, this means that  $\text{PSEMs}_{\mathcal{A}}(\bar{\mu})$ , the possible outputs of [Construction 5.4](#), are precisely the randomized PSEMs of dependency structure  $\mathcal{A}$  that can give rise to  $\bar{\mu}$ . This is already substantial evidence that causal models  $\mathcal{M} \in \text{PSEMs}_{\mathcal{A}}(\bar{\mu})$  are closely related to the QIM-compatibility witness  $\bar{\mu}$ . But everything we have seen so far describes only the correspondence in the absence of intervention, a setting in which many causal models are indistinguishable. We now show that the correspondence goes deeper, by extending it to interventions. In any randomized PSEM  $M$ , we can define an event

$$\text{do}_M(\mathbf{X}=\mathbf{x}) := \bigcap_{X \in \mathbf{X}} \bigcap_{\mathbf{s} \in \mathcal{V}(\mathbf{Pa}(X))} f_X(U_X, \mathbf{s}) = \mathbf{x}[X], \quad \begin{array}{l} \text{where } \mathbf{x}[X] \text{ is the} \\ \text{value of } X \text{ in } \mathbf{x}. \end{array} \quad (5.1)$$

This is intuitively the event in which the randomness is such that  $\mathbf{X} = \mathbf{x}$  regardless of the values of the parent variables.<sup>3</sup> As we now show, conditioning on  $\text{do}_M(\mathbf{X}=\mathbf{x})$  has the effect of intervention.

**Theorem 5.6.** Suppose that  $\bar{\mu}$  is a witness to  $\mu \models \Diamond \mathcal{A}$ ,  $\mathcal{M} \in \text{PSEM}_{\mathcal{A}}(\bar{\mu})$ ,  $\mathbf{X} \subseteq \mathcal{X}$  and  $\mathbf{x} \in \mathcal{V}(\mathbf{X})$ . If  $\bar{\mu}(\text{do}_{\mathcal{M}}(\mathbf{X}=\mathbf{x})) > 0$ , then:

link to  
proof

(a)  $\bar{\mu}(\mathcal{X} \mid \text{do}_{\mathcal{M}}(\mathbf{X}=\mathbf{x}))$  can arise from  $\mathcal{M}_{\mathbf{X} \leftarrow \mathbf{x}}$ ;

(b) for all events  $\varphi \subseteq \mathcal{V}(\mathcal{X})$ ,  $\Pr_{\mathcal{M}}([\mathbf{X} \leftarrow \mathbf{x}] \varphi) \leq \bar{\mu}(\varphi \mid \text{do}_{\mathcal{M}}(\mathbf{X}=\mathbf{x})) \leq \Pr_{\mathcal{M}}([\langle \mathbf{X} \leftarrow \mathbf{x} \rangle] \varphi)$

and all three are equal when  $\mathcal{M} \models \mathcal{U} \twoheadrightarrow \mathcal{X}$  (such as when  $\mathcal{M}$  is acyclic).

Theorem 5.6 shows that the relationship between witnesses and causal models extends to interventions. Even when  $\text{do}_M(\mathbf{X}=\mathbf{x})$  has probability zero, it is always possible to find a nearly equivalent setting where the bounds of the theorem apply.<sup>4</sup> Intervention and conditioning are conceptually very different, so it may seem surprising that conditioning can have the effect of intervention (and also that the Pearl's  $\text{do}(\cdot)$  notation actually corresponds to an event (Hitchcock 2024)). We emphasize that the conditioning (on  $\text{do}_M(\mathbf{X}=\mathbf{x})$ ) is on the randomness  $U_X$  and not  $X$  itself; intervening on  $\mathbf{X}=\mathbf{x}$  is indeed fundamentally different from conditioning on  $\mathbf{X}=\mathbf{x}$ .

---

<sup>3</sup>This is essentially the event in which, for each  $X \in \mathbf{X}$ , the response variable  $\hat{U}_X := \lambda s. f_X(s, U_X)$ , whose possible values  $\mathcal{V}(\hat{U}_X)$  are functions from  $\mathcal{V}(\text{Pa}_M(X))$  to  $\mathcal{V}(X)$  (Rubin 1974; Balke and Pearl 1994), takes on the constant function  $\lambda p. x$ .

<sup>4</sup>More precisely, for all  $\epsilon > 0$ , there exists some  $\mathcal{M}'$  that differs from  $\mathcal{M}$  on the probabilities all causal formulas by at most  $\epsilon$ , and a distribution  $\bar{\mu}'$  that is  $\epsilon$ -close to  $\bar{\mu}$ , such that  $\bar{\mu}'(\text{do}_{\mathcal{M}'}(\mathbf{X}=\mathbf{x})) > 0$ . As a result, Theorem 5.6 places bounds on the conditional probabilities that are possible limits of sequences of distributions  $(\nu_k)_{k \geq 0}$  where  $\nu_k(\text{do}_M(\mathbf{X}=\mathbf{x})) > 0$ , i.e., the possible outcomes of conditioning a *non-standard* probability measure (Halpern 2017) on this probability-zero event.

## 5.4 QIM-Compatibility and Information Theory

The fact that the dependency structure of a (causal) Bayesian network describes the independencies of the distribution it induces is fundamental to both causality and probability. It makes explicit the distributional consequences of BN structure. Yet, despite substantial interest (Baier et al. 2022), generalizing the BN case to more complex (e.g., cyclic) dependency structures remains largely an open problem. In Section 5.4.1, we generalize the BN case by providing an information-theoretic constraint, capable of capturing conditional independence, functional dependence, and more, on the distributions that can arise from an *arbitrary* dependency structure. This connection between causality and information theory has implications for both fields. It grounds the cyclic dependency structures found in causality in concrete constraints on the distributions they represent. At the same time, it allows us to resolve longstanding confusion about structure in information theory, clarifying the meaning of the so-called “interaction information”, and recasting a standard counterexample to substantiate the claim it was intended to oppose. In Section 5.4.2, we strengthen this connection. Using entropy to measure distance to (in)dependence, we develop a scoring function to measure how far a distribution is from being QIM-compatible with a given dependency structure. This function turns out to have an intimate relationship with the qualitative PDG scoring function  $SDef$ , which we use to show that our information-theoretic constraints degrade gracefully on “near-compatible” distributions.

We now review a few critical information-theoretic concepts and their relationships to (in)dependence (see Section 2.6.1 for a full primer). Conditional entropy  $H_\mu(Y|X)$  measures how far  $\mu$  is from satisfying the functional depen-

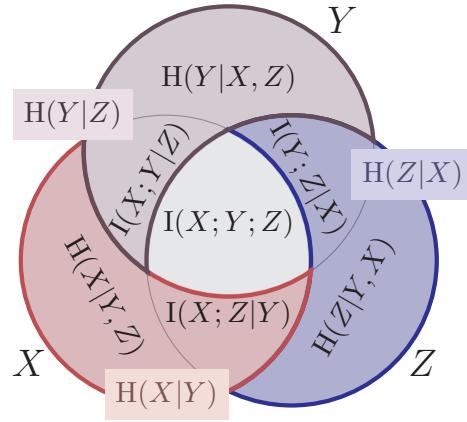


Figure 5.1: A graphical illustration that  $I(X; Y; Z) = H(X, Y, Z) - H(X|Y) + H(Y|Z) + H(Z|X)$ , on an information diagram.

dency  $X \rightarrow Y$ . Conditional mutual information  $I_\mu(Y; Z|X)$  measures how far  $\mu$  is from satisfying the conditional independence  $Y \perp\!\!\!\perp Z | X$ . Linear combinations of these quantities (for  $X, Y, Z \subseteq \mathcal{X}$ ) can be viewed as the inner product between a coefficient vector  $v$  and a  $2^{|\mathcal{X}|} - 1$  dimensional vector  $\mathbf{I}_\mu$  that we will call the *information profile* of  $\mu$ . For three variables, the components of this vector were illustrated in Figure 2.2; for convenience, we give another illustration of it in Figure 5.1. It is not hard to see that an arbitrary conjunction of (conditional) (in)dependencies can be expressed as a constraint  $\mathbf{I}_\mu \cdot v \geq 0$ , for some appropriate vector  $v$ .

We now return to the qualitative PDG scoring function  $SDef$ , which interprets a hypergraph structure  $\mathcal{A}$  as a function of the form  $\mathbf{I}_\mu \cdot v_{\mathcal{A}}$ . Recall that this *structural information deficiency*, given by

$$SDef_{\mathcal{A}}(\mu) = \mathbf{I}_\mu \cdot v_{\mathcal{A}} := -H_\mu(\mathcal{X}) + \sum_{a \in \mathcal{A}} H_\mu(T_a | S_a), \quad (5.2)$$

is the difference between the number of bits needed to (independently) specify the randomness in  $\mu$  along the hyperarcs of  $\mathcal{A}$ , and the number of bits needed to specify a sample of  $\mu$  according to its own structure ( $\emptyset \rightarrow \mathcal{X}$ ). While we have

seen that  $SDef$  has some nice properties,<sup>5</sup> it can also behave unintuitively in some cases; for instance, it can be negative. Clearly, it does not measure how close  $\mu$  is to being structurally compatible with  $\mathcal{A}$ , in general. Nevertheless, there is still a fundamental relationship between  $SDef$  and QIM-compatibility, as we now show.

### 5.4.1 A Necessary Condition for QIM-Compatibility

What constraints does QIM-compatibility with  $\mathcal{A}$  place on a distribution  $\mu$ ? When  $G$  is a dag, we have seen that if  $\mu \models \Diamond\mathcal{A}_G$ , then  $\mu$  must satisfy the independencies of the corresponding Bayesian network ([Theorem 5.1](#)); we have also seen that additional hyperarcs impose functional dependencies ([Theorem 5.2](#)). But these results apply only when  $\mathcal{A}$  is of a very special form. More generally,  $\mu \models \Diamond\mathcal{A}$  implies that  $\mu$  can arise from some randomized causal model whose equations have dependency structure  $\mathcal{A}$  ([Propositions 5.3 and 5.4](#)). Still, unless  $\mathcal{A}$  has a particularly special form, it is not obvious whether or not this says something about  $\mu$ . The primary result of this section is an information-theoretic bound ([Theorem 5.7](#)) that generalizes most of the concrete consequences of QIM-compatibility we have seen so far ([Theorems 5.1 and 5.2](#)). The result is a connection between information theory and causality; it yields an information-theoretic test for complex causal dependency structures, and enables causal notions of structure to dispel misconceptions in information theory.

**Theorem 5.7.** *If  $\mu \models \Diamond\mathcal{A}$ , then  $SDef_{\mathcal{A}}(\mu) \leq 0$ .*

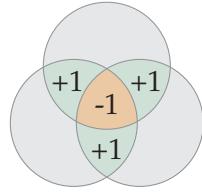
[ link to  
proof ]

---

<sup>5</sup>For instance, it captures BN independencies and the dependencies of [Theorem 5.2](#), reduces to maximum entropy for the empty hypergraph, and combines with the quantitative PDG scoring function ([Richardson and Halpern 2021](#)) to capture factor graphs.

[Theorem 5.7](#) applies to all hypergraphs, and subsumes every general-purpose technique we know of for proving that  $\mu \not\models \mathcal{A}$ . Indeed, the negative directions of [Theorems 5.1](#) and [5.2](#) are immediate consequences of it. To illustrate some of its subtler implications, let's return to the 3-cycle in [Example 5.4](#).

**Example 5.5.** It is easy to see (e.g., by inspecting [Figure 5.1](#)) that  $SDef_{3\text{-cycle}}(\mu) = H_\mu(Y|X) + H_\mu(Z|Y) + H_\mu(X|Z) - H_\mu(XYZ) = -I_\mu(X; Y; Z)$ . [Theorem 5.7](#) therefore tells us that a distribution  $\mu$  that is QIM-compatible with the 3-cycle cannot have negative interaction information  $I_\mu(X; Y; Z)$ . What does this mean? Overall, conditioning on one variable can only reduce the amount of remaining information in other variables (in expectation). However, when  $I(X; Y; Z) < 0$ , conditioning on one variable causes the other two to share more information than they did before. The most extreme instance is  $\mu_{xor}$ , the distribution in which two variables are independent and the third is their parity (illustrated on the right). It seems intuitively clear that  $\mu_{xor}$  cannot arise from the 3-cycle, a causal model with only pairwise dependencies. This is difficult to prove directly, but is an immediate consequence of [Theorem 5.7](#).  $\triangle$

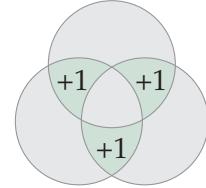


For many, there is an intuition that  $I(X; Y; Z) < 0$  should require a fundamentally “3-way” interaction between the variables, and should not arise through pairwise interactions alone ([James 2018](#)). This has been a source of conflict ([Williams and Beer 2010; MacKay 2003; leonbloy 2015; Cover and Thomas 1991](#)), because traditional ways of making precise “pairwise interactions” (e.g., maximum entropy subject to pairwise marginal constraints and pairwise factorization) do not ensure that  $I(X; Y; Z) \geq 0$ . But QIM-compatibility does. One can verify by enumeration that the 3-cycle is the most expressive causal structure with no joint dependencies, and we have already proven that QIM-compatibility with that

hypergraph implies non-negative interaction information. QIM-compatibility has another even more noteworthy clarifying effect on information theory.

There is a school of thought that contends that *all* structural information in  $\mu(\mathcal{X})$  is captured by its information profile  $I_\mu$ . This position has fallen out of favor in some communities due to standard counterexamples: distributions that have intuitively different structures yet share an information profile ([James and Crutchfield 2017](#)). However, with “structure” explicated by compatibility, the prototypical counterexample of this kind suddenly supports the very notion it was meant to challenge, suggesting in an unexpected way that the information profile may yet capture the essence of probabilistic structure.

**Example 5.6.** Let  $A$ ,  $B$ , and  $C$  be variables with  $\mathcal{V}(A), \mathcal{V}(B), \mathcal{V}(C) = \{0, 1\}^2$ . Using independent fair coin flips  $X_1$ ,  $X_2$ , and  $X_3$ , define two joint distributions,  $P$  and  $Q$ , over  $A, B, C$  as follows. Define  $P(A, B, C)$  by letting  $A := (X_1, X_2)$ ,  $B := (X_2, X_3)$ , and  $C := (X_3, X_1)$ . Define  $Q$  by letting  $A := (X_1, X_2)$ ,  $B := (X_1, X_3)$ , and  $C := (X_1, X_2 \oplus X_3)$ . Structurally,  $P$  and  $Q$  appear to be very different. According to  $P$ , the first components of the three variables  $(A, B, C)$  are independent, yet they are identical according to  $Q$ . Moreover,  $P$  has only simple pairwise interactions between the variables, while  $P$  has  $\mu_{xor}$  (a clear 3-way interaction) embedded within it. Yet  $P$  and  $Q$  have identical information profiles (see right): in both cases, each of  $\{A, B, C\}$  is determined by the values of the other two, each pair share one bit of information given the third, and  $I(A; B; C) = 0$ .



This example has been used to argue that multivariate Shannon information does not take into account important structural differences between distributions

(James and Crutchfield 2017). We are now in a position to give a novel and far deeper response, by appealing to QIM-compatibility.<sup>6</sup> Unsurprisingly,  $P$  is compatible with the 3-cycle; it is clearly consists of “2-way” interactions, as each pair of variables shares a bit. But, counterintuitively, the distribution  $Q$  is *also* compatible with the 3-cycle! (The reader is encouraged to verify that  $U_1 = X_3 \oplus X_1$ ,  $U_2 = X_2$ , and  $U_3 = X_3$  serves as a witness.) To emphasize: this is despite the fact that  $Q$  is just  $\mu_{xor}$  (which is certainly not compatible with the 3-cycle) together with a seemingly irrelevant random bit  $X_1$ . By the results of Section 5.3, this means there is a causal model without joint dependence giving rise to  $Q$ —so, despite appearances,  $Q$  does not require a 3-way interaction. Indeed,  $P$  and  $Q$  are QIM-compatible with precisely the same hypergraphs over  $\{A, B, C\}$ , suggesting that they don’t have a structural difference after all.  $\triangle$

In light of Example 5.6, one might reasonably conjecture that the converse of Theorem 5.7 holds. Unfortunately, it does not (see Section 5.B.2); the quantity  $SDef_{\mathcal{A}}(\mu)$  does not completely determine whether or not  $\mu \models \Diamond \mathcal{A}$ . We now pursue a new (entropy-based) scoring function that does. This will allow us to generalize Theorem 5.7 to distributions that are only “near-compatible” with  $\mathcal{A}$ .

---

<sup>6</sup>Note that  $P$  and  $Q$  no longer have the same profile if we split each variable into its two components. Since the notion of “component” is based on the assignment  $\mathcal{V}$  of variables to possible values, our view that  $\mathcal{V}$  is not structural information diffuses this counterexample by assumption—but the present argument is much stronger.

### 5.4.2 A Scoring Function for QIM-Compatibility

Here is a function that measures how far a distribution  $\mu$  is from being QIM-compatible with  $\mathcal{A}$ .

$$\text{QIMInc}_{\mathcal{A}}(\mu) := \inf_{\substack{\nu(\mathcal{U}, \mathcal{X}) \\ \nu(\mathcal{X}) = \mu(\mathcal{X})}} -H_\nu(\mathcal{U}) + \sum_{a \in \mathcal{A}} H_\nu(U_a) + \sum_{a \in \mathcal{A}} H_\nu(T_a | S_a, U_a). \quad (5.3)$$

$\text{QIMInc}$  is a direct translation of [Definition 5.1](#) (a-c); it measures the (optimal) quality of an extended distribution  $\nu$  as a witness. The infimum restricts the search to  $\nu$  satisfying (a), the first two terms measure  $\nu$ 's discrepancy of with (b), and the last term measures  $\nu$ 's discrepancy with (c). Therefore:

**Proposition 5.8.**  $\text{QIMInc}_{\mathcal{A}}(\mu) \geq 0$ , with equality iff  $\mu \models \mathcal{A}$ .

[ [link to proof](#) ]

Although they seem to be very different,  $\text{QIMInc}$  and  $SDef$  turn out to be closely related. In fact, modulo the infimum,  $\text{QIMInc}_{\mathcal{A}}$  is a special case of  $SDef$ —not for the hypergraph  $\mathcal{A}$ , but rather for a transformed one  $\mathcal{A}^\dagger$  that models the noise variables explicitly. To construct  $\mathcal{A}^\dagger$  from  $\mathcal{A}$ , add new nodes  $\mathcal{U} = \{U_a\}_{a \in \mathcal{A}}$ , and replace each hyperarc



Finally, add one additional hyperarc  $\mathcal{U} \rightarrow \mathcal{X}$ . (Intuitively, this hyperarc creates functional dependencies in the spirit of [Theorem 5.2](#).) With these definitions in place, we can state a theorem that bounds  $\text{QIMInc}$  above and below with information deficiencies. The lower bound generalizes [Theorem 5.7](#) by giving an upper limit on  $SDef_{\mathcal{A}}(\mu)$  even for distributions  $\mu$  that are not QIM-compatible with  $\mathcal{A}$ . The upper bound is tight in general, and shows that  $\text{QIMInc}_{\mathcal{A}}$  can be equivalently defined as a minimization over  $SDef_{\mathcal{A}^\dagger}$ .

**Theorem 5.9.** (a) If  $(\mathcal{X}, \mathcal{A})$  is a hypergraph,  $\mu(\mathcal{X})$  is a distribution, and  $\nu(\mathcal{X}, \mathcal{U})$  is an extension of  $\nu$  to additional variables  $\mathcal{U} = \{U_a\}_{a \in \mathcal{A}}$  indexed by  $\mathcal{A}$ , then:

$$SDef_{\mathcal{A}}(\mu) \leq QIMInc_{\mathcal{A}}(\mu) \leq SDef_{\mathcal{A}^\dagger}(\nu).$$

(b) For all  $\mu$  and  $\mathcal{A}$ , there is a choice of  $\nu$  that achieves the upper bound. That is,

$$QIMInc_{\mathcal{A}}(\mu) = \min \left\{ SDef_{\mathcal{A}^\dagger}(\nu) : \begin{array}{l} \nu \in \Delta\mathcal{V}(\mathcal{X}, \mathcal{U}) \\ \nu(\mathcal{X}) = \mu(\mathcal{X}) \end{array} \right\}.$$

## 5.5 Discussion

We have shown how directed hypergraphs can be used to represent structural aspects of distributions. Moreover, they can do so in a way that generalizes conditional independencies and functional dependencies and has deep connections to causality and information theory. Many open questions remain. A major one is that of more precisely understanding QIM-compatibility in cyclic models. We do not yet know, for example, whether the same set of distributions are QIM-compatible with the clockwise and counter-clockwise 3-cycles. A related problem is to find an efficient procedure that can determine whether a given distribution is QIM-compatible with a hypergraph. We hope to explore all these questions in future work.

## APPENDICES FOR CHAPTER 5

### 5.A Proofs

We begin with a de-randomization construction, that will be useful for the proofs.

#### 5.A.1 From CPDs to Distributions over Functions

Compare two objects:

- a cpd  $p(Y|X)$ , and
- a distribution  $q(Y^X)$  over functions  $g : \mathcal{V}X \rightarrow \mathcal{V}Y$ .

The latter is significantly larger — if both  $|\mathcal{V}X| = |\mathcal{V}Y| = N$ , then  $q$  is a  $N^N$  dimensional object, while  $p$  is only dimension  $N^2$ . A choice of distribution  $q(Y^X)$  corresponds to a unique choice cpd  $p(Y|X)$ , according to

$$p(Y=y | X=x) := q(Y^X(x) = y).$$

**Claim 1.** 1. *The definition above in fact yields a cpd, i.e.,  $\sum_y p(Y=y|X=x) = 1$  for all  $x \in \mathcal{V}X$ .*

2. *This definition of  $p(Y|X)$  is the conditional marginal of any joint distribution  $\mu(X, Y, Y^X)$  satisfying  $\mu(Y^X) = q$  and  $\mu(Y = Y^X(X)) = 1$ .*

Both  $p$  and  $q$  give probabilistic information about  $Y$  conditioned on  $X$ . But  $q(Y^X)$  contains strictly more information. Not only does it specify the distribution over  $Y$  given  $X=x$ , but it also contains counter-factual information about

the distribution of  $Y$  if  $X$  were equal to  $x'$ , conditioned on the fact that, in reality,  $X=x$ .

Is there a natural construction that goes in the opposite direction, intuitively making as many independence assumptions as possible? It turns out there is:

$$q(Y^X = g) = \prod_{x \in V^X} p(Y=g(x) \mid X=x).$$

Think of  $Y^X$  as a collection of variables  $\{Y^x : x \in V^X\}$  describing the value of the function for each input, so that  $q$  is a joint distribution over them. This construction simply asks that these variables be independent. Specifying a distribution with these independences amounts to a choice of “marginal” distribution  $q(Y^x)$  for each  $x \in V^X$ , and hence is essentially a function of type  $V^X \rightarrow \Delta V^Y$ , the same as  $p$ . In addition, if we apply the previous construction, we recover  $p$ , since:

$$\begin{aligned} q(Y^X(x) = y) &= \sum_{g: V^X \rightarrow V^Y} \mathbb{1}[g(x) = y] \prod_{x' \in V^X} p(Y=g(x') \mid X=x') \\ &= \sum_{g: V^X \rightarrow V^Y} \mathbb{1}[g(x) = y] p(Y=g(x) \mid X=x) \prod_{x' \neq x} p(Y=g(x') \mid X=x') \\ &= p(Y=y \mid X=x) \sum_{g: V^X \rightarrow V^Y} \mathbb{1}[g(x) = y] \prod_{x' \neq x} p(Y=g(x') \mid X=x') \\ &= p(Y=y \mid X=x) \sum_{g: V^X \setminus \{x\} \rightarrow V^Y} \prod_{x' \in V^X \setminus \{x\}} p(Y=g(x') \mid X=x') \\ &= p(Y=y \mid X=x). \end{aligned}$$

The final equality holds because the remainder of the terms can be viewed as the probability of selecting any function from  $X \setminus \{x\}$  to  $Y$ , under an analogous measure; thus, it equals 1. This will be a useful construction for us in general.

### 5.A.2 Results on (In)dependence

**Lemma 5.10.** Suppose  $X_1, \dots, X_n$  are variables,  $Y_1, \dots, Y_n$  are sets, and for each  $i \in \{1, \dots, n\}$ , we have a function  $f_i : V(X_i) \rightarrow Y_i$ . Then if  $X_1, \dots, X_n$  are mutually independent (according to a joint distribution  $\mu$ ), then so are  $f_1(X_1), \dots, f_n(X_n)$ .

*Proof.* This is an intuitive fact, but we provide a proof for completeness. Explicitly, mutual independence of  $X_1, \dots, X_n$  means that, for all joint settings  $\mathbf{x} = (x_1, \dots, x_n)$ , we have  $\mu(X_1=x_1, \dots, X_n=x_n) = \prod_{i=1}^n \mu(X_i=x_i)$ . So, for any joint setting  $\mathbf{y} = (y_1, \dots, y_n) \in Y_1 \times \dots \times Y_n$ , we have

$$\begin{aligned}
\mu(f_1(X_1)=y_1, \dots, f_n(X_n)=y_n) &= \mu(\{\mathbf{x} : \mathbf{f}(\mathbf{x}) = \mathbf{y}\}) \\
&= \sum_{\substack{(x_1, \dots, x_n) \in V(X_1, \dots, X_n) \\ f_1(x_1)=y_1, \dots, f_n(x_n)=y_n}} \mu(X_1=x_1, \dots, X_n=x_n) \\
&= \sum_{\substack{x_1 \in V X_1 \\ f_1(x_1)=y_1}} \cdots \sum_{\substack{x_n \in V X_n \\ f_n(x_n)=y_n}} \mu(X_1=x_1, \dots, X_n=x_n) \\
&= \sum_{\substack{x_1 \in V X_1 \\ f_1(x_1)=y_1}} \cdots \sum_{\substack{x_n \in V X_n \\ f_n(x_n)=y_n}} \prod_{i=1}^n \mu(X_i=x_i) \\
&= \left( \sum_{\substack{x_1 \in V X_1 \\ f_1(x_1)=y_1}} \mu(X_1=x_1) \right) \cdots \left( \sum_{\substack{x_n \in V X_n \\ f_n(x_n)=y_n}} \mu(Y_1=y_1) \right) \\
&= \prod_{i=1}^n \mu(f_i(X_i)=y_i). \quad \square
\end{aligned}$$

**Lemma 5.11** (properties of determination).

1. If  $\nu \models A \twoheadrightarrow B$  and  $\nu \models A \twoheadrightarrow C$ , then  $\nu \models A \twoheadrightarrow (B, C)$ .
2. If  $\nu \models A \twoheadrightarrow B$  and  $\nu \models B \twoheadrightarrow C$ , then  $\nu \models A \twoheadrightarrow C$ .

*Proof.*  $\nu \models X \twoheadrightarrow Y$ , means there exists a function  $f : V(A) \rightarrow V(B)$  such that

$\nu(f(Y) = X) = 1$ , i.e., the event  $f(A) = B$  occurs with probability 1.

1. Let  $f : \mathcal{V}(A) \rightarrow \mathcal{V}(B)$  and  $g : \mathcal{V}(A) \rightarrow \mathcal{V}(C)$  be such that  $\nu(f(A) = B) = 1 = \nu(g(A) = C)$ . Since both events happen with probability 1, so must the event  $f(A) = B \cap g(A) = C$ . Thus the event  $(f(A), g(A)) = (B, C)$  occurs with probability 1. Therefore,  $\nu \models A \twoheadrightarrow (B, C)$ .
2. The same ideas, but faster: we have  $f : \mathcal{V}(A) \rightarrow \mathcal{V}(B)$  as before, and  $g : \mathcal{V}(B) \rightarrow \mathcal{V}(C)$ , such that the events  $f(A) = B$  and  $g(B) = C$  occur with probability 1. By the same logic, it follows that their conjunction holds with probability 1, and hence  $C = f(g(A))$  occurs with probability 1. So  $\nu \models A \twoheadrightarrow C$ . □

**Theorem 5.1.** *If  $G$  is a directed acyclic graph and  $\mathcal{I}(G)$  consists of the independencies of its corresponding Bayesian network, then  $\mu \models \Diamond \mathcal{A}_G$  if and only if  $\mu$  satisfies  $\mathcal{I}(G)$ .*

*Proof.* Label the vertices of  $G = (\mathcal{N}, E)$  by natural numbers so that they are a topological sort of  $G$ —that is, without loss of generality, suppose  $\mathcal{N} = [n] := \{1, 2, \dots, n\}$ , and  $i < j$  whenever  $i \rightarrow j \in E$ . By the definition of  $\mathcal{A}_G$ , the arcs  $\mathcal{A}_G = \{S_i \xrightarrow{i} i\}_{i=1}^n$  are also indexed by integers. Finally, write  $\mathcal{X} = (X_1, \dots, X_n)$  for the variables  $\mathcal{X}$  corresponding to  $\mathcal{N}$  over which  $\mu$  is defined.

( $\implies$ ). Suppose  $\mu \models \mathcal{A}_G$ . This means there is an extension of  $\bar{\mu}(\mathcal{X}, \mathcal{U})$  of  $\mu(\mathcal{X})$  to additional independent variables  $\mathcal{U} = (U_1, \dots, U_n)$ , such that  $\bar{\mu} \models (S_i, U_i) \twoheadrightarrow i$  for all  $i \in [n]$ .

First, we claim that if  $\bar{\mu}$  is such a witness, then  $\bar{\mu} \models (U_1, \dots, U_k) \twoheadrightarrow (X_1, \dots, X_k)$  for all  $k \in [n]$ , and so in particular,  $\bar{\mu} \models \mathcal{U} \twoheadrightarrow \mathcal{X}$ . This follows from QIM-compatibility's condition (c) and the fact that  $G$  is acyclic, by induction. In more

detail: The base case of  $k = 0$  holds vacuously. Suppose that  $\bar{\mu} \models (X_1, \dots, X_k)$  for some  $k < n$ . Now, condition (c) of [Definition 5.1](#) says  $\bar{\mu} \models (S_{k+1}, U_{k+1}) \rightarrow\!\!\!\rightarrow X_{k+1}$ . Because the variables are sorted in topological order, the parent variables  $S_{k+1}$  are a subset of  $\{X_1, \dots, X_n\}$ , which are determined by  $\mathcal{U}$  by the induction hypothesis; at the same time clearly  $\bar{\mu} \models (U_1, \dots, U_{k+1}) \rightarrow\!\!\!\rightarrow U_{k+1}$  as well. So, by two instances of [Lemma 5.11](#),  $\bar{\mu} \models (U_1, \dots, U_{k+1}) \rightarrow\!\!\!\rightarrow X_{k+1}$ . Combining with our inductive hypothesis, we find that  $\bar{\mu} \models (U_1, \dots, U_{k+1}) \rightarrow\!\!\!\rightarrow (X_1, \dots, X_{k+1})$ . So, by induction,  $\bar{\mu} \models (U_1, \dots, U_k) \rightarrow\!\!\!\rightarrow (X_1, \dots, X_k)$  for  $k \in [n]$ , and in particular,  $\bar{\mu} \models \mathcal{U} \rightarrow\!\!\!\rightarrow \mathcal{X}$ .

With this in mind, we now return to proving that  $\mu$  has the required independencies. It suffices to show that  $\mu(\mathcal{X}) = \prod_{i=1}^n \mu(X_i \mid S_i)$ . We do so by showing that, for all  $k \in [n]$ ,  $\mu(X_1, \dots, X_k) = \mu(X_1, \dots, X_{k-1})\mu(X_k \mid S_k)$ . By QIM-compatibility witness condition (c), we know that  $\bar{\mu} \models (S_k, U_k) \rightarrow\!\!\!\rightarrow X_k$ , and so there exists a function  $f_k : \mathcal{V}(S_k) \times \mathcal{V}(U_k) \rightarrow \mathcal{V}(X_k)$  for which the event  $f_k(S_k, U_k) = X_k$  occurs with probability 1. Since  $\bar{\mu} \models (U_1, \dots, U_{k-1}) \rightarrow\!\!\!\rightarrow (X_1, \dots, X_{k-1})$ , and  $U_k$  is independent of  $(U_1, \dots, U_{k-1})$ , it follows from [Lemma 5.10](#) that  $\bar{\mu} \models (X_1, \dots, X_{k-1}) \perp\!\!\!\perp U_k$ . Thus

$$\begin{aligned}\mu(X_1, \dots, X_{k-1}, X_k) &= \sum_{u \in \mathcal{V}(U_k)} \mu(X_1, \dots, X_{k-1})\bar{\mu}(U_k = u) \cdot \mathbb{1}[X_k = f_k(S_k, u)] \\ &= \mu(X_1, \dots, X_{k-1}) \sum_{u \in \mathcal{V}(U_k)} \bar{\mu}(U_k = u) \cdot \mathbb{1}[X_k = f_k(S_k, u)]\end{aligned}$$

Observe that the quantity on the right, including the sum, is a function of  $X_k$  and  $S_k$ , but no other variables; let  $\varphi(X_k, S_k)$  denote this quantity. Because  $\mu$  is a probability distribution, know that  $\varphi(X_k, S_k)$  must be the conditional probability of  $X_k$  given  $X_1, \dots, X_{k-1}$ , and it depends only on the variables  $S_k$ . Thus  $\mu(X_1, \dots, X_k) = \mu(X_1, \dots, X_{k-1})\mu(X_k \mid S_k)$ .

Therefore  $\nu(\mathcal{X}) = \mu(\mathcal{X})$  factors as required by the BN  $G$ , meaning that  $\mu$  has

the independencies specified by  $G$ . (See Koller & Friedman Thm 3.2, for instance.)

( $\Leftarrow$ ). Suppose  $\mu$  satisfies the independencies of  $G$ , meaning that each node is conditionally independent of its non-descendants given its parents. We now repeatedly apply the construction [Section 5.A.1](#) to construct a QIM-compatibility witness. Specifically, for  $k \in \{1, \dots, n\}$ , let  $U_k$  be a variable whose values  $\mathcal{V}(U_k) := \mathcal{V}(X_k)^{\mathcal{V}(S_k)}$  are functions from values of  $X_k$ 's parents, to values of  $X_k$ . Let  $\mathcal{U}$  denote the joint variable  $(U_1, \dots, U_n)$ , and observe that a setting  $\mathbf{g} = (g_1, \dots, g_n)$  of  $\mathcal{U}$  uniquely picks out a value of  $\mathcal{X}$ , by evaluating each function in order. Let's call this function  $f : \mathcal{V}(\mathcal{U}) \rightarrow \mathcal{V}(\mathcal{X})$ .

To be more precise, we now construct  $f(\mathbf{g})$  inductively. The first component we must produce is  $X_1$ , but since  $X_1$  has no parents,  $g_1$  effectively describes a single value of  $X_1$ , so we define the first component  $f(\mathbf{g})[X_1]$  to be that value. More generally, assuming that we have already defined the components  $X_1, \dots, X_{i-1}$ , among which are the variables  $S_k$  on which  $X_i$  depends, we can determine the value of  $X_i$ ; formally, this means defining

$$f(\mathbf{g})[X_i] := g_i(f(\mathbf{g})[S_i]),$$

which, by our inductive assumption, is well-defined. Note that, for all  $\mathbf{g} \in \mathcal{V}(\mathcal{U})$  and  $\mathbf{x} \in \mathcal{V}(\mathcal{X})$ , the function  $f$  is characterized by the property

$$f(\mathbf{g}) = \mathbf{x} \iff \bigwedge_{i=1}^n g_i(\mathbf{x}[S_i]) = \mathbf{x}[X_i]. \quad (5.4)$$

To quickly verify this: if  $f(\mathbf{g}) = \mathbf{x}$ , then in particular, for  $i \in [n]$ , then  $\mathbf{x}[X_i] = f(\mathbf{g})[X_i] = g_i(\mathbf{x}[S_i])$  by the definition above. Conversely, if the right-hand side of (5.4) holds, then we can prove  $f(\mathbf{g}) = \mathbf{x}$  by induction over our construction of  $f$ : if  $f(\mathbf{g})[X_j] = \mathbf{x}[X_j]$  for all  $j < i$ , then  $f(\mathbf{g})[X_i] = g_i(f(\mathbf{g})[S_i]) = g_i(\mathbf{x}[S_i]) = \mathbf{x}[X_i]$ .

Next, we define an unconditional probability over each  $U_k$  according to

$$\bar{\mu}_i(U_i = g) := \prod_{\mathbf{s} \in \mathcal{V}(S_k)} \mu(X_i = g(s) \mid S_i = \mathbf{s}),$$

which, as verified in [Section 5.A.1](#), is indeed a conditional probability, and has the property that  $\bar{\mu}_i(U_i(\mathbf{s}) = x) = \mu(X_i = x \mid S_i = \mathbf{s})$  for all  $x \in \mathcal{V}(X_i)$  and  $\mathbf{s} \in \mathcal{V}(S_i)$ . By taking an independent combination (tensor product) of each of these unconditional distributions, we obtain a joint distribution  $\bar{\mu}(\mathcal{U}) = \prod_{i=1}^n \bar{\mu}_i(U_i)$ . Finally, we extend this distribution to a full joint distribution  $\bar{\mu}(\mathcal{U}, \mathcal{X})$  via the pushforward of  $\bar{\mu}(\mathcal{U})$  through the function  $f$  defined by induction above. In this distribution, each  $X_i$  is determined by  $U_i$  and  $S_i$ .

By construction, the variables  $\mathcal{U}$  are mutually independent (for [Definition 5.1\(b\)](#)), and satisfy  $(S_k, U_k) \twoheadrightarrow X_k$  for all  $k \in [n]$  ([Definition 5.1\(c\)](#)). It remains only to verify that the marginal of  $\bar{\mu}$  on the variables  $\mathcal{X}$  is the original distribution  $\mu$  ([Definition 5.1\(a\)](#)). Here is where we rely on the fact that  $\mu$  satisfies the independencies of  $G$ , which means that we can factor  $\mu(\mathcal{X})$  as

$$\mu(\mathcal{X}) = \prod_{i=1}^n \mu(X_i \mid S_i).$$

$$\begin{aligned}
\bar{\mu}(\mathcal{X}=\mathbf{x}) &= \sum_{\mathbf{g} \in \mathcal{V}(\mathcal{U})} \bar{\mu}(\mathcal{U}=\mathbf{g}) \cdot \delta f(\mathbf{x} \mid \mathbf{g}) \\
&= \sum_{(g_1, \dots, g_n) \in \mathcal{V}(\mathcal{U})} \mathbb{1}[\mathbf{x} = f(\mathbf{g})] \prod_{i=1}^n \bar{\mu}(U_i=g_i) \\
&= \sum_{(g_1, \dots, g_n) \in \mathcal{V}(\mathcal{U})} \mathbb{1}\left[\bigwedge_{i=1}^n g_i(\mathbf{x}[S_i]) = \mathbf{x}[X_i]\right] \prod_{i=1}^n \bar{\mu}(U_i=g_i) \quad [\text{by (5.4)}] \\
&= \prod_{i=1}^n \sum_{g \in \mathcal{V}(U_i)} \mathbb{1}[g(\mathbf{x}[S_i]) = \mathbf{x}[X_i]] \cdot \bar{\mu}(U_i = g) \\
&= \prod_{i=1}^n \bar{\mu}\left(\left\{g \in \mathcal{V}(U_i) \mid g(\mathbf{s}_i) = x_i\right\}\right) \quad \text{where } \begin{array}{l} x_i := \mathbf{x}[X_i], \\ \mathbf{s}_i := \mathbf{x}[S_i] \end{array} \\
&= \prod_{i=1}^n \bar{\mu}(U_i(\mathbf{s}_i) = x_i) \\
&= \prod_{i=1}^n \mu(X_i = x_i \mid S_i = \mathbf{s}_i) \\
&= \mu(\mathcal{X} = \mathbf{x}).
\end{aligned}$$

Therefore, when  $\mu$  satisfies the independencies of a BN  $G$ , it is QIM-compatible with  $\mathcal{A}_G$ .  $\square$

Before we move on to proving the other results in the chapter, we first illustrate how this relatively substantial first half of the proof of [Theorem 5.1](#) can be dramatically simplified by relying on two information-theoretic arguments.

*Alternate, information-based proof.* ( $\implies$ ). Let  $G$  be a dag. If  $\mu \models \mathcal{A}_G$ , then by [Theorem 5.7](#),  $SDef_{\mathcal{A}_G}(\mu) \leq 0$ . In the proof of [Theorem 3.5](#), it is shown that  $SDef_{\mathcal{A}_G}(\mu) \geq 0$  with equality iff  $\mu$  satisfies the BN's independencies. Thus  $\mu$  must satisfy the appropriate independencies.  $\square$

### Theorem 5.2.

- (a)  $\mu \models X \rightarrow\!\!\!\rightarrow Y \wedge \Diamond \mathcal{A}$  if and only if  $\forall n \geq 0. \mu \models \Diamond \mathcal{A} \sqcup_{X \rightarrow Y}^{(+n)}$ .
- (b) if  $\mathcal{A} = \mathcal{A}_G$  for a dag  $G$ , then  $\mu \models X \rightarrow\!\!\!\rightarrow Y \wedge \Diamond \mathcal{A}$  if and only if  $\mu \models \Diamond \mathcal{A} \sqcup_{X \rightarrow Y}^{(+1)}$ .
- (c) if  $\exists a \in \mathcal{A}$  such that  $S_a = \emptyset$  and  $X \in T_a$ , then  $\mu \models X \rightarrow\!\!\!\rightarrow Y \wedge \Diamond \mathcal{A}$  iff  $\mu \models \Diamond \mathcal{A} \sqcup_{X \rightarrow Y}^{(+2)}$ .

*Proof.* (a). The forward direction is straightforward. Suppose that  $\mu \models \mathcal{A}$  and  $\mu \models X \rightarrow\!\!\!\rightarrow Y$ . The former condition gives us a witness  $\nu(\mathcal{X}, \mathcal{U})$  in which  $\mathcal{U} = \{U_a\}_{a \in \mathcal{A}}$  are mutually independent variables indexed by  $\mathcal{A}$ , that determine their respective edges. “Extend”  $\nu$  in the unique way to  $n$  additional constant variables  $U_1, \dots, U_n$ , each of which can only take on one value. We claim that this “extended” distribution  $\nu'$ , which we conflate with  $\nu$  because it is not meaningfully different, is a witness to  $\mu \models \Diamond \mathcal{A} \sqcup_{X \rightarrow Y}^{(+n)}$ . Since  $\mu \models X \rightarrow\!\!\!\rightarrow Y$  it must also be that  $\nu \models X \rightarrow\!\!\!\rightarrow Y$ , and it follows that  $\nu \models (X, U_i) \rightarrow\!\!\!\rightarrow Y$  for all  $i \in \{1, \dots, n\}$ , demonstrating that the new requirements of  $\nu'$  imposed by Definition 5.1(c) hold. (The remainder of the requirements for condition (c), namely that  $\nu' \models (S_a, U_a) \rightarrow\!\!\!\rightarrow T_a$  for  $a \in \mathcal{A}$ , still hold because  $\nu'$  is an extension of  $\nu$ , which we know has this property.) Finally, since  $\mathcal{U}$  are mutually independent and each  $U_i$  is a constant (and hence independent of everything), the variables  $\mathcal{U}' := \mathcal{U} \sqcup \{U_i\}_{i=1}^n$  are also mutually independent. Thus  $\nu$  (or, more precisely, an isomorphic “extension” of it to additional trivial variables) is a witness of  $\mu \models \Diamond \mathcal{A} \sqcup_{X \rightarrow Y}^{(+n)}$ .

The reverse direction is difficult to prove directly, yet it is a straightforward application of Theorem 5.7. Suppose that  $\mu \models \mathcal{A} \sqcup_{X \rightarrow Y}^{(+n)}$  for all  $n \geq 0$ . By

Theorem 5.7, we know that

$$0 \geq SDef_{\mathcal{A} \sqcup \overset{(+n)}{X \rightarrow Y}}(\mu) = SDef_{\mathcal{A}}(\mu) + n H_\mu(Y|X).$$

Because  $SDef_{Ar}(\mu)$  is bounded below (by  $-\log |\mathcal{V}(\mathcal{X})|$ ), it cannot be the case that  $H_\mu(Y|X) > 0$ ; otherwise, the inequality above would not hold for large  $n$  (specifically, for  $n > \log |\mathcal{V}(\mathcal{X})|/H_\mu(Y|X)$ ). By Gibbs inequality,  $H_\mu(Y|X)$  is non-negative, and thus it must be the case that  $H_\mu(Y|X) = 0$ . Thus  $\mu \models X \twoheadrightarrow Y$ . It is also true that  $\mu \models \diamond \mathcal{A}$  by monotonicity (Theorem 10.2), which is itself a direct application of Theorem 5.7

**(b).** Now  $\mathcal{A} = \mathcal{A}_G$  for some graph  $G$ . The forward direction of the equivalence is strictly weaker than the one we already proved in part (a); we have shown  $\mu \models \diamond \mathcal{A} \sqcup \overset{(+n)}{X \rightarrow Y}$  for all  $n \geq 0$ , and needed only to show it for  $n = 1$ . The reverse direction is what's interesting. As before, we will take a significant shortcut by using Theorem 5.7. Suppose  $\mu \models \diamond \mathcal{A} \sqcup \overset{(+1)}{X \rightarrow Y}$ . In this case where  $\mathcal{A} = \mathcal{A}_G$ , we have already shown (in the proof of Theorem 3.5) that  $SDef_{\mathcal{A}}(\mu) \geq 0$ . It follows that

$$0 \stackrel{\text{(Theorem 5.7)}}{\geq} SDef_{\mathcal{A} \sqcup \overset{(+n)}{X \rightarrow Y}}(\mu) = SDef_{\mathcal{A}}(\mu) + H_\mu(Y|X) \geq 0,$$

and thus  $H_\mu(Y|X) = 0$ , meaning that  $\mu \models X \twoheadrightarrow Y$  as promised. As before, we also have  $\mu \models \diamond \mathcal{A}$  by monotonicity.

**(c).** As in part (b), the forward direction is a special case of the forward direction of part (a), and it remains only to prove the reverse direction. Equipped with the additional information that  $\mathcal{A} \rightsquigarrow \{\rightarrow \{X\}\}$ , suppose that  $\mu \models \diamond \mathcal{A} \sqcup \overset{(+2)}{X \rightarrow Y}$ . By monotonicity, this means  $\mu \models \mathcal{A}$  and also that  $\mu \models \rightarrow \boxed{X} \not\rightarrow \boxed{Y}$ . Let  $\mathcal{A}'$  denote this hypergraph. Once again by appeal to Theorem 5.7, we have that

$$0 \geq SDef_{\mathcal{A}'} = -H_\mu(X, Y) + H(X) + 2H_\mu(Y|X) = H_\mu(Y|X) \geq 0.$$

It follows that  $H_\mu(Y|X) = 0$ , and thus  $\mu \models X \rightarrow\!\!\!\rightarrow Y$ . As mentioned above, we also know that  $\mu \models \mathcal{A}$ , and thus  $\mu \models \Diamond \mathcal{A} \wedge X \rightarrow\!\!\!\rightarrow Y$  as promised.  $\square$

### 5.A.3 Causality Results of Section 5.3

**Proposition 5.3.** *Given a graph  $G$  and a distribution  $\mu$ ,  $\mu \models \Diamond \mathcal{A}_G$  iff there exists a fully randomized PSEM of dependency structure  $\mathcal{A}_G$  from which  $\mu$  can arise.*

*Proof.* ( $\implies$ ). Suppose  $\mu \models \mathcal{A}_G$ . Thus there exists some witness  $\bar{\mu}(\mathcal{X}, \mathcal{U})$  to this fact, satisfying conditions (a-c) of Definition 5.1. Because  $\mathcal{A}_G$  is partitional, the elements of  $\text{PSEMs}_{\mathcal{A}_G}(\bar{\mu})$  are ordinary (i.e., not generalized) randomized PSEMs. We claim that every  $\mathcal{M} = (M, P) \in \text{PSEMs}_{\mathcal{A}_G}(\bar{\mu})$  that is a randomized PSEM from which  $\mu$  can arise, and also has the property that  $\text{Pa}_M(Y) \subseteq \text{Pa}_G(Y) \cup \{U_Y\}$  for all  $Y \in \mathcal{X}$ .

- The hyperarcs of  $\mathcal{A}_G$  correspond to the vertices of  $G$ , which in turn correspond to the variables in  $\mathcal{X}$ ; thus  $\mathcal{U} = \{U_X\}_{X \in \mathcal{X}}$ . By property (b) of QIM-compatibility witnesses (Definition 5.1), these variables  $\{U_X\}_{X \in \mathcal{X}}$  are mutually independent according to  $\bar{\mu}$ . Furthermore, because  $\mathcal{M} = (M, P) \in \text{PSEMs}_{\mathcal{A}_G}(\bar{\mu})$ , we know that  $\bar{\mu}(\mathcal{U}) = P$ , and thus the variables in  $\mathcal{U}$  must be mutually independent according to  $P$ . By construction, in causal models  $\mathcal{M} \in \text{PSEMs}_{\mathcal{A}_G}(\bar{\mu})$  the equation  $f_Y$  can depend only on  $S_Y = \text{Pa}_G(Y) \subseteq \mathcal{X}$  and  $U_Y$ . So, in particular,  $f_Y$  does not depend on  $U_X$  for  $X \neq Y$ .

Altogether, we have shown that  $\mathcal{M}$  contains exogenous variables  $\{U_X\}_{X \in \mathcal{X}}$  that are mutually independent according to  $P$ , and that  $f_Y$  does not depend on  $U_X$  when  $X \neq Y$ . Thus,  $\mathcal{M}$  is a randomized PSEM.

- By condition (a) on QIM-compatibility witnesses (Definition 5.1), we know that  $\bar{\mu}(\mathcal{X}) = \mu$ . By Proposition 5.5(a), we know that  $\mu \in \{\mathcal{M}\}$ . Together, the previous two sentences mean that  $\mu$  can arise from  $\mathcal{M}$ .
- Finally, as mentioned in the first bullet item, the equation  $f_Y$  in  $M$  can depend only on  $S_Y = \text{Pa}_G(Y)$  and on  $U_Y$ . Thus  $\text{Pa}_M(Y) \subseteq \text{Pa}_G(Y) \cup \{U_Y\}$  for all  $Y \in \mathcal{X}$ .

Under the assumption that  $\mu \models \mathcal{A}_G$ , we have now shown that there exists a randomized causal model  $\mathcal{M}$  from which  $\mu$  can arise, with the property that  $\text{Pa}_{\mathcal{M}}(Y) \subseteq \text{Pa}_G(Y) \cup \{U_Y\}$  for all  $Y \in \mathcal{X}$ .

( $\Leftarrow$ ). Conversely, suppose there is a randomized PSEM  $\mathcal{M} = (M = (\mathcal{Y}, \mathcal{U}, \mathcal{F}), P)$  with the property that  $\text{Pa}_M(Y) \subseteq \text{Pa}_G(Y) \cup \{U_Y\}$  for all  $Y$ , from which  $\mu$  can arise. The last clause means there exists some  $\nu \in \{\mathcal{M}\}$  such that  $\nu(\mathcal{X}) = \mu$ . We claim that this  $\nu$  is a witness to  $\mu \models \mathcal{A}_G$ . We already know that condition (a) of being a QIM-compatibility witness is satisfied, since  $\nu(\mathcal{X}) = \mu$ . Condition (b) holds because of the assumption that  $\{U_X\}_{X \in \mathcal{X}}$  are mutually independent in the distribution  $P$  for a randomized PSEM (and the fact that  $\nu(\mathcal{U}) = P$ , since  $\nu \in \{\mathcal{M}\}$ ). Finally, we must show that (c) for each  $Y \in \mathcal{X}$ ,  $\nu \models \text{Pa}_G(Y) \cup \{U_Y\} \rightarrowtail Y$ . Since  $\nu \in \{\mathcal{M}\}$ , we know that  $M$ 's equation holds with probability 1 in  $\nu$ , and so it must be the case that  $\nu \models \text{Pa}_M(Y) \rightarrowtail Y$ . Note that, in general, if  $A \subseteq B$  and  $A \rightarrowtail C$ , then  $B \rightarrowtail C$ . By assumption,  $\text{Pa}_M(Y) \subseteq \text{Pa}_G(Y) \cup \{U_Y\}$ , and thus  $\nu \models \text{Pa}_G(Y) \cup \{U_Y\} \rightarrowtail Y$ .

Thus  $\nu$  satisfies all conditions (a-c) for a QIM-compatibility witness, and hence  $\mu \models \mathcal{A}_G$ . □

**Proposition 5.4.**  $\mu \models \Diamond \mathcal{A}$  iff there exists a generalized randomized PSEM with

structure  $\mathcal{A}$  from which  $\mu$  can arise.

*Proof.* ( $\implies$ ). Suppose  $\mu \models \mathcal{A}$ , meaning there exists a witness  $\nu(\mathcal{X}, \mathcal{U})$  with property [Definition 5.1\(c\)](#), meaning that, for all  $a \in \mathcal{A}$ , there is a functional dependence  $(S_a, U_a) \twoheadrightarrow T_a$ . Thus, there is some set of functions  $\mathcal{F}$  with these types that holds with probability 1 according to  $\nu$ . Meanwhile, by [Definition 5.1\(b\)](#),  $\nu(\mathcal{U})$  are mutually independent, so defining  $P_a(U_a) := \nu(U_a)$ , we have  $\nu(\mathcal{U}) = \prod_{a \in \mathcal{A}} P_a(U_a)$ . Together, the previous two conditions (non-deterministically) define a generalized randomized PSEM  $\mathcal{M}$  of shape  $\mathcal{A}$  for which  $\nu \in \{\mathcal{M}\}$ . Finally, by [Definition 5.1\(a\)](#), we know that  $\mu$  can arise from  $\mathcal{M}$ .

( $\Leftarrow$ ). Conversely, suppose there is a generalized randomized SEM  $\mathcal{M}$  of shape  $\mathcal{A}$  from which  $\mu(\mathcal{X})$  can arise. Thus, there is some  $\nu \in \{\mathcal{M}\}$  whose marginal on  $\mathcal{X}$  is  $\mu$ . We claim that this  $\nu$  is also a witness that  $\mu \models \mathcal{A}$ . The marginal constraint from [Definition 5.1\(a\)](#) is clearly satisfied. Condition (b) is immediate as well, because  $\nu(\mathcal{U}) = \prod_a P_a(U_a)$ . Finally, condition (c) is satisfied, because the equations of  $\mathcal{M}$  hold with probability 1, ensuring the appropriate functional dependencies.  $\square$

**Proposition 5.5.** *If  $\bar{\mu}(\mathcal{X}, \mathcal{U}_{\mathcal{A}})$  is a witness for QIM-compatibility with  $\mathcal{A}$  and  $\mathcal{M}$  is a PSEM with dependency structure  $\mathcal{A}$ , then  $\bar{\mu} \in \{\mathcal{M}\}$  if and only if  $\mathcal{M} \in \text{PSEMs}_{\mathcal{A}}(\bar{\mu})$ .*

*Proof.* (a) is straightforward. Suppose  $\mathcal{M} \in \text{PSEMs}(\nu)$ . By construction, the equations of  $\mathcal{M}$  reflect functional dependencies in  $\nu$ , and hence hold with probability 1.<sup>7</sup> Furthermore, the distribution  $P(\mathcal{U})$  in all  $\mathcal{M} \in \text{PSEMs}(\nu)$  is equal to  $\nu(\mathcal{U})$ .

---

<sup>7</sup>When the probability of some combination of source variables is zero, there is typically more than one choice of functions that holds with probability 1; the choice of functions is essentially the choice of  $\mathcal{M} \in \text{PSEMs}(\nu)$ .

These two facts, demonstrate that  $\nu$  satisfies the two constraints required for membership in  $\{\mathcal{M}\}$ .

(b). We do the two directions separately. First, suppose  $\mathcal{M} \in \text{PSEMs}(\nu)$ . We have already shown (in part (a)) that  $\nu \in \{\mathcal{M}\}$ . The construction of  $\text{PSEMs}(\nu)$  depends on the hypergraph  $\mathcal{A}$  (even if the dependence is not explicitly clear from our notation) in such a way that  $f_X$  does not depend on any variables beyond  $U_a$  and  $S_{a_X}$ . Thus,  $\text{Pa}_{\mathcal{M}}(X) \subseteq S_{a_X} \cup \{U_{a_X}\}$ .

Conversely, suppose  $\mathcal{M} = (\mathcal{X}, \mathcal{U}, \mathcal{F})$  is a PSEM satisfying  $\nu \in \{\mathcal{M}\}$  and  $\text{Pa}_{\mathcal{M}}(X) \subseteq S_{a_X} \cup \{U_{a_X}\}$ . We would like to show that  $\mathcal{M} \in \text{PSEMs}(\nu)$ . Because  $\nu \in \{\mathcal{M}\}$ , we know that the distribution  $P(\mathcal{U})$  over the exogenous variables in the PSEM  $\mathcal{M}$  is equal to  $\nu(\mathcal{U})$ , matching the first part of our construction. What remains is to show that the equations  $\mathcal{F}$  are consistent with our transformation. Choose any  $X \in \mathcal{X}$ . Because  $\mathcal{A}$  is subpartitional, there is a unique  $a_X \in \mathcal{A}$  such that  $X \in T_{a_X}$ . Now choose any values  $s \in \mathcal{V}(S_{a_X})$  and  $u \in \mathcal{V}(U_{a_X})$ . If  $\nu(s, u) > 0$ , then we know there is a unique value of  $x \in \mathcal{V}(X)$  such that  $\nu(s, u, x) > 0$ . Since  $\mathcal{M}$ 's equation for  $X$ ,  $f_X$ , depends only on  $s$  and  $u$ , and holds with probability 1, we know that  $f_X(s, u) = t$ , as required. On the other hand, if  $\nu(s, u) = 0$ , then any choice of  $f_X(s, u)$  is consistent with our procedure. Since this is true for all  $X$ , and all possible inputs to the equation  $f_X$ , we conclude that the equations  $\mathcal{F}$  can arise from the procedure described in the main text, and therefore  $\mathcal{M} \in \text{PSEMs}(\nu)$ .  $\square$

**Theorem 5.6.** Suppose that  $\bar{\mu}$  is a witness to  $\mu \models \Diamond \mathcal{A}$ ,  $\mathcal{M} \in \text{PSEMs}_{\mathcal{A}}(\bar{\mu})$ ,  $\mathbf{X} \subseteq \mathcal{X}$  and  $\mathbf{x} \in \mathcal{V}(\mathbf{X})$ . If  $\bar{\mu}(\text{do}_{\mathcal{M}}(\mathbf{X}=\mathbf{x})) > 0$ , then:

(a)  $\bar{\mu}(\mathcal{X} \mid \text{do}_{\mathcal{M}}(\mathbf{X}=\mathbf{x}))$  can arise from  $\mathcal{M}_{\mathbf{X} \leftarrow \mathbf{x}}$ ;

(b) for all events  $\varphi \subseteq \mathcal{V}(\mathcal{X})$ ,  $\Pr_{\mathcal{M}}([\mathbf{X} \leftarrow \mathbf{x}] \varphi) \leq \bar{\mu}(\varphi \mid \text{do}_{\mathcal{M}}(\mathbf{X}=\mathbf{x})) \leq$

$$\Pr_{\mathcal{M}} (\langle \mathbf{X} \leftarrow \mathbf{x} \rangle \varphi)$$

and all three are equal when  $\mathcal{M} \models \mathcal{U} \twoheadrightarrow \mathcal{X}$  (such as when  $\mathcal{M}$  is acyclic).

*Proof.* **(part a).** Let  $(M, P) := \mathcal{M}$  be the SEM and probability over exogenous variable in the PSEM  $\mathcal{M}$ , and  $\mathcal{F} = \{f_Y\}_{Y \in \mathcal{X}}$  be its set of equations. Because we have assumed  $\nu(\text{do}_M(\mathbf{X}=\mathbf{x})) > 0$ , the conditional distribution

$$\nu \mid \text{do}_M(\mathbf{X}=\mathbf{x}) = \nu(\mathcal{U}, \mathcal{X}) \cdot \prod_{X \in \mathbf{X}} \mathbb{1}[\forall \mathbf{s}. f_X(U_X, \mathbf{s}) = \mathbf{x}[X]] \Big/ \nu(\text{do}_M(\mathbf{X}=\mathbf{x}))$$

is defined. By assumption,  $\mathcal{M} \in \text{PSEMs}(\nu)$  and  $\nu$  is a witness to  $\mu \models \mathcal{A}$ . Thus, by Proposition 5.5, we know that  $\nu \in \{\mathcal{M}\}$ . So in particular, all equations of  $\mathcal{M}$  hold for all joint settings  $(\mathbf{u}, \omega) \in \mathcal{V}(\mathcal{X} \cup \mathcal{U})$  in the support of  $\nu$ . But the support of the conditional distribution  $\nu \mid \text{do}_M(\mathbf{X}=\mathbf{x})$  is a subset of the support of  $\nu$ , so all equations of  $\mathcal{M}$  also hold in the conditioned distribution. Furthermore, the event  $\text{do}_M(\mathbf{X}=\mathbf{x})$  is the event in which, for all  $X \in \mathbf{X}$ , the variable  $U_X$  takes on a value such that  $f_X(\dots, U_X, \dots) = \mathbf{x}[X]$ . Thus the equations corresponding to  $\mathbf{X} = \mathbf{x}$  also hold with probability 1 in  $\nu \mid \text{do}_M(\mathbf{X}=\mathbf{x})$ .

This shows that all equations of  $\mathcal{M}_{\mathbf{X} \leftarrow \mathbf{x}}$  hold with probability 1 in  $\nu \mid \text{do}_M(\mathbf{X}=\mathbf{x})$ . However, the marginal distribution  $\nu(\mathcal{U} \mid \text{do}_M(\mathbf{X}=\mathbf{x}))$  over  $\mathcal{U}$  will typically not be the distribution  $P(\mathcal{U})$ —indeed, we have altered collapsed distribution of the variables  $\mathcal{U}_{\mathbf{X}} := \{U_X : X \in \mathbf{X}\}$ . So, strictly speaking,  $\nu \mid \text{do}_M(\mathbf{X}=\mathbf{x}) \notin \{\mathcal{M}_{\mathbf{X} \leftarrow \mathbf{x}}\}$ . Our objective, therefore, is to show that there is a *different* distribution  $\nu' \in \{\mathcal{M}_{\mathbf{X} \leftarrow \mathbf{x}}\}$  such that  $\nu'(\mathcal{X}) = \nu(\mathcal{X} \mid \text{do}_M(\mathbf{X}=\mathbf{x}))$ . Let  $\mathbf{Z} := \mathcal{X} \setminus \mathbf{X}$ , and  $\mathcal{U}_{\mathbf{Z}} := \{U_Z : Z \in \mathbf{Z}\}$ . We can define  $\nu'$  according to

$$\nu'(\mathcal{X}, \mathcal{U}_{\mathbf{X}}, \mathcal{U}_{\mathbf{Z}}) := \nu(\mathcal{X}, \mathcal{U}_{\mathbf{Z}} \mid \text{do}_M(\mathbf{X}=\mathbf{x})) P(\mathcal{U}_{\mathbf{X}}).$$

This distribution satisfies three critical properties:

1. Clearly  $\nu'$  has the appropriate marginal  $\nu'(\mathcal{X}) = \nu(\mathcal{X} \mid \text{do}_M(\mathbf{X}=\mathbf{x}))$  on exogenous variables  $\mathcal{X}$ , by construction.
2. At the same time, the marginal on exogenous variables is

$$\begin{aligned}
\nu'(\mathcal{U}) &= \nu'(\mathcal{U}_{\mathbf{X}}, \mathcal{U}_{\mathbf{Z}}) \\
&= \int_{\mathcal{V}(\mathcal{X})} \nu(\omega, \mathcal{U}_{\mathbf{Z}} \mid \text{do}_M(\mathbf{X}=\mathbf{x})) P(\mathcal{U}_{\mathbf{X}}) d\omega \\
&= P(\mathcal{U}_{\mathbf{X}}) \nu(\mathcal{U}_{\mathbf{Z}} \mid \text{do}_M(\mathbf{X}=\mathbf{x})) \\
&= P(\mathcal{U}_{\mathbf{X}}) P(\mathcal{U}_{\mathbf{Z}} \mid \text{do}_M(\mathbf{X}=\mathbf{x})) \quad [\text{since } \text{do}_M(\mathbf{X}=\mathbf{x}) \text{ depends only on } \mathcal{U}] \\
&= P(\mathcal{U}_{\mathbf{X}}) P(\mathcal{U}_{\mathbf{Z}}) \quad \left[ \begin{array}{l} \text{since } \text{do}_M(\mathbf{X}=\mathbf{x}) \text{ depends only} \\ \text{on } \mathcal{U}_{\mathbf{X}}, \text{ while } \mathcal{U}_{\mathbf{X}} \text{ and } \mathcal{U}_{\mathbf{Z}} \text{ are in-} \\ \text{dependent in } \nu \text{ (by the witness} \\ \text{condition).} \end{array} \right] \\
&= P(\mathcal{U}_{\mathbf{X}}, \mathcal{U}_{\mathbf{Z}}) \quad [\text{for the same reason as above}]. 
\end{aligned}$$

3. Finally,  $\nu'$  satisfies all equations of  $\mathcal{M}_{\mathbf{X} \leftarrow \mathbf{x}}$ . It satisfies the equations for the variables  $\mathbf{X}$  because  $\mathbf{X} = \mathbf{x}$  holds with probability 1. At the same time, the equations in  $\mathcal{M}_{\mathbf{X} \leftarrow \mathbf{x}}$  corresponding to the variables  $\mathbf{Z}$  hold with probability 1, because the marginal  $\nu'(\mathcal{U}_{\mathbf{Z}}, \mathcal{X})$  is shared with the distribution  $\nu \mid \text{do}_M(\mathbf{X}=\mathbf{x})$ —and that distribution satisfies these equations. (It suffices to show that they share this particular marginal because the equations for  $\mathbf{Z}$  do not depend on  $\mathcal{U}_{\mathbf{X}}$ .)

Together, items 2 and 3 show that  $\nu' \in \{\mathcal{M}_{\mathbf{X} \leftarrow \mathbf{x}}\}$ , and item 1 shows that  $\nu(\mathcal{X} \mid \text{do}_M(\mathbf{X}=\mathbf{x}))$  can arise from  $\mathcal{M}_{\mathbf{X} \leftarrow \mathbf{x}}$ .

**(part b).** We will again make use of the distribution  $\nu'$  defined in part (a), and its three critical properties listed above. Given a setting  $\mathbf{u} \in \mathcal{V}(\mathcal{U})$  of the

exogenous variables, let

$$\mathcal{F}_{\mathbf{X} \leftarrow \mathbf{x}}(\mathbf{u}) := \left\{ \omega \in \mathcal{V}(\mathcal{X}) \mid \begin{array}{ll} \forall X \in \mathbf{X}. & \omega[X] = \mathbf{x}[X] \\ \forall Y \in \mathcal{X} \setminus \mathbf{X}. & \omega[Y] = f_X(\omega[\mathcal{X} \setminus Y], \mathbf{u}) \end{array} \right\}$$

denote the set of joint settings of endogenous variables that are consistent with the equations of  $\mathcal{M}_{\mathbf{X} \leftarrow \mathbf{x}}$ .

If  $\mathbf{u} \in \mathcal{V}(\mathcal{U})$  is such that

$$\begin{aligned} (M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}] \varphi &\iff (M_{\mathbf{X} \leftarrow \mathbf{x}}, \mathbf{u}) \models \varphi \\ &\iff \forall \omega \in \mathcal{F}_{\mathbf{X} \leftarrow \mathbf{x}}(\mathbf{u}). \omega \in \varphi \\ &\iff \mathcal{F}_{\mathbf{X} \leftarrow \mathbf{x}}(\mathbf{u}) \subseteq \varphi, \end{aligned}$$

then  $\phi$  holds at all points that satisfy the equations of  $M_{\mathbf{X} \leftarrow \mathbf{x}}$ . So, since  $\nu'$  is supported only on such points (property 3), it must be that  $\nu'(\varphi) = 1$ . By property 1,  $\nu'(\varphi) = \nu(\varphi \mid \text{do}_M(\mathbf{X}=\mathbf{x}))$ .

Furthermore, if  $\nu'(\varphi) > 0$ , then there must exist some  $\omega \in \mathcal{F}_{\mathbf{X} \leftarrow \mathbf{x}}(\mathbf{u})$  satisfying  $\varphi$ , and thus  $(M, \mathbf{u}) \models \langle \mathbf{X} \leftarrow \mathbf{x} \rangle \varphi$ . Putting both of these observations together, and with a bit more care to the symbolic manipulation, we find that:

$$\begin{aligned} \Pr_{\mathcal{M}}([\mathbf{X} \leftarrow \mathbf{x}] \varphi) &= P(\{\mathbf{u} \in \mathcal{V}(\mathcal{U}) : (M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}] \varphi\}) \\ &= \sum_{\mathbf{u} \in \mathcal{V}(\mathcal{U})} P(\mathbf{u}) \mathbb{1}[\mathcal{F}_{\mathbf{X} \leftarrow \mathbf{x}}(\mathbf{u}) \subseteq \varphi] \\ &\leq \sum_{\mathbf{u} \in \mathcal{V}(\mathcal{U})} P(\mathbf{u}) \nu'(\varphi \mid \mathbf{u}) &= \nu'(\varphi) = \nu(\varphi \mid \text{do}_M(\mathbf{X}=\mathbf{x})) \\ &\leq \sum_{\mathbf{u} \in \mathcal{V}(\mathcal{U})} P(\mathbf{u}) \mathbb{1}[\mathcal{F}_{\mathbf{X} \leftarrow \mathbf{x}}(\mathbf{u}) \cap \varphi \neq \emptyset] \\ &= P(\{\mathbf{u} \in \mathcal{V}(\mathcal{U}) : (M, \mathbf{u}) \models \langle \mathbf{X} \leftarrow \mathbf{x} \rangle \varphi\}) \\ &= \Pr_{\mathcal{M}}(\langle \mathbf{X} \leftarrow \mathbf{x} \rangle \varphi), \quad \text{as desired.} \end{aligned}$$

Finally, if  $\nu \models \mathcal{U} \twoheadrightarrow \mathcal{X}$ , then  $\mathcal{F}_{\mathbf{X} \leftarrow \mathbf{x}}(\mathbf{u})$  is a singleton for all  $\mathbf{u}$ , and hence  $\varphi$  holding for all  $\omega \in \mathcal{F}_{\mathbf{X} \leftarrow \mathbf{x}}$  and for some  $\omega \in \mathcal{F}_{\mathbf{X} \leftarrow \mathbf{x}}$  are equivalent. So, in this case,

$$(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}] \varphi \iff (M, \mathbf{u}) \models \langle \mathbf{X} \leftarrow \mathbf{x} \rangle \varphi,$$

and thus the probability of both formulas are the same—and it must also equal  $\nu(\varphi \mid \text{do}_M(\mathbf{X} = \mathbf{x}))$  which we have shown lies between them.  $\square$

#### 5.A.4 Information Theoretic Results of Section 5.4

To prove [Theorem 5.7](#) and [Theorem 5.9\(a\)](#), we will need the following Lemma.

**Lemma 5.12.** *Consider a set of variables  $\mathbf{Y} = \{Y_1, \dots, Y_n\}$ , and another (set of) variable(s)  $X$ . Every joint distribution  $\mu(X, \mathbf{Y})$  over the values of  $X$  and  $\mathbf{Y}$  satisfies*

$$\sum_{i=1}^n I_\mu(X; Y_i) \leq I_\mu(X; \mathbf{Y}) + \sum_{i=1}^n H_\mu(Y_i) - H_\mu(\mathbf{Y}).$$

*Proof.* Since there is only one joint distribution in scope, we omit the subscript  $\mu$ , writing  $I(-)$  instead of  $I_\mu(-)$  and  $H(-)$  instead of  $H_\mu(-)$ , in the body of this proof. The following fact, known as the *chain rule for mutual information* will also be critical for our calculations:

$$I(A; B, C) = I(A; C) + I(A; B \mid C) \quad (\text{the MI chain rule}). \quad (5.5)$$

We prove this by induction on  $n$ . In the base case ( $n = 1$ ), we must show that  $I(X; Y) \leq I(X; Y) + H(Y) - H(Y)$ , which is an obvious tautology. Now, suppose inductively that

$$\sum_{i=1}^k I(X; Y_i) \leq I(X; \mathbf{Y}_{1:k}) + \sum_{i=1}^k H(Y_i) - H(\mathbf{Y}_{1:k}) \quad (\text{IH}_k)$$

for some  $k < n$ , where  $\mathbf{Y}_{1:k} = (Y_1, \dots, Y_k)$ . We now prove that the analogue for  $k + 1$  also holds. Some calculation reveals that

$$\begin{aligned}
& I(X; Y_{k+1}) \\
&= I(X; \mathbf{Y}_{1:k+1}) - I(X; \mathbf{Y}_{1:k} \mid Y_{k+1}) && \left[ \text{by MI chain rule (5.5)} \right] \\
&\leq I(X; \mathbf{Y}_{1:k+1}) && \left[ \text{since } I(X; \mathbf{Y}_{1:k} \mid Y_{k+1}) \geq 0 \right] \\
&= I(X; Y_{k+1} \mid \mathbf{Y}_{1:k}) + I(\mathbf{Y}_{1:k}; Y_{k+1}) && \left[ \text{by MI chain rule (5.5)} \right] \\
&= \begin{cases} I(X; \mathbf{Y}_{1:k+1}) & + H(Y_{k+1}) - H(\mathbf{Y}_{1:k+1}) \\ -I(X; \mathbf{Y}_{1:k}) & + H(\mathbf{Y}_{1:k}) \end{cases} && \left[ \begin{array}{l} \text{left: one more MI chain rule (5.5);} \\ \text{right: defn of mutual information} \end{array} \right].
\end{aligned}$$

Observe: adding this inequality to our inductive hypothesis  $(IH_k)$  yields  $(IH_{k+1})$ ! So, by induction, the lemma holds for all  $k$ .  $\square$

**Theorem 5.7.** *If  $\mu \models \Diamond \mathcal{A}$ , then  $SDef_{\mathcal{A}}(\mu) \leq 0$ .*

*Proof.* Suppose that  $\mu \models \mathcal{A}$ , meaning that there is a witness  $\nu(\mathcal{X}, \mathcal{U})$  that extends  $\mu$ , and has properties (a-c) of Definition 5.1. For each hyperarc  $a$ , since  $\nu \models (S_a, U_a) \rightarrow\!\!\!\rightarrow T_a$ , we have  $H_{\nu}(T_a \mid S_a, U_a) = 0$ , and so

$$H_{\mu}(T_a \mid S_a) = H_{\nu}(T_a \mid S_a, U_a) + I_{\nu}(T_a; U_a \mid S_a) = I_{\nu}(T_a; U_a \mid S_a).$$

Thus, we compute

$$\begin{aligned}
& \sum_{a \in \mathcal{A}} H_\mu(T_a \mid S_a) \\
&= \sum_{a \in \mathcal{A}} I_\nu(U_a; T_a \mid S_a) && \text{by the above} \\
&= \sum_{a \in \mathcal{A}} I_\nu(U_a; T_a, S_a) - I_\nu(U_a; S_a) && \text{by MI chain rule (5.5)} \\
&\leq \sum_{a \in \mathcal{A}} I_\nu(U_a; T_a, S_a) && \text{since } I_\nu(U_a; S_a) \geq 0 \\
&\leq \sum_{a \in \mathcal{A}} I_\nu(U_a; \mathcal{X}) && \text{since } \mathcal{X} \rightarrowtail (S_a, T_a) \\
&\leq I_\nu(\mathcal{X}; \mathcal{U}) + \sum_{a \in \mathcal{A}} H_\nu(U_a) - H_\nu(\mathcal{U}) && \text{by Lemma 5.12} \\
&= I_\nu(\mathcal{X}; \mathcal{U}) && \text{since } \mathcal{U} \text{ are independent} \\
&&& (\text{per condition (b) of Definition 5.1}) \\
&\leq H_\nu(\mathcal{X}) = H_\mu(\mathcal{X}). && \text{by condition (a) of Definition 5.1}
\end{aligned}$$

Thus,  $SDef_{\mathcal{A}}(\mu) \leq 0$ . □

**Proposition 5.8.**  $QIMInc_{\mathcal{A}}(\mu) \geq 0$ , with equality iff  $\mu \models \mathcal{A}$ .

*Proof.* The first term in the definition of  $QIMInc$  be written as

$$\left( -H_\nu(\mathcal{U}) + \sum_{a \in \mathcal{A}} H_\nu(U_a) \right) = \mathbb{E}_\nu \left[ \log \frac{\nu(\mathcal{U})}{\prod_a \nu(U_a)} \right]$$

and is therefore the relative entropy between  $\nu(\mathcal{U})$  and the independent product distribution  $\prod_{a \in \mathcal{A}} \nu(U_a)$ . Thus, it is non-negative. The remaining terms of  $QIMInc_{\mathcal{A}}(\mu)$ , are all conditional entropies, and hence non-negative as well. Thus  $QIMInc_{\mathcal{A}}(\mu) \geq 0$ .

Now, suppose  $\mu$  is s2-comaptible with  $\mathcal{A}$ , i.e., there exists some  $\nu(\mathcal{U}, \mathcal{X})$  such that (a)  $\nu(\mathcal{X}) = \mu(\mathcal{X})$ , (b)  $H_\nu(T_a | S_a, U_a) = 0$ , and (d)  $\{U_a\}_{a \in \mathcal{A}}$  are mutually independent. Then clearly  $\nu$  satisfies the condition under the infemum, every

$H_\nu(T_a | S_a, U_a)$  is zero. It is also immediate that the final term is zero as well, because it equals  $D(\nu(\mathcal{U}) \parallel \prod_a \nu(U_a))$ , and  $\nu(\mathcal{U}) = \prod_a \nu(U_a)$ , per the definition of mutual independence. Thus,  $\nu$  witnesses that  $\text{QIMInc}_{(\mathcal{A}, \lambda)} = 0$ .

Conversely, suppose  $\text{QIMInc}_{(\mathcal{A}, \lambda)} = 0$ . Because the feasible set is closed and bounded, as is the function, the infimum is achieved by some joint distribution  $\nu(\mathcal{X}, \mathcal{A})$  with marginal  $\mu(\mathcal{X})$ . In this distribution  $\nu$ , we know that every  $H_\nu(T_a | S_a, U_a) = 0$  and  $D(\nu(\mathcal{U}) \parallel \prod_a \nu(U_a)) = 0$ —because if any of these terms were positive, then the result would be positive as well. So  $\nu$  satisfies (a) and (b) by definition. And, because relative entropy is zero iff its arguments are identical we have  $\nu(\mathcal{U}) = \prod_a \nu(U_a)$ , so the  $U_a$ 's are mutually independent, and  $\nu$  satisfies (d) as well.  $\square$

### Theorem 5.9.

(a) If  $(\mathcal{X}, \mathcal{A})$  is a hypergraph,  $\mu(\mathcal{X})$  is a distribution, and  $\nu(\mathcal{X}, \mathcal{U})$  is an extension of  $\nu$  to additional variables  $\mathcal{U} = \{U_a\}_{a \in \mathcal{A}}$  indexed by  $\mathcal{A}$ , then:

$$SDef_{\mathcal{A}}(\mu) \leq \text{QIMInc}_{\mathcal{A}}(\mu) \leq SDef_{\mathcal{A}^\dagger}(\nu).$$

(b) For all  $\mu$  and  $\mathcal{A}$ , there is a choice of  $\nu$  that achieves the upper bound. That is,

$$\text{QIMInc}_{\mathcal{A}}(\mu) = \min \left\{ SDef_{\mathcal{A}^\dagger}(\nu) : \begin{array}{l} \nu \in \Delta^{\mathcal{V}}(\mathcal{X}, \mathcal{U}) \\ \nu(\mathcal{X}) = \mu(\mathcal{X}) \end{array} \right\}.$$

*Proof.* Part (a). The left-hand side of the theorem ( $SDef_{\mathcal{A}}(\nu) \leq \text{QIMInc}_{\mathcal{A}}(\mu)$ ) is a strengthening of the argument used to prove [Theorem 5.7](#). Specifically, letting  $\nu^*$  be a minimizer of the optimization problem defining  $\text{QIMInc}$ ,

$$\begin{aligned} & \text{QIMInc}_{\mathcal{A}}(\mu) - SDef_{\mathcal{A}}(\mu) \\ &= \left( \sum_{a \in \mathcal{A}} H_{\nu^*}(T_a | S_a, U_a) - H_{\nu^*}(\mathcal{U}) + \sum_{a \in \mathcal{A}} H_{\nu^*}(U_a) \right) - \left( \sum_{a \in \mathcal{A}} H_\mu(T_a | S_a) - H_\mu(\mathcal{X}) \right) \end{aligned}$$

$$\begin{aligned}
&= \sum_{a \in \mathcal{A}} \left( H_{\nu^*}(T_a | S_a, U_a) - H_{\nu^*}(T_a | S_a) \right) + H_\mu(\mathcal{X}) - H_{\nu^*}(\mathcal{U}) + \sum_{a \in \mathcal{A}} H_{\nu^*}(U_a) \\
&= - \sum_{a \in \mathcal{A}} I_{\nu^*}(T_a; U_a | S_a) + H_\mu(\mathcal{X}) - H_{\nu^*}(\mathcal{U}) + \sum_{a \in \mathcal{A}} H_{\nu^*}(U_a).
\end{aligned}$$

The argument given in the first five lines of the proof of [Theorem 5.7](#), gives us a particularly convenient bound for the first group of terms on the left:

$$\sum_{a \in \mathcal{A}} I_{\nu^*}(U_a; T_a | S_a) \leq I_{\nu^*}(\mathcal{X}; \mathcal{U}) + \sum_{a \in \mathcal{A}} H_{\nu^*}(U_a) - H_{\nu^*}(\mathcal{U}).$$

Substituting this into our previous expression, we have:

$$\begin{aligned}
&\text{QIMInc}_{\mathcal{A}}(\mu) - SDef_{\mathcal{A}}(\mu) \\
&\geq - \left( I_{\nu^*}(\mathcal{X}; \mathcal{U}) + \sum_{a \in \mathcal{A}} H_{\nu^*}(U_a) - H_{\nu^*}(\mathcal{U}) \right) + H_\mu(\mathcal{X}) - H_{\nu^*}(\mathcal{U}) + \sum_{a \in \mathcal{A}} H_{\nu^*}(U_a) \\
&= H_\mu(\mathcal{X}) - I_{\nu^*}(\mathcal{X}; \mathcal{U}) \\
&\geq 0.
\end{aligned}$$

The final inequality holds because of our assumption that the marginal  $\nu^*(\mathcal{X})$  equals  $\mu(\mathcal{X})$ . Thus,  $\text{QIMInc}_{\mathcal{A}}(\mu) \geq SDef_{\mathcal{A}}(\mu)$ , as promised.

We now turn to the right-hand inequality, and part (b) of the theorem. Recall that  $\nu^*$  is defined to be a minimizer of the optimization problem defining  $\text{QIMInc}$ . For the right inequality ( $\text{QIMInc}_{\mathcal{A}}(\mu) \leq SDef_{\mathcal{A}^\dagger}(\nu)$ ) of part (a), observe that

$$\begin{aligned}
SDef_{\mathcal{A}^\dagger}(\nu) &= -H_\nu(\mathcal{X}, \mathcal{U}) + \sum_{a \in \mathcal{A}} H_\nu(U_a) + \sum_{a \in \mathcal{A}} H_\nu(T_a | S_a, U_a) + H_\nu(\mathcal{X} | \mathcal{U}) \\
&= \left( -H_\nu(\mathcal{U}) + \sum_{a \in \mathcal{A}} H_\nu(U_a) \right) + \sum_{a \in \mathcal{A}} H_\nu(T_a | S_a, U_a) \\
&\geq \left( -H_{\nu^*}(\mathcal{U}) + \sum_{a \in \mathcal{A}} H_{\nu^*}(U_a) \right) + \sum_{a \in \mathcal{A}} H_{\nu^*}(T_a | S_a, U_a) \\
&= \text{QIMInc}(\mu).
\end{aligned}$$

This proves the right-hand side of the inequality of part (a). Moreover, because

the one inequality holds with equality when  $\nu = \nu^*$  is a minimizer of this quantity (subject to having marginal  $\mu(\mathcal{X})$ ) we have shown part (b) as well.  $\square$

## 5.B Constructions and Counterexamples

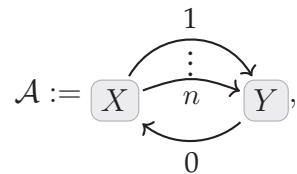
### 5.B.1 Parallel Arcs without Functional Dependency

We now give a counterexample to a simpler previously conjectured strengthening of [Theorem 5.2](#), in which part (a) is an if-and-only-if. In the unconditional case, it is true that, two arcs  $\{\overset{1}{\rightarrow} X, \overset{2}{\rightarrow} X\}$  precisely encode that  $X$  is a constant, as illustrated by [Example 5.2](#). The following, slightly more general result, is an immediate corollary of [Theorem 5.2\(c\)](#).

**Proposition 5.13.**  $\mu \models \mathcal{A} \sqcup \{\overset{1}{\rightarrow} X, \overset{2}{\rightarrow} X\}$  if and only if  $\mu \models \mathcal{A}$  and  $\mu \models \emptyset \twoheadrightarrow X$ .

One can be forgiven for imagining that the conditional case would be analogous—that QIM-compatibility with a hypergraph that has two parallel arcs from  $X$  to  $Y$  would imply that  $Y$  is a function of  $X$ . But this is not the case. Furthermore, our counterexample also shows that neither of the two properties we consider in the main text (requiring that  $\mathcal{A}$  is partitional, or that the QIM-compatibility with  $\mu$  is even) are enough to ensure this. That is, there are partitional graphs  $\mathcal{A}$  such that  $\mu \models^e \mathcal{A}$  but  $\mu \not\models \mathcal{A} \sqcup \{X \overset{1}{\rightarrow} Y, X \overset{2}{\rightarrow} Y\}$ .

**Example 5.7.** We will construct a witness of SIM-compatibility for the hypergraph



in which  $Y$  is *not* a function of  $X$ , which for  $n = 3$  will disprove the analogue of [Theorem 5.2](#) for the partitional context  $\mathcal{A}'$  equal to the 2-cycle.

Let  $\mathcal{U} = (U_0, U_1, \dots, U_n)$  be a vector of  $n$  mutually independent random coins, and  $A$  is one more independent random coin. For notational convenience, define the random vector  $\mathbf{U} := (U_0, \dots, U_n)$  consisting of all variables  $U_i$  except for  $U_0$ . Then, define variables  $X$  and  $Y$  according to:

$$\begin{aligned} X &:= (A \oplus U_1, \dots, A \oplus U_n, U_0 \oplus U_1, U_0 \oplus U_2, \dots, U_0 \oplus U_n) \\ &= (A \oplus \mathbf{U}, U_0 \oplus \mathbf{U}) \\ Y &:= (A, U_0 \oplus \mathbf{U}) = (A, U_0 \oplus U_1, U_0 \oplus U_2, \dots, U_0 \oplus U_n), \end{aligned}$$

where and the operation  $Z \oplus \mathbf{V}$  is element-wise xor (or addition in  $\mathbb{F}_2^n$ ), after implicitly converting the scalar  $Z$  to a vector by taking  $n$  copies of it. Call the resulting distribution  $\nu(X, Y, \mathcal{U})$ .

If we now show that  $\nu$  witnesses that its marginal on  $X, Y$  is QIM-compatible with  $\mathcal{A}$ , which is straightforward.

(b)  $\mathcal{U}$  are mutually independent by assumption;

(c.0)  $Y = (A, \mathbf{B})$  and  $U_0$  determine  $X$  according to:

$$\begin{aligned} g(A, \mathbf{B}, U_0) &= (A \oplus U_0 \oplus \mathbf{B}, \mathbf{B}) \\ &= (A \oplus U_0 \oplus U_0 \oplus \mathbf{U}, U_0 \oplus \mathbf{U}) \quad \text{since } \mathbf{B} = U_0 \oplus \mathbf{U} \\ &= (A \oplus \mathbf{U}, U_0 \oplus \mathbf{U}) = X \end{aligned}$$

(c.1–n) for  $i \in \{1, \dots, n\}$ ,  $U_i$  and  $X = (\mathbf{V}, \mathbf{B})$  together determine  $Y$  according to

$$f_i(\mathbf{V}, \mathbf{B}, U_i) := (V_i \oplus U_i, \mathbf{B}) = (A \oplus U_i \oplus U_i, U_0 \oplus \mathbf{U}) = Y.$$

In addition, this distribution  $\nu(\mathcal{U}, X, Y)$  satisfies condition

- (d)  $\nu(X, Y \mid \mathcal{U}) = \frac{1}{2}\mathbb{1}[g(Y, U_0) = X] \prod_{i=1}^n \mathbb{1}[f_i(X, U_i) = Y]$ , since, for all joint settings of  $\mathcal{U}$ , there are two possible values of  $(X, Y)$ , corresponding to the two values of  $A$ , and both happen with probability  $\frac{1}{2}$ .

Thus, we have constructed a distribution that witnessing the fact that  $\mu(X, Y) \models^e \mathcal{A}$ . But observe that  $X$  alone does not determine  $Y$  in this distribution, because  $X$  alone is not enough to determine  $A$  (without also knowing some  $U_i$ ).

For completeness, note that the bound of [Theorem 5.7](#) tells that we must satisfy

$$\begin{aligned} 0 &\geq SDef_{\mathcal{A}}(\mu) = -H_{\mu}(X, Y) + nH_{\mu}(Y \mid X) + H_{\mu}(X \mid Y) \\ &= -I_{\mu}(X; Y) + (n - 1)H_{\mu}(Y \mid X) \end{aligned}$$

Indeed, this distribution has information profile

$$H(X \mid Y) = 1 \text{ bit}, \quad I(X; Y) = n \text{ bits}, \quad H(Y \mid X) = 1 \text{ bit},$$

and so  $SDef_{\mathcal{A}}(\mu) = -1$  bit. Intuitively, this one missing bit corresponds to the value of  $A$  that is not determined by the structure of  $\mathcal{A}$ .  $\triangle$

### 5.B.2 Counter-Examples to the Converse of [Theorem 5.7](#)

In light of [Example 5.6](#) and its connections to  $SDef$  through [Theorem 5.7](#), one might hope this criterion is not just a bound, but a precise characterization of the distributions that are QIM-compatible with the 3-cycle. Unfortunately, it does not, and the converse of [Theorem 5.7](#) is false.

**Example 5.8.** Suppose  $\mu(X, Y, Z) = \text{Unif}(X, Z)\delta\text{id}(Y|X)$  and  $\mathcal{A} = \{\rightarrow X, \rightarrow Y\}$ , where all variables are binary. Then  $S\text{Def}_{\mathcal{A}}(\mu) = 0$ , but  $X$  and  $Y$  are not independent.  $\triangle$

Here is another counter-example, of a very different kind.

**Example 5.9.** Suppose  $A, B, C$  are binary variables. It can be shown by enumeration (see appendix) that no distribution supported on seven of the eight possible joint settings of  $\mathcal{V}(A, B, C)$  can be QIM-compatible with the 3-cycle  $\mathcal{A}_{3o}$ . Yet it is easy to find examples of such distributions  $\mu$  that have positive interaction information  $I(A; B; C)$ , and thus  $S\text{Def}_{\mu}(\mathcal{A}_{3o}) \leq 0$  for such distributions.  $\triangle$

## 5.C From Causal Models to Witnesses

We now return to the “easy” direction of the correspondence between QIM-compatibility witnesses and causal models, mentioned at the beginning of [Section 5.3.2](#). Given a (generalized) randomized PSEM  $\mathcal{M}$ , we now show that distributions  $\nu \in \{\mathcal{M}\}$  are witnesses to the QIM-compatibility of the marginals of  $\nu$  with the hypergraph  $\mathcal{A}_{\mathcal{M}}$ . More formally:

**Proposition 5.14.** *If  $\mathcal{M} = (\mathcal{U}, \mathcal{V}, \mathcal{F}, P)$  is a randomized PSEM, then every  $\nu \in \{\mathcal{M}\}$  witnesses the QIM-compatibility of its marginal on its exogenous variables, with the dependency structure of  $\mathcal{M}$ . That is, for all  $\nu \in \{\mathcal{M}\}$  and  $\mathcal{Y} \subseteq \mathcal{U} \cup \mathcal{V}$ ,  $\nu(\mathcal{Y}) \models \Diamond \mathcal{A}_{\mathcal{M}}$ .*

The proof is straightforward: by definition, if  $\nu \in \{\mathcal{M}\}$ , then it must satisfy the equations, and so automatically fulfills condition (c). Condition (a) is also satisfied trivially, by assumption: the distribution we’re considering is defined to

be a marginal of  $\nu$ . Finally, (b) is also satisfied by construction: we assumed that  $\mathcal{U}_{\mathcal{A}} = \{U_a\}_{a \in \mathcal{A}}$  are independent.

## **Part II**

# **A Universal Objective**

## CHAPTER 6

### LOSS AS THE INCONSISTENCY OF A PDG: CHOOSE YOUR MODEL, NOT YOUR LOSS

In a world blessed with a great diversity of loss functions, we argue that that choice between them is not a matter of taste or pragmatics, but of model. Recall that there is a natural way to measure the degree of a PDG’s inconsistency. In this chapter, we prove that many standard loss functions arise as the inconsistency of a natural PDG describing the appropriate scenario, and use the same approach to justify a well-known connection between regularizers and priors. We also show that the PDG inconsistency captures a large class of statistical divergences, and detail benefits of thinking of them in this way, including an intuitive visual language for deriving inequalities between them. In variational inference, we find that the ELBO, a somewhat opaque objective for latent variable models, and variants of it arise for free out of uncontroversial modeling assumptions—as do simple graphical proofs of their corresponding bounds. Finally, we observe that inconsistency becomes the log partition function (free energy) in the setting where PDGs are factor graphs.

#### 6.1 Introduction

Many tasks in artificial intelligence have been fruitfully cast as optimization problems, but often the choice of objective is not unique. For instance, a key component of a machine learning system is a loss function which the system must minimize, and a wide variety of losses are used in practice. Each implicitly represents different values and results in different behavior, so the choice between them can be quite important (Wang et al. 2020; Jadon 2020). Yet, because it’s

unclear how to choose a “good” loss function, the choice is usually made by empirics, tradition, and an instinctive calculus acquired through the practice—not by explicitly laying out beliefs. Furthermore, there is something to be gained by fiddling with these loss functions: one can add regularization terms, to (dis)incentivize (un)desirable behavior. But the process of tinkering with the objective until it works is often unsatisfying. It can be a tedious game without clear rules or meaning, while results so obtained are arguably overfitted and difficult to motivate.

By contrast, a choice of *model* admits more principled discussion, in part because models are testable; it makes sense to ask if a model is accurate. This observation motivates our proposal: instead of specifying a loss function directly, one articulates a situation that gives rise to it, in the (more interpretable) language of probabilistic beliefs and certainties. Concretely, we use the machinery of Probabilistic Dependency Graphs (PDGs), a particularly expressive class of graphical models that can incorporate arbitrary (even inconsistent) probabilistic information in a natural way, and comes equipped with a well-motivated measure of inconsistency ([Richardson and Halpern 2021](#)).

A primary goal of this paper is to show that PDGs and their associated inconsistency measure can provide a “universal” model-based loss function. Towards this end, we show that many standard objective functions—cross entropy, square error, many statistical distances, the ELBO, regularizers, and the log partition function—arise naturally by measuring the inconsistency of the appropriate underlying PDG. This is somewhat surprising, since PDGs were not designed with the goal of capturing loss functions at all. Specifying a loss function indirectly like this is in some ways more restrictive, but it is also more intuitive (it no technical familiarity with losses, for instance), and admits more grounded

defense and criticism.

For a particularly powerful demonstration, consider the variational autoencoder (VAE), an enormously successful class of generative model that has enabled breakthroughs in image generation, semantic interpolation, and unsupervised feature learning ([Kingma and Welling 2014](#)). Structurally, a VAE for a space  $X$  consists of a (smaller) latent space  $Z$ , a prior distribution  $p(Z)$ , a decoder  $d(X|Z)$ , and an encoder  $e(Z|X)$ . A VAE is not considered a “graphical model” for two reasons. The first is that the encoder  $e(Z|X)$  has the same target variable as  $p(Z)$ , so something like a Bayesian Network cannot simultaneously incorporate them both (besides, they could be inconsistent with one another). The second reason: it is not a VAE’s structure, but rather its *loss function* that makes it tick. A VAE is typically trained by maximizing the “ELBO”, a somewhat difficult-to-motivate function of a sample  $x$ , originating in variational calculus. We show that  $-\text{ELBO}(x)$  is also precisely the inconsistency of a PDG containing  $x$  and the probabilistic information of the autoencoder ( $p$ ,  $d$ , and  $e$ ). We can form such a PDG precisely because PDGs allow for inconsistency. Thus, PDG semantics simultaneously legitimize the strange structure of the VAE, and also justify its loss function, which can be thought of as a property of the model itself (its inconsistency), rather than some mysterious construction borrowed from physics.

Representing objectives as model inconsistencies, in addition to providing a principled way of selecting an objective, also has beneficial pedagogical side effects, because of the *structural* relationships between the underlying models. For instance, these relationships will allow us to derive simple and intuitive visual proofs of technical results, such as the variational inequalities that traditionally motivate the ELBO, and the monotonicity of Rényi divergence.

In the remainder of this chapter, we show in more detail how this concept of inconsistency, beyond simply providing a permissive and intuitive modeling framework, reduces exactly to many standard objectives used in machine learning and to measures of statistical distance. We demonstrate that this framework clarifies the relationships between them, by providing clear derivations of otherwise opaque inequalities.

**Preliminaries.** This chapter focuses primarily on *quantitative* aspects of PDGs—after all, the idea behind modern machine learning is to always defer to data and empirics. So, unless we say otherwise, we will take the inconsistency of a PDG refer to the purely observational inconsistency ( $\gamma = 0$ ), dropping the subscript and writing  $\langle\!\langle \mathbf{m} \rangle\!\rangle := \inf_{\mu} OInc_{\mathbf{m}}(\mu)$ .

Intuitively, believing more things can't make you any less inconsistent. **Lemma 6.1** captures this formally: adding cpds or increasing confidences cannot decrease a PDG's inconsistency.

**Lemma 6.1** (Monotonicity of Inconsistency). *Suppose PDGs  $\mathbf{m}$  and  $\mathbf{m}'$  differ only in their edges (resp.  $\mathcal{A}$  and  $\mathcal{A}'$ ) and confidences (resp.  $\beta$  and  $\beta'$ ). If  $\mathcal{A} \subseteq \mathcal{A}'$  and  $\beta_a \leq \beta'_a$  for all  $a \in \mathcal{A}$ , then  $\langle\!\langle \mathbf{m} \rangle\!\rangle_{\gamma} \leq \langle\!\langle \mathbf{m}' \rangle\!\rangle_{\gamma}$  for all  $\gamma$ .<sup>1</sup>*

[ link to proof ]

As we will see, this tool is sufficient to derive many interesting relationships between loss functions.

---

<sup>1</sup>All proofs can be found in [Section 6.C](#).

## 6.2 Standard Metrics as Inconsistencies

### 6.2.1 Three Dimensions of Log-Likelihood

Suppose that you believe that  $X$  is distributed according to  $p(X)$ , and also that it (certainly) equals some value  $x$ . These beliefs are consistent if  $p(X=x) = 1$  but become less so as  $p(X=x)$  decreases. In fact, this inconsistency is equal to the information content  $I_p[X=x] := -\log p(X=x)$ , or *surprisal* (Tribus 1961), of the event  $X=x$ , according to  $p$ .<sup>2</sup> In machine learning,  $I_p$  is usually called “negative log likelihood”, and is the de-facto standard training objective for training generative models (Grover and Ermon 2018; Myung 2003).

**Proposition 6.2.** Consider a distribution  $p(X)$ . The inconsistency of the PDG comprising  $p$  and  $X=x$  equals the surprisal  $I_p[X=x]$ . That is,

$$I_p[X=x] = \langle\!\langle \xrightarrow{p} X \leftarrow^x \rangle\!\rangle.$$

(Recall that  $\langle\!\langle m \rangle\!\rangle$  is the inconsistency of the PDG  $m$ .)

In some ways, this result is entirely unsurprising, given that (3.1) is a flexible formula built out of information-theoretic primitives. Even so, note that the inconsistency of believing both a distribution and an event happens to be the standard measure of discrepancy between the two—and is even named after “surprise”, a particular expression of epistemic conflict.

Still, we have a ways to go before this amounts to any more than a curiosity. One concern is that this picture is incomplete; we train probabilistic models with

<sup>2</sup>This construction requires the event  $X=x$  to be measurable, and is only useful if it has nonzero probability. One can get similar, but subtler, results for probability densities; see Section 6.A.

more than one sample. What if we replace  $x$  with an empirical distribution over many samples?

**Proposition 6.3.** *If  $p(X)$  is a probabilistic model of  $X$ , and  $\mathcal{D} = \{x_i\}_{i=1}^m$  is a dataset with empirical distribution  $\Pr_{\mathcal{D}}$ , then*

$$\text{CrossEntropy}(\Pr_{\mathcal{D}}, p) := \frac{1}{m} \sum_{i=1}^m I_p[X=x_i] = \left\langle \overrightarrow{p} \boxed{X} \xleftarrow{(\infty)} \Pr_{\mathcal{D}} \right\rangle + H(\Pr_{\mathcal{D}}).$$

*Remark.* As usual,  $H(\Pr_{\mathcal{D}})$  is a constant depending only on the data, so is irrelevant for the purposes of optimizing  $p$ .

Essentially the only choices we've made in specifying the PDG of [Proposition 6.3](#) are the confidences. But  $\text{CrossEntropy}(\Pr_{\mathcal{D}}, p)$  is the expected code length per sample from  $\Pr_{\mathcal{D}}$ , when using codes optimized for the (incorrect) distribution  $p$ . So implicitly, a modeler using cross-entropy has already articulated a belief the data distribution  $\Pr_{\mathcal{D}}$  is the "true one". To get the same effect from a PDG, the modeler must make this belief explicit by placing infinite confidence in the data distribution  $\Pr_{\mathcal{D}}$ .

Now consider an orthogonal generalization of [Proposition 6.2](#), in which the sample  $x$  is only a partial observation of  $(x, z)$  from a joint model  $p(X, Z)$ .

**Proposition 6.4.** *If  $p(X, Z)$  is a joint distribution, then the information content of the partial observation  $X = x$  is given by*

$$I_p[X=x] = \left\langle \boxed{Z} \xleftarrow{p} \boxed{X} \xleftarrow{x} \right\rangle. \quad (6.1)$$

Intuitively, the inconsistency of the PDG on the right hand side of Equation (6.1) is localized to the variable  $X$ , where the observation  $x$  conflicts with  $p(X)$ . In other words, the only relevant aspect of the distribution  $p(X, Z)$  is its

marginal on  $X$ . [Propositions 6.3](#) and [6.4](#) extend the surprisal result ([Proposition 6.2](#)) in two orthogonal directions: to settings with multiple observations, and with partial observation. The multi-sample partial-observation generalization also holds; see [Section 6.B.3](#). We now introduce a third axis.

So far we have considered models of an unconditional distribution  $p(X)$ . Because they are unconditional, such models must describe how to generate a complete sample  $X$  without input, and so are called *generative*; the process of training them is called *unsupervised* learning ([Hastie et al. 2009](#)). In the (more common) *supervised* setting, we train *discriminative* models to predict  $Y$  from  $X$ , via labeled samples  $\{(x_i, y_i)\}_i$ . There, cross entropy loss is perhaps even more dominant—and it is essentially the inconsistency of a PDG consisting of the predictor  $h(Y|X)$  together with high-confidence data.

**Proposition 6.5** (Cross Entropy, Supervised). *The inconsistency of the PDG comprising a probabilistic predictor  $h(Y|X)$ , and a high-confidence empirical distribution  $\Pr_{\mathcal{D}}$  of a dataset  $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m$  equals the cross-entropy loss (minus the empirical uncertainty in  $Y$  given  $X$ , a constant depending only on  $\mathcal{D}$ ). That is,*

$$\left\langle\!\!\left\langle \begin{array}{c} \Pr_{\mathcal{D}} \downarrow^{(\infty)} \\ X \xrightarrow[h]{} Y \end{array} \right\rangle\!\!\right\rangle = \frac{1}{m} \sum_{i=1}^m \log \frac{1}{h(y_i | x_i)} - H_{\Pr_{\mathcal{D}}}(Y|X).$$

[ link to proof ]

[Proposition 6.5](#) describes the multi-sample conditional instantiation of log-likelihood, which in practice is the most common setting for machine learning altogether. Analogous results hold for all combinations of the three axes we have described:

$$\left\{ \begin{array}{ll} \text{unconditional} & \text{conditional} \\ (\text{generative}), & (\text{discriminative}) \end{array} \right\} \times \left\{ \begin{array}{ll} \text{single-} & \text{multi-sample} \\ \text{sample'} & (\text{full dataset}) \end{array} \right\} \times \left\{ \begin{array}{ll} \text{full} & \text{partial info} \\ \text{info}, & (\text{latent variable}) \end{array} \right\};$$

See figure [Figure 6.1](#) for an illustration.

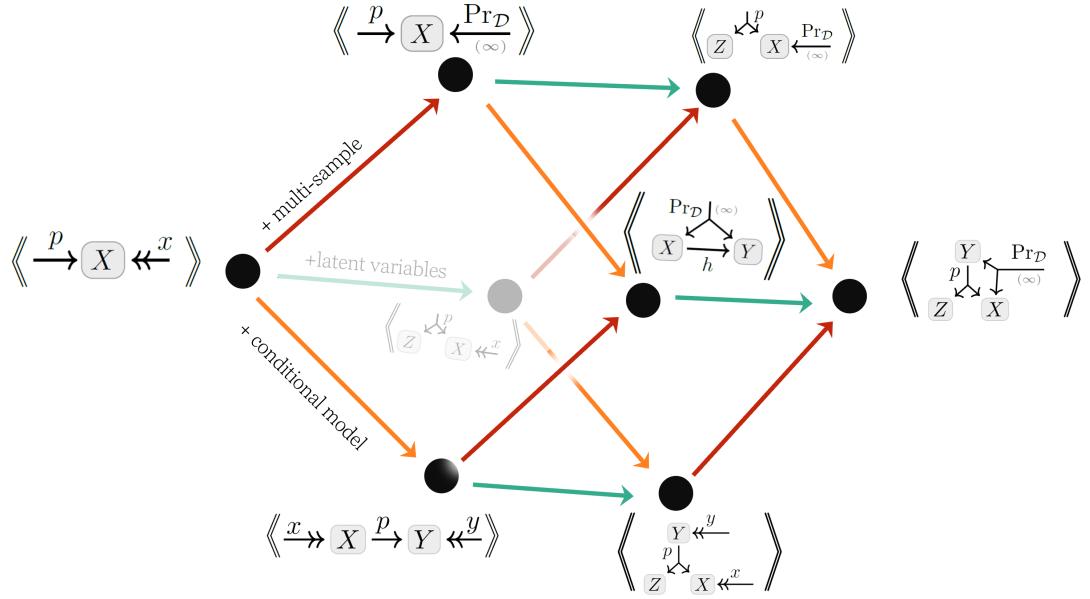


Figure 6.1: Variants of log-probability based losses, across three orthogonal dimensions: conditional vs unconditional, multi-sample vs single-sample, and latent-variable vs full-information

### 6.2.2 Accuracy and Square Loss

Simple evaluation metrics, such as the accuracy of a classifier, and the mean squared error of a regressor, also arise naturally as inconsistencies.

**Proposition 6.6** (Log Accuracy as Inconsistency). *Consider functions  $f, h : X \rightarrow Y$  from inputs to labels, where  $h$  is a predictor and  $f$  generates the true labels. The inconsistency of believing  $f$  and  $h$  (with any confidences), and a distribution  $D(X)$  with confidence  $\beta$ , is  $\beta$  times the log accuracy of  $h$ . That is,*

$$\begin{aligned} \langle\langle D_{(\beta)} \xrightarrow{(r)} X \xrightarrow{(s)} Y \rangle\rangle &= -\beta \log \Pr_{x \sim D} (f(x) = h(x)) \\ &= \beta \mathbb{I}_D[f = h]. \end{aligned} \tag{6.2}$$

One often speaks of the accuracy of a hypothesis  $h$ , leaving the true labels  $f$  and empirical distribution  $D$  implicit. Yet Proposition 6.6 suggests that there

link to proof

is a sense in which  $D(X)$  plays the primary role: the inconsistency in (6.2) is scaled by the confidence in  $D$ , and does not depend on the confidences in  $h$  or  $f$ . Why should this be this the case? Expressing  $(x, y)$  such that  $y \neq f(x)$  with codes optimized for  $f$  is not just inefficient, but impossible. The same is true for  $h$ , so we can only consider  $\mu$  such that  $\mu(f=h)=1$ . In other words, the only way to form a joint distribution *at all* compatible with both the predictor  $h$  and the labels  $f$ , is to throw out samples that the predictor gets wrong—and the cost of throwing out samples scales with your confidence in  $D$ , not in  $h$ . This illustrates why accuracy gives no gradient information for training  $h$ . It is worth noting that this is precisely the opposite of what happened in [Proposition 6.5](#): there we were unwilling to budge on the input distribution, and the inconsistency scaled with the confidence in  $h$ .

Observe how even properties of these simple metrics—relationships with one another and features of gradients—can be clarified by an underlying model.

When  $Y \cong \mathbb{R}^n$ , an estimator  $h(Y|X)$  is referred to as a regressor instead of a classifier. In this setting, most answers are incorrect, but some more so than others. A common way of measuring incorrectness is with mean squared error (MSE):  $\mathbb{E}|f(X) - Y|^2$ . MSE is also the inconsistency of believing that the labels and predictor have Gaussian noise—often a reasonable assumption because of the central limit theorem.

**Proposition 6.7** (MSE as Inconsistency).

[ link to proof ]

$$\left\langle\!\!\left\langle \begin{array}{c} D \\ \xrightarrow{(\infty)} \\ X \end{array} \xrightarrow{\begin{array}{c} f \\ h \end{array}} \begin{array}{c} \mu_f \\ \mu_h \end{array} \xrightarrow{\mathcal{N}_1} Y \right\rangle\!\!\right\rangle = \frac{1}{2} \mathbb{E}_D |f(X) - h(X)|^2 =: \text{MSE}_D(f, h),$$

where  $\mathcal{N}_1(Y|\mu)$  is a unit Gaussian on  $Y$  with mean  $\mu$ .

In the appendix, we treat general univariate Gaussian predictors, with arbitrary variances and confidences.

### 6.3 Regularizers and Priors as Inconsistencies

Regularizers are extra terms added to loss functions, which provide a source of inductive bias towards simple model parameters. There is a well-known correspondence between using a regularizer and doing maximum *a posteriori* inference with a prior (see [Section 6.A](#)), in which L2 regularization corresponds to a Gaussian prior ([Rennie 2003](#)), while L1 regularization corresponds to a Laplacian prior ([Williams 1995](#)). Note that the ability to make principled modeling choices about regularizers is a primary benefit of this correspondence. Our approach provides a new justification of it.

**Proposition 6.8.** Suppose you have a parameterized model  $p(Y|\Theta)$ , a prior  $q(\Theta)$ , and a trusted distribution  $D(Y)$ . The inconsistency of also believing  $\Theta = \theta$  is the cross entropy loss, plus the regularizer:  $\log \frac{1}{q(\theta)}$  times your confidence in  $q$ . That is,

$$\left\langle \begin{array}{c} q \\ (\beta) \rightarrow \\ \theta \end{array} \rightarrow \Theta \xrightarrow{p} Y \right\rangle_{D \uparrow (\infty)} = \mathbb{E}_{y \sim D} \log \frac{1}{p(y|\theta)} + \beta \log \frac{1}{q(\theta)} - H(D) \quad (6.3)$$

If our prior is  $q(\theta) = \frac{1}{k} \exp(-\frac{1}{2}\theta^2)$ , a (discretized) unit gaussian, then the right hand side of [\(6.3\)](#) becomes

$$\underbrace{\mathbb{E}_D \log \frac{1}{p(Y|\theta)}}_{\text{Cross entropy loss}} + \underbrace{\frac{\beta}{2} \theta_0^2}_{\substack{\text{L2 regularizer} \\ (\text{complexity cost of } \theta)}} + \underbrace{\beta \log k - H(D)}_{\text{constant in } p \text{ and } \theta},$$

which is the L2 regularized version of [Proposition 6.3](#). Moreover, the regularization strength corresponds exactly to the confidence  $\beta$ . What about other

link to proof

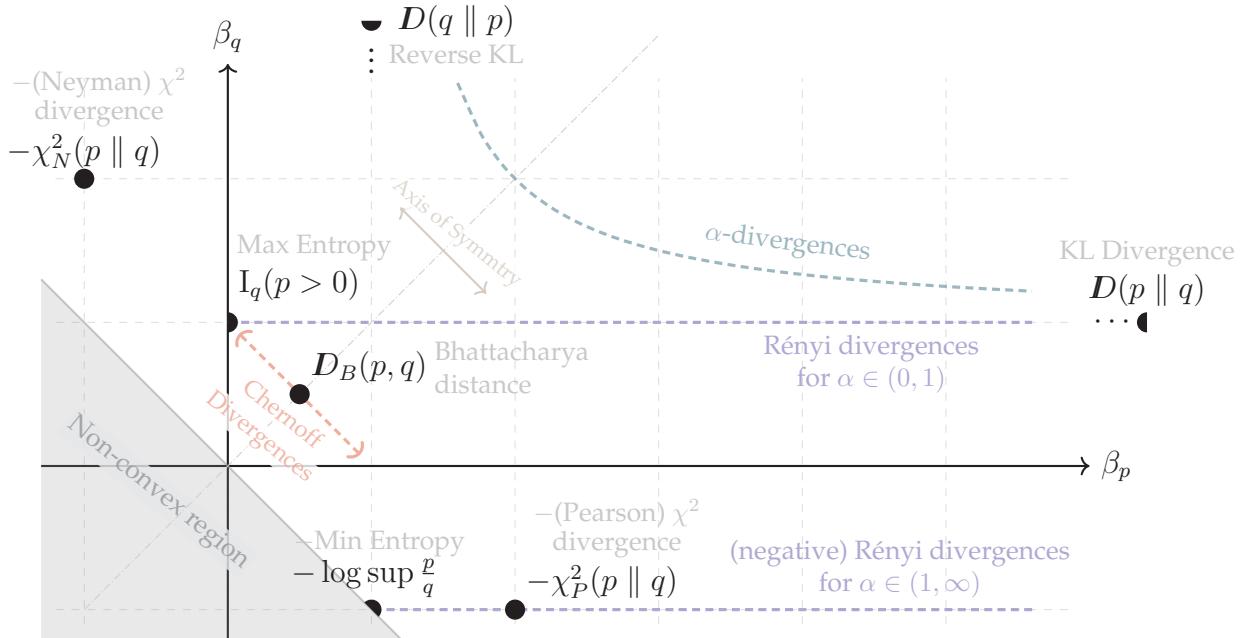


Figure 6.1: A map of the inconsistency of the PDG comprising  $p(X)$  and  $q(X)$ , as we vary their respective confidences  $\beta_p$  and  $\beta_q$ . Solid circles indicate well-known named measures, semicircles indicate limiting values, and the heavily dashed lines are well-established classes.

priors? It is not difficult to see that if we use a (discretized) unit Laplacian prior,  $q(\theta) \propto \exp(-|\theta|)$ , the second term instead becomes  $\beta|\theta_0|$ , which is L1 regularization. More generally, to consider a complexity measure  $U(\theta)$ , we need only include the Gibbs distribution  $\Pr_U(\theta) \propto \exp(-U(\theta))$  into our PDG. We remark that nothing here is specific to cross entropy; any of the objectives we describe can be regularized in this way.

#### 6.4 Statistical Distances as Inconsistencies

Suppose you are concerned with a single variable  $X$ . One friend has told you that it is distributed according to  $p(X)$ ; another has told you that it follows  $q(X)$ . You adopt both beliefs. Your mental state will be inconsistent if (and only if)  $p \neq q$ , with more inconsistency the more  $p$  and  $q$  differ. Thus the inconsistency

of a PDG comprising  $p$  and  $q$  is a measure of divergence. Recall that a PDG also allows us to specify the confidences  $\beta_p$  and  $\beta_q$  of each cpd, so we can form a PDG divergence  $D_{(r,s)}^{\text{PDG}}(p\|q)$  for every setting  $(r, s)$  of  $(\beta_p, \beta_q)$ . It turns out that a large class of statistical divergences arise in this way. We start with a familiar one.

**Proposition 6.9** (KL Divergence as Inconsistency). *The inconsistency of believing  $p$  with complete certainty, and also  $q$  with some finite certainty  $\beta$ , is  $\beta$  times the KL Divergence (or relative entropy) of  $q$  with respect to  $p$ . That is,*

$$\left\langle\left\langle \frac{p}{(\infty)} \rightarrow [X] \leftarrow \frac{q}{(\beta)} \right\rangle\right\rangle = \beta D(p \| q).$$

This result gives us an(other) intuitive interpretation of the asymmetry of relative entropy (i.e., KL divergence), and a prescription for when it makes sense to use it.  $D(p \| q)$  is the inconsistency of a mental state containing both  $p$  and  $q$ , when absolutely certain of  $p$  (and not willing to budge on it). This concords with the standard intuition that  $D(p \| q)$  reflects the amount of information required to change  $q$  into  $p$ , which is why it is usually called the relative entropy “from  $q$  to  $p$ ”. We now consider the general case of a PDG comprising  $p(X)$  and  $q(X)$  with arbitrary confidences.

**Lemma 6.10.** *The inconsistency  $D_{(r,s)}^{\text{PDG}}(p\|q)$  of a PDG comprising  $p(X)$  with confidence  $r$  and  $q(X)$  with confidence  $s$  is given in closed form by*

$$D_{(r,s)}^{\text{PDG}}(p\|q) := \left\langle\left\langle \frac{p}{(r)} \rightarrow [X] \leftarrow \frac{q}{(s)} \right\rangle\right\rangle = -(r+s) \log \sum_x \left( p(x)^r q(x)^s \right)^{\frac{1}{r+s}}.$$

Of the many generalizations of KL divergence, Rényi divergences, first characterized by Alfréd Rényi 1961 are perhaps the most significant, as few others have found either application or an interpretation in terms of coding theory

(Van Erven and Harremos 2014). The Rényi divergence of order  $\alpha$  between two distributions  $p(X)$  and  $q(X)$  is given by

$$D_\alpha(p \parallel q) := \frac{1}{1-\alpha} \log \sum_{x \in \mathcal{V}(X)} p(x)^\alpha q(x)^{1-\alpha}. \quad (6.4)$$

Rényi introduced this measure in the same paper as the more general class of  $f$ -divergences, but directs his attention towards those of the form (6.4), because they satisfy a natural weakening of standard postulates for Shannon entropy due to Fadeev (1957). Concretely, every symmetric, continuous measure that additively separates over independent events, and with a certain “mean-value property”, up to scaling, is of the form (6.4) for some  $\alpha$  (Rényi 1961). It follows from Lemma 6.10 that every Rényi divergence is a PDG divergence, and every (non-limiting) PDG divergence is a (scaled) Rényi divergence.

**Corollary 6.10.1** (Rényi Divergences). *For all  $r, s \in \bar{\mathbb{R}}$ , with  $r + s \geq 0$ , and all  $\alpha \in [0, \infty]$ , we have that*

$$\left\langle\left\langle \frac{p}{(r)} \rightarrow X \leftarrow \frac{q}{(s)} \right\rangle\right\rangle = s \cdot D_{\frac{r}{r+s}}(p \parallel q) \quad \text{and} \quad D_\alpha(p \parallel q) = \left\langle\left\langle \frac{p}{(\frac{\alpha}{1-\alpha})} \rightarrow X \leftarrow \frac{q}{(1-\alpha)} \right\rangle\right\rangle$$

However, the two classes are not identical, because the PDG divergences have extra limit points. One big difference is that the reverse KL divergence  $D(q \parallel p)$  is not a Rényi divergence  $D_\alpha(p \parallel q)$  for any value (or limit) of  $\alpha$ . This lack of symmetry has led others (e.g., Cichocki and Amari 2010) to work instead with a symmetric variant called  $\alpha$ -divergence, rescaled by an additional factor of  $\frac{1}{\alpha}$ . The relationships between these quantities can be seen in Figure 6.1.

The Chernoff divergence measures the tightest possible exponential bound on probability of error (Nielsen 2011) in Bayesian hypothesis testing. It also happens to be the smallest possible inconsistency of simultaneously believing  $p$  and  $q$ , with confidences whose sum equals one.

**Corollary 6.10.2.** *The Chernoff Divergence between  $p$  and  $q$  equals*

$$\inf_{\beta \in (0,1)} \left\langle \left\langle \xrightarrow[\text{(\beta)}]{p} X \xleftarrow[\text{(1-\beta)}]{q} Y \right\rangle \right\rangle.$$

One significant consequence of representing divergences as inconsistencies is that we can use [Lemma 6.1](#) to derive relationships between them. The following facts follow directly from [Figure 6.1](#), by inspection.

**Corollary 6.10.3.**

1. Rényi entropy is monotonic in its parameter  $\alpha$ .
2.  $D(p \parallel q) \geq 2D_B(p, q) \leq D(q \parallel p)$ .
3. If  $q(p > 0) < 1$  (i.e.,  $q \not\ll p$ ), then  $D(q \parallel p) = \infty$ .

These divergences correspond to PDGs with only two edges and one variable. What about more complex graphs? For a start, conditional divergences

$$D_{(r,s)}^{\text{PDG}}(p(Y|X) \parallel q(Y|X) \mid r(X)) := \mathbb{E}_{x \sim r} D_{(r,s)}^{\text{PDG}}(p(Y|x) \parallel q(Y|x))$$

can be represented straightforwardly as

$$D_{(r,s)}^{\text{PDG}}(p \parallel q \mid r) = \left\langle \left\langle \xrightarrow[\text{(\infty)}]{r} X \xrightarrow[p(r)]{q(s)} Y \right\rangle \right\rangle.$$

Other structures are useful intermediates. [Lemma 6.1](#), plus some structural manipulation, gives visual proofs of many divergence properties; [Figure 6.2](#) features such a proof of the data-processing inequality. And in general, PDG inconsistency can be viewed as a vast generalization of divergences to arbitrary structured objects.

$$\begin{aligned}
\left\langle \frac{p}{(\beta)} \rightarrow X \xleftarrow{q} \right\rangle &= \left\langle \frac{p}{(\beta)} \rightarrow X \xleftarrow{q} \xrightarrow{f(\beta+\zeta)} Y \right\rangle = \left\langle \frac{p}{(\beta)} \rightarrow X_1 = X_2 \xleftarrow{q} \right\rangle \\
&\geq \left\langle \frac{p}{(\beta)} \rightarrow X_1 \xrightarrow{f(\beta)} Y \xleftarrow{f} X_2 \xleftarrow{q} \right\rangle = \left\langle \frac{f \circ p}{(\beta)} \rightarrow X \xleftarrow{f \circ q} \right\rangle
\end{aligned}$$

Figure 6.2: A visual, monotonicity-based proof of the data-processing inequality for all PDG divergences:  $D_{(\beta,\zeta)}^{\text{PDG}}(p \parallel q) \geq D_{(\beta,\zeta)}^{\text{PDG}}(f \circ p \parallel f \circ q)$ . In words: the cpd  $f(Y|X)$  can always be satisfied, so adds no inconsistency. It is then equivalent to split  $f$  and the variable  $X$  into  $X_1$  and  $X_2$  with edges enforcing  $X_1 = X_2$ . But removing such edges can only decrease inconsistency. Finally, compose the remaining cpds to give the result. See the [Section 6.C.1](#) for a full justification.

## 6.5 Variational Objectives and Bounds

The fact that the incompatibility of  $\mathcal{M}$  with a *specific* joint distribution  $\mu$  is an upper bound on the inconsistency is not a deep one, but it is of a variational flavor. Here, we focus on the more surprising converse: PDG semantics capture general aspects of variational inference. Moreover, PDGs provide the basis of an intuitive graphical proof language for variational bounds.

### 6.5.1 PDGs and Variational Approximations

We begin by recounting the standard development of the ‘Evidence Lower BOund’ (ELBO), a standard objective for training latent variable models ([Blei et al. 2017, §2.2](#)). Suppose we have a model  $p(X, Z)$ , but only have access to observations of  $x$ . In service of adjusting  $p(X, Z)$  to make our observations more likely, we would like to maximize  $\log p(X = x)$ , the “evidence” of  $x$  ([Proposition 6.4](#)). Unfortunately, computing  $p(X) = \sum_z p(X, Z=z)$  requires summing

over all of  $Z$ , which can be intractable. The variational approach is as follows: fix a family of distributions  $\mathcal{Q}$  that is easy to sample from, choose some  $q(Z) \in \mathcal{Q}$ , and define  $\text{ELBO}_{p,q}(x) := \mathbb{E}_{z \sim q} \log \frac{p(x,z)}{q(z)}$ . This is something we can estimate, since we can sample from  $q$ . By Jensen's inequality,

$$\text{ELBO}_{p,q}(x) = \mathbb{E}_q \log \frac{p(x,Z)}{q(Z)} \leq \log \left[ \mathbb{E}_q \frac{p(x,Z)}{q(Z)} \right] = \log p(x),$$

with equality if  $q(Z) = p(Z)$ . So to find  $p$  maximizing  $p(x)$ , it suffices to adjust  $p$  and  $q$  to maximize  $\text{ELBO}_{p,q}(x)$ <sup>3</sup> provided  $\mathcal{Q}$  is expressive enough.

The formula for the ELBO is somewhat difficult to make sense of.<sup>4</sup> Nevertheless, it arises naturally as the inconsistency of the appropriate PDG.

**Proposition 6.11.** *The negative ELBO of  $x$  is the inconsistency of the PDG containing  $p, q$ , and  $X=x$ , with high confidence in  $q$ . That is,*

$$-\text{ELBO}_{p,q}(x) = \left\langle \begin{array}{c} q \\ \xrightarrow{(\infty)} \end{array} Z \xrightarrow[p]{\quad} X \xleftarrow{x} \right\rangle.$$

[ link to proof ]

Owing to its structure, a PDG is often more intuitive and easier to work with than the formula for its inconsistency. To illustrate, we now give a simple and visually intuitive proof of the bound traditionally used to motivate the ELBO, via Lemma 6.1:

$$\log \frac{1}{p(x)} = \left\langle \begin{array}{c} p \\ \xleftarrow{\quad} \end{array} Z \xrightarrow{x} X \right\rangle \leq \left\langle \begin{array}{c} q \\ \xleftarrow{\infty} \end{array} Z \xrightarrow[p]{\quad} X \xleftarrow{x} \right\rangle = -\text{ELBO}_{p,q}(x).$$

The first and last equalities are Propositions 6.4 and 6.11 respectively. Now to reap some pedagogical benefits. The second PDG has more edges so it is clearly at least as inconsistent. Furthermore, it's easy to see that equality holds when

---

<sup>3</sup>or for many iid samples:  $\max_{p,q} \sum_{x \in \mathcal{D}} \text{ELBO}_{p,q}(x)$ .

<sup>4</sup>Especially if  $p, q$  are densities. See Section 6.A.

$q(Z) = p(Z)$ : the best distribution for the left PDG has marginal  $p(Z)$  anyway, so insisting on it incurs no further cost.

### 6.5.2 Variational Auto-Encoders and PDGs

An autoencoder is a probabilistic model intended to compress a variable  $X$  (e.g., an image) to a compact latent representation  $Z$ . Its structure is given by two conditional distributions: an encoder  $e(Z|X)$ , and a decoder  $d(X|Z)$ . Of course, not all pairs of cpds fill this role equally well. One important consideration is the *reconstruction error* (6.5): when we decode an encoded image, we would like it to resemble the original.

$$\text{Rec}(x) := \mathbb{E}_{z \sim e(Z|x)} \underbrace{\mathbb{I}_{d(X|z)}(x)}_z = \sum_z e(z|x) \log \frac{1}{d(x|z)} \quad (6.5)$$

$\left( \begin{array}{l} \text{additional bits required to} \\ \text{decode } x \text{ from its encoding } z \end{array} \right)$

There are other desiderata as well. Perhaps good latent representations  $Z$  have uncorrelated components, and are normally distributed. We encode such wishful thinking as a belief  $p(Z)$ , known as a variational prior.

The data of a Variational Auto-Encoder ([Kingma and Welling 2014](#); [Rezende et al. 2014](#)), or VAE, consists of  $e(Z|X)$ ,  $d(X|Z)$ , and  $p(Z)$ . The encoder  $e(Z|X)$  can be used as a variational approximation of  $Z$ , differing from  $q(Z)$  of [Section 6.5.1](#) only in that it can depend on  $X$ . VAEs are trained with the analogous form of the ELBO:

$$\begin{aligned} \text{ELBO}_{p,e,d}(x) &:= \mathbb{E}_{z \sim e(Z|x)} \left[ \log \frac{p(z)d(x|z)}{e(z|x)} \right] \\ &= -\text{Rec}(x) - D(e(Z|x) \| p). \end{aligned}$$

This gives us the following analog of [Proposition 6.11](#).

**Proposition 6.12.** *The VAE loss of a sample  $x$  is the inconsistency of the PDG comprising the encoder  $e$  (with high confidence, as it defines the encoding), decoder  $d$ , prior  $p$ , and  $x$ . That is,*

$$-\text{ELBO}_{p,e,d}(x) = \left\langle \begin{array}{c} p \\ \rightarrow \\ Z \\ \downarrow \\ e \\ (\infty) \\ \leftarrow \\ d \\ X \\ \leftarrow \\ x \end{array} \right\rangle.$$

We now give a visual proof of the analogous variational bound. Let  $\Pr_{p,d}(X, Z) := p(Z)d(X|Z)$  be the distribution that arises from decoding the prior. Then:

$$\log \frac{1}{\Pr_{p,d}(x)} = \left\langle \begin{array}{c} p \\ \downarrow \\ Z \\ \leftarrow \\ e \\ (\infty) \\ \leftarrow \\ d \\ X \\ \leftarrow \\ x \end{array} \right\rangle \leq \left\langle \begin{array}{c} p \\ \downarrow \\ Z \\ \leftarrow \\ e \\ (\infty) \\ \leftarrow \\ d \\ X \\ \leftarrow \\ x \end{array} \right\rangle = -\text{ELBO}_{p,e,d}(x).$$

The first and last equalities are [Propositions 6.4](#) and [6.12](#), and the inequality is [Lemma 6.1](#). See the appendix for multi-sample analogs of the bound and [Proposition 6.12](#).

### 6.5.3 The $\beta$ -VAE Objective

The ELBO is not the only objective that has been used to train networks with a VAE structure. In the most common variant, due to [Higgins et al. \(2016\)](#), one weights the reconstruction error (6.5) and the ‘KL term’ differently, resulting in a loss function of the form

$$\beta\text{-ELBO}_{p,e,d}(x) := -\text{Rec}(x) - \beta D(e(Z|x) \| p),$$

which, when  $\beta=1$ , is the ELBO as before. The authors view  $\beta$  as a regularization strength, and argue that it sometimes helps to have a stronger prior. Sure enough:

**Proposition 6.13.**  $-\beta\text{-ELBO}_{p,e,d}(x)$  is the inconsistency of the same PDG, but with confidence  $\beta$  in  $p(Z)$ .

## 6.6 Free Energy as Factor Graph Inconsistency

A weighted factor graph  $\Psi = (\phi_J, \theta_J)_{J \in \mathcal{J}}$ , where each  $\theta_J$  is a real-valued weight,  $J$  is associated with a subset of variables  $\mathbf{X}_J$ , and  $\phi_J : \mathcal{V}(\mathbf{X}_J) \rightarrow \mathbb{R}$ , determines a distribution by

$$\Pr_\Psi(\mathbf{x}) = \frac{1}{Z_\Psi} \prod_{J \in \mathcal{J}} \phi_J(\mathbf{x}_J)^{\theta_J}.$$

$Z_\Psi$  is the constant  $\sum_{\mathbf{x}} \prod_{J \in \mathcal{J}} \phi_J(\mathbf{x}_J)^{\theta_J}$  required to normalize the distribution, and is known as the *partition function*. Computing  $\log Z_\Psi$  is intimately related to probabilistic inference in factor graphs (Ma et al. 2013). We will revisit this point in Chapter 9. Following Richardson and Halpern (2021), let  $\mathbf{m}_\Psi$  be the PDG with edges  $\{\overset{J}{\rightarrow} \mathbf{X}_J\}_{\mathcal{J}}$ , cpds  $p_J(\mathbf{X}_J) \propto \phi_J(\mathbf{X}_J)$ , and weights  $\alpha_J, \beta_J := \theta_J$ . There, it is shown that  $\Pr_\Psi$  is the unique minimizer of  $\|\mathbf{m}_\Psi\|_1$ . But what about the corresponding inconsistency,  $\langle\!\langle \mathbf{m}_\Psi \rangle\!\rangle_1$ ?

If the factors are normalized and all variables are edge targets, then  $Z_\Psi \leq 1$ , so  $\log \frac{1}{Z_\Psi} \geq 0$  measures how far the product of factors is from being a probability distribution. So in a sense, it measures  $\Psi$ 's inconsistency.

**Proposition 6.14.** For all weighted factor graphs  $\Psi$ , we have that  $\langle\!\langle \mathbf{m}_\Psi \rangle\!\rangle_1 = -\log Z_\Psi$ .

[ link to proof ]

The exponential families generated by weighted factor graphs are a cornerstone of statistical mechanics, where  $-\log Z_\Psi$  is known as the (Heimholz) free energy. It is also an especially natural quantity to minimize: the principle of free-energy minimization has been enormously successful in describing of not only

chemical and biological systems ([Chipot and Pohorille 2007](#)), but also cognitive ones ([Friston 2009](#)).

## 6.7 Beyond Standard Losses: A Concrete Example

In contexts where a loss function is standard, it is usually for good reason—which is why we have focused on recovering standard losses. But most situations are non-standard, and even if they have standard sub-components, those components may interact with one another in more than one way. Correspondingly, there is generally more than one way to cobble standard loss functions together. How should you choose between them? By giving a principled model of the situation.

Suppose we want to train a predictor network  $h(Y|X)$  from two sources of information: partially corrupted data with distribution  $d(X, Y)$ , and a simulation with distribution  $s(X, Y)$ . If the simulation is excellent and the data unsalvageable, we would have high confidence in  $s$  and low confidence in  $d$ , in which case we would train with cross entropy with respect to  $s$ ,  $\mathcal{L}_{\text{sim}} := \mathbb{E}_s[\log^1/h(Y|X)]$ . Conversely, if the simulation were bad and the data mostly intact, we would use  $\mathcal{L}_{\text{dat}}$ , the cross entropy with respect to  $d$ . What if we’re not so confident in either?

One approach a practitioner might find attractive is to make a dataset from samples of both  $s$  and  $d$ , or equivalently, train with a convex combination of the two previous losses,  $\mathcal{L}_1 := \lambda_s \mathcal{L}_{\text{sim}} + \lambda_d \mathcal{L}_{\text{dat}}$  for some  $\lambda_s, \lambda_d > 0$  with  $\lambda_s + \lambda_d = 1$ . This amounts to training  $h$  with cross entropy with respect to the mixture  $\lambda_s s + \lambda_d d$  ([Claim 6.7.1](#) in the appendix). Doing so treats  $d$  and  $s$  as completely unrelated, and so redundancy is not used to correct errors—a fact on display

when we present the modeling choices in PDG form, such as

$$\mathcal{L}_1 = \left\langle \left. \begin{array}{c} \lambda \\ (\infty) \end{array} \right\rightarrow \boxed{\begin{array}{c} Z \\ \bullet \quad \bullet \\ \text{sim dat} \end{array}} \right. \xrightarrow[\text{(}\infty\text{)}]{} \begin{array}{c} X \\ h \\ Y \end{array} \right\rangle,$$

in which a switch variable  $Z$  with possible values  $\{\text{sim}, \text{dat}\}$  controls whether samples come from  $s$  or  $d$ , and is distributed according to  $\lambda(Z = \text{sim}) = \lambda_s$ .

Our practitioner now tries a different approach: draw data samples  $(x, y) \sim d$  but discount  $h$ 's surprisal when the simulator finds the point unlikely, via loss  $\mathcal{L}_2 := \mathbb{E}_d[s(X, Y) \log^{1/h(Y|X)}]$ . This is the cross entropy with respect to the (un-normalized) product density  $ds$ , which in many ways is appropriate. However, by this metric, the optimal predictor is  $h^*(Y|x) \propto d(Y|x)s(Y|x)$ , which is *uncalibrated* (Dawid 1982). If the data and simulator agree ( $d = s$ ), then we would want  $h(Y|x) = s(Y|x)$  for all  $x$ , but instead we get  $h^*(Y|x) \propto s(Y|x)^2$ . So  $h^*$  is overconfident. What went wrong?  $\mathcal{L}_2$  cannot be written as the (ordinary  $\gamma=0$ ) inconsistency of a PDG containing only  $s$ ,  $h$ , and  $d$ , but for a large fixed  $\gamma$ , it is essentially the  $\gamma$ -inconsistency

$$\mathcal{L}_2 \approx C \left\langle \left. \begin{array}{c} s \\ (\alpha:1) \\ \beta:\gamma \end{array} \right\leftarrow \begin{array}{c} X \\ h \\ Y \end{array} \right. \xleftarrow[\text{(}\beta:\gamma\text{)}]{} \begin{array}{c} d \\ (\alpha:1) \end{array} \right\rangle_\gamma + \text{const},$$

where  $C$  is the constant required to normalize the joint density  $sd$ , and *const* does not depend on  $h$  (Claim 6.7.2). However, the values of  $\alpha$  in this PDG indicate an over-determination of  $XY$  (it is determined in two different ways), and so  $h^*$  is more deterministic than intended. By contrast,

$$\mathcal{L}_3 := \left\langle \left. \begin{array}{c} s \\ (\lambda_s) \end{array} \right\leftarrow \begin{array}{c} X \\ h \\ Y \end{array} \right. \xleftarrow[\text{(}\lambda_d\text{)}]{} \begin{array}{c} d \end{array} \right\rangle,$$

does not have this issue: the optimal predictor  $h^*$  according to  $\mathcal{L}_3$  is proportional to the  $\lambda$ -weighted geometric mean of  $s$  and  $d$  (Claim 6.7.3). It seems that our

approach, in addition to providing a unified view of standard loss functions, can also suggest more appropriate loss functions in practical situations.

## 6.8 Reverse-Engineering a Loss Function?

Given an *arbitrary* loss function, can we find a PDG that gives rise to it? The answer, as we discovered in [Section 4.2.2](#), appears to be yes—although not without making unsavory modeling choices. Without affecting its semantics, one may add the variable  $T$  that takes values  $\{t, f\}$ , and the event  $T = t$ , to any PDG. Now, given a cost function  $c : \mathcal{V}(X) \rightarrow \mathbb{R}_{\geq 0}$ , define the cpd  $\hat{c}(T|X)$  by  $\hat{c}(t|x) := e^{-c(x)}$ . By threatening to generate the falsehood  $f$  with probability dependent on the cost of  $X$ ,  $\hat{c}$  ties the value of  $X$  to inconsistency.

**Proposition 6.15.**  $\left\langle \xrightarrow[\infty]{p} X \xrightarrow{\hat{c}} T \xleftarrow{t} \right\rangle = \mathbb{E}_{x \sim p}[c(x)]$ .

[[link to proof](#)]

Setting confidence  $\beta_p := \infty$  may not be realistic since we’re still training the model  $p$ , but doing so is necessary to recover  $\mathbb{E}_p c$ .<sup>5</sup> Any mechanism that generates inconsistency based on the value of  $X$  (such as this one) also works in reverse: the PDG “squirms”, contorting the probability of  $X$  to disperse the inconsistency. One cannot simply “emit loss” without affecting the rest of the model, as one does with utility in an Influence Diagram ([Howard 1983](#)). Even setting every  $\beta := \infty$  may not be enough to prevent the squirming. To illustrate, consider a model  $\mathcal{S}$  of the supervised learning setting (predict  $Y$  from  $X$ ), with labeled data  $\mathcal{D}$ , model  $h$ , and a loss function  $\ell$  on pairs of output labels.

---

<sup>5</sup>If  $\beta_p$  were instead equal to 1, we would have obtained  $-\log \mathbb{E}_p \exp(-c(X))$ , with optimal distribution  $\mu(X) \neq p(X)$  ([Claim 6.8.1](#) in the appendix).

Concretely, define:

$$\mathcal{S} := \Pr_{\mathcal{D}}[Y] \xrightarrow{\hat{\ell}} \text{T} \quad \text{and} \quad \mathcal{L} := \mathbb{E}_{\substack{(x,y) \sim \Pr_{\mathcal{D}} \\ y' \sim p(Y'|x)}} [\ell(y, y')].$$

Given [Proposition 6.15](#), one might imagine  $\langle\langle \mathcal{S} \rangle\rangle = \mathcal{L}$ , but this is not so. In some ways,  $\langle\langle \mathcal{S} \rangle\rangle$  is actually preferable. The optimal  $h(Y'|X)$  according to  $\mathcal{L}$  is a degenerate cpd that places all mass on the label(s)  $y_X^*$  minimizing expected loss, while the optimal  $h(Y'|X)$  according to  $\langle\langle \mathcal{S} \rangle\rangle$  is  $\Pr_{\mathcal{D}}(Y|X)$ , which means that it is calibrated, unlike  $\ell$ . If, in addition, we set  $\alpha_p, \alpha_{\Pr_{\mathcal{D}}} := 1$  and strictly enforce the qualitative picture, finally no more squirming is possible, as we arrive at  $\lim_{\gamma \rightarrow \infty} \langle\langle \mathcal{S} \rangle\rangle_{\gamma} = \mathcal{L}$  ([Claim 6.8.2](#)).

In the process, we have given up our ability to tolerate inconsistency by setting all probabilistic modeling choices in stone. What's more, we've dragged in the global parameter  $\gamma$ , further handicapping our ability to compose this model with others. To summarize: while model inconsistency readily generates appropriate loss functions, the converse does not work as well. Reverse-engineering a loss may require making questionable modeling choices with absolute certainty, resulting in brittle models with limited potential for composition. In the end, we must confront our modeling choices; good loss functions come from good models.

## 6.9 Conclusions

We seen that that PDG semantics not only capture Bayesian Networks and Factor Graphs ([Richardson and Halpern 2021](#)), but also generate many standard loss functions, including some non-trivial ones. In each case, the appropriate

loss arises simply by articulating modeling assumptions, and then measuring inconsistency. Viewing loss functions in this way also has beneficial side effects, including an intuitive visual proof language for reasoning about the relationships between them.

This “universal loss”, which provides a principled way of choosing an optimization objective, may be of particular interest to the AI alignment community.

## APPENDICES FOR CHAPTER 6

### 6.A Notes

**The Fine Print for Probability Densities.** Several of our results ([Propositions 6.2 to 6.5, 6.11, 6.12, 6.17, 6.19](#) and [6.20](#)) technically require the distribution to be represented with a mass function (as opposed to a probability density function, or pdf). A PDG containing both pdf and a finitely supported distribution on the same variable will typically have infinite inconsistency—but this is not just a quirk of the PDG formalism.

Probability density is not dimensionless (like probability mass), but rather has inverse  $X$ -units (e.g., probability per meter), so depends on an arbitrary choice of scale (the pdf for probability per meter and per centimeter will yield different numbers). In places where the objective does not have units that cancel before we take a logarithm, the use of a probability density  $p(X)$  becomes sensitive to this arbitrary choice of parameterization. For instance, the analog of surprisal,  $-\log p(x)$  for a pdf  $p$ , or its expectation, called differential entropy, both depend on an underlying scheme of measurement (an implicit base measure).

On the other hand, this choice of scale ultimately amounts to an additive constant. Moreover, beyond a certain point, decreasing the discretization size  $k$  of a discretized approximation  $\tilde{p}_k(X)$  *also* contributes a constant that depends only on  $k$ . But such constants are irrelevant for optimization, and so, even though such quantities are ill-defined and arguably meaningless in the continuous limit, the use of the continuous analogs as loss functions is still justified.

The bottom line is that all our results hold in a uniform way for every dis-

cretization size — yet in the limit as the discretization becomes smaller, an inconsistency may diverge to infinity. However, this divergence stems from an additive constant that depends only on the discretization size, which is irrelevant to its employment as a loss function. As a result, using one of these “unbalanced” functions involving densities where the units do not work out properly, results in a morally equivalent loss function, except without a diverging constant. This is a direct analogue of our discussion of entropy for continuous variables (Section 2.6.1).

**Maximum A Posteriori (MAP) updating, Priors, and Regularization.** The usual telling of the correspondence between regularizers and priors is something like the following. Suppose you have a parameterized family of distributions  $\Pr(X|\Theta)$  and have observed evidence  $X$ , but do not know the parameter  $\Theta$ . The maximum-likelihood estimate of  $\Theta$  is then

$$\theta^{\text{MLE}}(X) := \arg \max_{\theta \in \Theta} \Pr(X | \theta) = \arg \max_{\theta \in \Theta} \log \Pr(X | \theta).$$

The logarithm is a monotonic transformation, so it does not change the argmax, but it has nicer properties, so that function is generally used instead. (Many of the loss functions in main body of the paper are log-likelihoods also.)

In some sense, better than estimating the maximum likelihood, is to perform a Bayesian update with the new information, to get a *distribution* over  $\Theta$ . If that’s too expensive, we could simply take the estimate with the highest posterior probability, which is called the Maximum A Posteriori (MAP) estimate. For any given  $\theta$ , the Bayesian reading of Bayes rule states that

$$\text{posterior } \Pr(\Theta|X) = \frac{\text{likelihood } \Pr(X|\Theta) \cdot \text{prior } \Pr(\Theta)}{\text{evidence } \Pr(X) = \sum_{\theta'} \Pr(X|\theta') \Pr(\theta')}.$$

So taking a logarithm,

$$\text{log-posterior} = \frac{\text{log-likelihood}}{\log \Pr(X|\Theta)} + \frac{\text{log-prior}}{\log \Pr(\Theta)} - \frac{\text{log-evidence}}{\log \Pr(X)}.$$

The final term does not depend on  $\theta$ , so it is not relevant for finding the optimal  $\theta$  by this metric. Swapping the signs so that we are taking a minimum rather than a maximum, the MAP estimate is then given by

$$\theta^{\text{MAP}}(X) := \arg \min_{\theta \in \Theta} \left\{ \log \frac{1}{\Pr(X|\theta)} + \log \frac{1}{\Pr(\theta)} \right\}.$$

Note that if negative log likelihood (or surprisal,  $-\log \Pr(X|\theta)$ ) was our original loss function, we have now added an arbitrary extra term, as a function of  $\Theta$ , to our loss function. It is in this sense that priors classically correspond to regularizers.

## 6.B Further Results and Generalizations

### 6.B.1 Full Characterization of Gaussian Predictors

The inconsistency of a PDG containing two univariate Gaussian regressors of with arbitrary parameters and confidences, is most cleanly articulated in terms of the geometric and quadratic means.

**Definition 6.1** (Weighted Power Mean). The weighted power mean  $M_p^w(\mathbf{r})$  of the collection of real numbers  $\mathbf{r} = r_1, \dots, r_n$  with respect to the convex weights  $w = w_1, \dots, w_n$  satisfying  $\sum_i w_i = 1$ , is given by

$$M_p^w(\mathbf{r}) := \left( \sum_{i=1}^n w_i (r_i)^p \right)^{\frac{1}{p}}.$$

We omit the superscript as a shorthand for the uniform weighting  $w_i = 1/N$ .  $\square$

Name	$p$	Formula
Harmonic	( $p = -1$ ):	$\text{HM}_w(\mathbf{r}) = \frac{1}{\sum_{i=1}^n w_i/r_i}$
Geometric	( $\lim p \rightarrow 0$ ):	$\text{GM}_w(\mathbf{r}) = \prod_{i=1}^n r_i^{w_i}$
Arithmetic	( $p = 1$ ):	$\text{AM}_w(\mathbf{r}) = \sum_{i=1}^n w_i r_i$
Quadratic	( $p = 2$ ):	$\text{QM}_w(\mathbf{r}) = \sqrt{\sum_{i=1}^n w_i r_i^2}$

Table 6.B.1: special cases of the  $p$ -power mean  $M_p^w(\mathbf{r})$

Many standard means, such as those in Table 6.B.1, are special cases. It is well known that  $M_p^w(\mathbf{r})$  is increasing in  $p$ , and strictly so if not all elements of  $\mathbf{r}$  are identical. In particular,  $\text{QM}_w(a, b) > \text{GM}_w(a, b)$  for all  $a \neq b$  and positive weights  $w$ . We now present the result.

**Proposition 6.16.** Consider a PDG containing two (distinct) conditional Gaussian distributions on a variable  $Y$ , whose parameters can both depend on a variable  $X$ . Its inconsistency takes the form

[ link to proof ]

$$\begin{aligned}
 &= \mathbb{E}_D \left[ (\beta_1 + \beta_2) \log \frac{\text{QM}_{\hat{\beta}}(\sigma_1, \sigma_2)}{\text{GM}_{\hat{\beta}}(\sigma_1, \sigma_2)} + \frac{1}{2} \frac{\beta_1 \beta_2}{\beta_1 + \beta_2} \left( \frac{\mu_1 - \mu_2}{\text{QM}_{\hat{\beta}}(\sigma_1, \sigma_2)} \right)^2 \right] \\
 &= \frac{1}{2} \mathbb{E}_{x \sim D} \left[ \frac{\beta_1 \beta_2}{2} \frac{(f(x) - h(x))^2}{\beta_2 s(x)^2 + \beta_1 t(x)^2} + \frac{\beta_1 + \beta_2}{2} \log \frac{\beta_2 s(x)^2 + \beta_1 t(x)^2}{\beta_1 + \beta_2} - \beta_2 \log s(x) - \beta_1 \log t(x) \right]
 \end{aligned} \tag{6.6}$$

where  $\hat{\beta} = (\frac{\beta_2}{\beta_1 + \beta_2}, \frac{\beta_1}{\beta_1 + \beta_2})$  represents the normalized and reversed vector of confidences  $\beta = (\beta_1, \beta_2)$  for the two distributions, and  $\mu_1 = f(X)$ ,  $\mu_2 = g(X)$ ,  $\sigma_1 = s(X)$ ,  $\sigma_2 = t(X)$  are random variables over  $X$ .

The PDG on the left is semantically equivalent to (and in particular has the same inconsistency as) the PDG

$$\begin{array}{c} \mathcal{N}(f(x), s(x)) \\ \xrightarrow[\infty]{D} X \xrightarrow{\quad} Y \\ \mathcal{N}(h(x), t(x)) \end{array}$$

This illustrates an orthogonal point: that PDGs handle composition of functions as one would expect, so that it is equivalent to model an entire process as a single arrow, or to break it into stages, ascribing an arrow to each stage, with one step of randomization.

As a bonus, Proposition 6.16 also gives a proof of the inequality of the weighted geometric and quadratic means.

**Corollary 6.16.1.** For all  $\sigma_1$  and  $\sigma_2$ , and all weight vectors  $\beta$ ,  $\text{QM}_{\hat{\beta}}(\sigma_1, \sigma_2) \geq \text{GM}_{\hat{\beta}}(\sigma_1, \sigma_2)$ .

### 6.B.2 Full-Dataset ELBO and Bounds

We now present the promised multi-sample analogs from Section 6.5.1.

**Proposition 6.17.** The following analog of Proposition 6.12 for a whole dataset  $\mathcal{D}$  holds:

[ link to proof ]

$$-\mathbb{E}_{\Pr_{\mathcal{D}}} \text{ELBO}_{p,e,d}(X) = \left\langle \xrightarrow[p]{Z} \xleftarrow[e]{\kappa^{(\infty)}} \xrightarrow[d]{X} \xleftarrow[(\infty)]{\Pr_{\mathcal{D}}} \right\rangle + H(\Pr_{\mathcal{D}}).$$

Propositions 6.3 and 6.17 then give us an analog of the visual bounds in the body of the main paper (Section 6.5.1) for many i.i.d. data points at once, with only a single application of the inequality:

$$\begin{aligned} -\log \Pr(\mathcal{D}) &= -\log \prod_{i=1}^m (\Pr(x^{(i)})) = -\frac{1}{m} \sum_{i=1}^m \log \Pr(x^{(i)}) = \\ &= H(\Pr_{\mathcal{D}}) + \left\langle \xrightarrow[p]{Z} \xrightarrow[d]{X} \xleftarrow[(\infty)]{\Pr_{\mathcal{D}}} \right\rangle \leq \left\langle \xrightarrow[p]{Z} \xleftarrow[e]{\kappa^{(\infty)}} \xrightarrow[d]{X} \xleftarrow[(\infty)]{\Pr_{\mathcal{D}}} \right\rangle + H(\Pr_{\mathcal{D}}) \\ &= -\mathbb{E}_{\Pr_{\mathcal{D}}, p,e,d} \text{ELBO}(X) \end{aligned}$$

We also have the following formal statement of Proposition 6.13.

**Proposition 6.18.** The negative  $\beta$ -ELBO objective for a prior  $p(X)$ , encoder  $e(Z | X)$ , decoder  $d(X | Z)$ , at a sample  $x$ , is equal to the inconsistency of the corresponding PDG,

[ link to proof ]

where the prior has confidence equal to  $\beta$ . That is,

$$-\beta\text{-ELBO}_{p,e,d}(x) = \left\langle \begin{array}{c} p \\ \xrightarrow{(\beta)} \\ Z \end{array} \xrightarrow[d]{\curvearrowright} X \xleftarrow[e]{\curvearrowleft(\infty)} \xleftarrow{x} \right\rangle$$

As a specific case (i.e., effectively by setting  $\beta_p := 0$ ), we get the reconstruction error as the inconsistency of an autoencoder (without a variational prior):

**Corollary 6.18.1** (reconstruction error as inconsistency).

$$-\text{Rec}_{ed,d}(x) := \mathbb{E}_{z \sim e(Z|x)} I_{d(X|z)}(x) = \left\langle \begin{array}{c} d \\ \curvearrowright \\ Z \end{array} \xrightarrow[e]{\curvearrowleft(\infty)} X \xleftarrow{x} \right\rangle$$

### 6.B.3 More Variants of Log Likelihood Results

First, we show that our cross entropy results hold for all  $\gamma$ , in the sense that  $\gamma$  contributes only a constant.

**Proposition 6.19.** *Given a model determining a probability distribution with mass function  $p(X)$ , and samples  $\mathcal{D} = \{x_i\}_{i=1}^m$  determining an empirical distribution  $\Pr_{\mathcal{D}}$ , the following are equal, for all  $\gamma \geq 0$ :*

link to  
proof

1. The average negative log likelihood  $\ell(p; \mathcal{D}) = -\frac{1}{m} \sum_{i=1}^m \log p(x_i)$
2. The cross entropy of  $p$  relative to  $\Pr_{\mathcal{D}}$
3.  $\llbracket p \rrbracket_{\gamma}(\Pr_{\mathcal{D}}) + (1 + \gamma) H(\Pr_{\mathcal{D}})$
4.  $\left\langle \xrightarrow[p]{\Pr_{\mathcal{D}}} X \xleftarrow[(\infty)]{} \right\rangle_{\gamma} + (1 + \gamma) H(\Pr_{\mathcal{D}})$

As promised, we now give the simultaneous generalization of the surprisal result ([Proposition 6.2](#)) to both multiple samples (like in [Proposition 6.3](#)) and partial observations (as in [Proposition 6.4](#)).

**Proposition 6.20.** *The average marginal negative log likelihood  $\ell(p; x) := -\frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \log \sum_z p(x, z)$  is the inconsistency of the PDG containing  $p$  and the data distribution  $\Pr_{\mathcal{D}}$ , plus the entropy of the data distribution (which is constant in  $p$ ). That is,*

$$\ell(p; \mathcal{D}) = \left\langle \begin{array}{c} Z \\ \nearrow p \\ X \\ \xleftarrow[\infty]{} \Pr_{\mathcal{D}} \end{array} \right\rangle + H(\Pr_{\mathcal{D}}).$$

link to  
proof

## 6.C Proofs

**Lemma 6.1.** *Suppose PDGs  $m$  and  $m'$  differ only in their edges (resp.  $A$  and  $A'$ ) and confidences (resp.  $\beta$  and  $\beta'$ ). If  $A \subseteq A'$  and  $\beta_a \leq \beta'_a$  for all  $a \in A$ , then  $\langle\!\langle m \rangle\!\rangle_\gamma \leq \langle\!\langle m' \rangle\!\rangle_\gamma$  for all  $\gamma$ .*

*Proof.* For every  $\mu$ , adding more edges only adds non-negative terms to (3.1), while increasing  $\beta$  results in larger coefficients on the existing (non-negative) terms of (3.1). So for every fixed distribution  $\mu$ , we have  $\|m\|_\gamma(\mu) \leq \|m'\|_\gamma(\mu)$ . So it must also be the case that the infimum over  $\mu$ , so we find that  $\langle\!\langle m \rangle\!\rangle \leq \langle\!\langle m' \rangle\!\rangle$ .  $\square$

**Proposition 6.2.** *Consider a distribution  $p(X)$ . The inconsistency of the PDG comprising  $p$  and  $X=x$  equals the surprisal  $I_p[X=x]$ . That is,*

$$I_p[X=x] = \left\langle \begin{array}{c} p \\ \rightarrow \\ X \\ \leftrightarrow^x \end{array} \right\rangle.$$

(Recall that  $\langle\!\langle m \rangle\!\rangle$  is the inconsistency of the PDG  $m$ .)

*Proof.* Any distribution  $\mu(X)$  that places mass on some  $x' \neq x$  will have infinite KL divergence from the point mass on  $x$ . Thus, the only possibility for a finite consistency arises when  $\mu = \delta_x$ , and so

$$\left\langle\left\langle \xrightarrow{p} [X] \xleftarrow{x} \right\rangle\right\rangle = \left[ \xrightarrow{p} [X] \xleftarrow{x} \right](\delta_x) = D(\delta_x \parallel p) = \log \frac{1}{p(x)} = I_p(x).$$

□

Proposition 6.19 is a generalization of Proposition 6.3, so we prove them at the same time.

**Proposition 6.3.** *If  $p(X)$  is a probabilistic model of  $X$ , and  $\mathcal{D} = \{x_i\}_{i=1}^m$  is a dataset with empirical distribution  $\Pr_{\mathcal{D}}$ , then*

$$\text{CrossEntropy}(\Pr_{\mathcal{D}}, p) := \frac{1}{m} \sum_{i=1}^m I_p[X=x_i] = \left\langle\left\langle \xrightarrow{p} [X] \xleftarrow{\Pr_{\mathcal{D}}} \right\rangle\right\rangle + H(\Pr_{\mathcal{D}}).$$

**Proposition 6.19.** *Given a model determining a probability distribution with mass function  $p(X)$ , and samples  $\mathcal{D} = \{x_i\}_{i=1}^m$  determining an empirical distribution  $\Pr_{\mathcal{D}}$ , the following are equal, for all  $\gamma \geq 0$ :*

1. *The average negative log likelihood  $\ell(p; \mathcal{D}) = -\frac{1}{m} \sum_{i=1}^m \log p(x_i)$*
2. *The cross entropy of  $p$  relative to  $\Pr_{\mathcal{D}}$*
3.  $\llbracket p \rrbracket_{\gamma}(\Pr_{\mathcal{D}}) + (1 + \gamma) H(\Pr_{\mathcal{D}})$
4.  $\left\langle\left\langle \xrightarrow{p} [X] \xleftarrow{\Pr_{\mathcal{D}}} \right\rangle\right\rangle_{\gamma} + (1 + \gamma) H(\Pr_{\mathcal{D}})$

*Proof.* The equality of 1 and 2 is standard. The equality of 3 and 4 can be seen by the fact that in the limit of infinite confidence on  $\Pr_{\mathcal{D}}$ , the optimal distribution

must also equal  $\Pr_{\mathcal{D}}$ , so the least inconsistency is attained at this value. Finally it remains to show that the first two and last two are equal:

$$\begin{aligned} \llbracket p \rrbracket_{\gamma}(\Pr_{\mathcal{D}}) + (1 + \gamma) H(\Pr_{\mathcal{D}}) &= D(\Pr_{\mathcal{D}} \parallel p) - \gamma H(\Pr_{\mathcal{D}}) + (1 + \gamma) H(\Pr_{\mathcal{D}}) \\ &= D(\Pr_{\mathcal{D}} \parallel p) + H(\Pr_{\mathcal{D}}) \\ &= \mathbb{E}_{\Pr_{\mathcal{D}}} \left[ \log \frac{\Pr_{\mathcal{D}}}{p} + \log \frac{1}{\Pr_{\mathcal{D}}} \right] = \mathbb{E}_{\Pr_{\mathcal{D}}} \left[ \log \frac{1}{p} \right], \end{aligned}$$

which is the cross entropy, as desired.  $\square$

**Proposition 6.4.** *If  $p(X, Z)$  is a joint distribution, then the information content of the partial observation  $X = x$  is given by*

$$I_p[X=x] = \left\langle \begin{array}{c} \text{Z} \xrightarrow{p} \text{X} \xleftarrow{x} \end{array} \right\rangle. \quad (6.1)$$

*Proof.* As before, all mass of  $\mu$  must be on  $x$  for it to have a finite score. Thus it suffices to consider joint distributions of the form  $\mu(X, Z) = \delta_x(X)\mu(Z)$ . We have

$$\begin{aligned} \left\langle \begin{array}{c} \text{Z} \xrightarrow{p} \text{X} \xleftarrow{x} \end{array} \right\rangle &= \inf_{\mu(Z)} \left[ \begin{array}{c} \text{Z} \xrightarrow{p} \text{X} \xleftarrow{x} \end{array} \right] (\delta_x(X)\mu(Z)) \\ &= \inf_{\mu(Z)} D(\delta_x(X)\mu(Z) \parallel p(X, Z)) \\ &= \inf_{\mu(Z)} \mathbb{E}_{z \sim \mu} \log \frac{\mu(z)}{p(x, z)} = \inf_{\mu(Z)} \mathbb{E}_{z \sim \mu} \log \frac{\mu(z)}{p(x, z)} \frac{p(x)}{p(x)} \\ &= \inf_{\mu(Z)} \mathbb{E}_{z \sim \mu} \left[ \log \frac{\mu(z)}{p(z \mid x)} + \log \frac{1}{p(x)} \right] \\ &= \inf_{\mu(Z)} \left[ D(\mu(Z) \parallel p(Z \mid x)) \right] + \log \frac{1}{p(x)} \\ &= \log \frac{1}{p(x)} = I_p(x) \qquad \qquad \qquad [\text{Gibbs Inequality}] \end{aligned}$$

$\square$

**Proposition 6.20.** *The average marginal negative log likelihood  $\ell(p; x) := -\frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \log \sum_z p(x, z)$  is the inconsistency of the PDG containing  $p$  and the data*

distribution  $\Pr_{\mathcal{D}}$ , plus the entropy of the data distribution (which is constant in  $p$ ). That is,

$$\ell(p; \mathcal{D}) = \left\langle \begin{array}{c} Z \\ \swarrow^p \\ X \end{array} \xleftarrow{\Pr_{\mathcal{D}}} \right\rangle + H(\Pr_{\mathcal{D}}).$$

*Proof.* The same idea as in [Proposition 6.4](#), but a little more complicated.

$$\begin{aligned} \left\langle \begin{array}{c} Z \\ \swarrow^p \\ X \end{array} \xleftarrow{\Pr_{\mathcal{D}}} \right\rangle &= \inf_{\mu(Z|X)} \left[ \begin{array}{c} Z \\ \swarrow^p \\ X \end{array} \xleftarrow{\Pr_{\mathcal{D}}} \right] (\Pr_{\mathcal{D}}(X) \mu(Z | X)) \\ &= \inf_{\mu(Z|X)} D(\Pr_{\mathcal{D}}(X) \mu(Z | X) \| p(X, Z)) \\ &= \inf_{\mu(Z|X)} \mathbb{E}_{\substack{x \sim \Pr_{\mathcal{D}} \\ z \sim \mu}} \log \frac{\mu(z | x) \Pr_{\mathcal{D}}(x)}{p(x, z)} \\ &= \frac{1}{|\mathcal{D}|} \inf_{\mu(Z|X)} \sum_{x \in \mathcal{D}} \mathbb{E}_{z \sim \mu(Z|x)} \log \frac{\mu(z | x) \Pr_{\mathcal{D}}(x)}{p(x, z)} \frac{p(x)}{p(x)} \\ &= \frac{1}{|\mathcal{D}|} \inf_{\mu(Z|X)} \sum_{x \in \mathcal{D}} \left[ \mathbb{E}_{z \sim \mu} \left[ \log \frac{\mu(z | x)}{p(z | x)} \right] + \log \frac{1}{p(x)} - \log \frac{1}{\Pr_{\mathcal{D}}(x)} \right] \\ &= \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \left[ \inf_{\mu(Z|x)} \mathbb{E}_{z \sim \mu} \left[ \log \frac{\mu(z | x)}{p(z | x)} \right] + \log \frac{1}{p(x)} - \log \frac{1}{\Pr_{\mathcal{D}}(x)} \right] \\ &= \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \left[ \inf_{\mu(Z)} [D(\mu(Z) \| p(Z | x))] + \log \frac{1}{p(x)} \right] - H(\Pr_{\mathcal{D}}) \\ &= \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \log \frac{1}{p(x)} - H(\Pr_{\mathcal{D}}) \\ &= \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} I_p(x) - H(\Pr_{\mathcal{D}}) \\ &\quad \left( \quad = D(\Pr_{\mathcal{D}} \| p) \quad \right) \quad \square \end{aligned}$$

**Proposition 6.5.** *The inconsistency of the PDG comprising a probabilistic predictor  $h(Y|X)$ , and a high-confidence empirical distribution  $\Pr_{\mathcal{D}}$  of a dataset  $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m$  equals the cross-entropy loss (minus the empirical uncertainty in  $Y$  given  $X$ , a constant depending only on  $\mathcal{D}$ ). That is,*

$$\left\langle \begin{array}{c} \Pr_{\mathcal{D}} \\ \xrightarrow{\quad h \quad} \\ X \end{array} \xrightarrow{(\infty)} Y \right\rangle = \frac{1}{m} \sum_{i=1}^m \log \frac{1}{h(y_i | x_i)} - H_{\Pr_{\mathcal{D}}}(Y|X).$$

*Proof.*  $\Pr_{\mathcal{D}}$  has high confidence, it is the only joint distribution  $\mu$  with finite score.

Since  $f$  is the only other edge, the inconsistency is therefore

$$\begin{aligned} \mathbb{E}_{x \sim \Pr_{\mathcal{D}}} D\left(\Pr_{\mathcal{D}}(Y | x) \parallel f(Y | x)\right) &= \mathbb{E}_{x, y \sim \Pr_{\mathcal{D}}} \left[ \log \frac{\Pr_{\mathcal{D}}(y | x)}{f(y | x)} \right] \\ &= \mathbb{E}_{x, y \sim \Pr_{\mathcal{D}}} \left[ \log \frac{1}{f(y | x)} - \log \frac{1}{\Pr_{\mathcal{D}}(y | x)} \right] \\ &= \frac{1}{|\mathcal{D}|} \sum_{(x, y) \in \mathcal{D}} \left[ \log \frac{1}{f(y | x)} \right] - H_{\Pr_{\mathcal{D}}}(Y | X) \square \end{aligned}$$

**Proposition 6.6.** Consider functions  $f, h : X \rightarrow Y$  from inputs to labels, where  $h$  is a predictor and  $f$  generates the true labels. The inconsistency of believing  $f$  and  $h$  (with any confidences), and a distribution  $D(X)$  with confidence  $\beta$ , is  $\beta$  times the log accuracy of  $h$ . That is,

$$\begin{aligned} \left\langle\!\!\left\langle \begin{array}{c} \xrightarrow[D]{(\beta)} X \\ \xrightarrow[f]{(s)} \end{array} \xrightarrow[h]{(r)} Y \right\rangle\!\!\right\rangle &= -\beta \log \Pr_{x \sim D}(f(x) = h(x)) \\ &= \beta I_D[f = h]. \end{aligned} \tag{6.2}$$

*Proof.* Because  $f$  is deterministic, for every  $x$  in the support of a joint distribution  $\mu$  with finite score, we must have  $\mu(Y | x) = \delta_{f(x)}$ , since if  $\mu$  were to place any non-zero mass  $\mu(x, y) = \epsilon > 0$  on a point  $(x, y)$  with  $y \neq f(x)$  results in an infinite contribution to the KL divergence

$$D(\mu(Y | x) \parallel \delta_{f(x)}) = \mathbb{E}_{x, y \sim \mu} \log \frac{\mu(y | x)}{\delta_{f(x)}} \geq \mu(y, x) \log \frac{\mu(x, y)}{\mu(x) \cdot \delta_{f(x)}(y)} = \epsilon \log \frac{\epsilon}{0} = \infty.$$

The same holds for  $h$ . Therefore, for any  $\mu$  with a finite score, and  $x$  with  $\mu(x) > 0$ , we have  $\delta_{f(x)} = \mu(Y | x) = \delta_{h(x)}$ , meaning that we need only consider  $\mu$  whose support is a subset of those points on which  $f$  and  $h$  agree. On all such points, the contribution to the score from the edges associated to  $f$  and  $h$  will be zero, since  $\mu$  matches the conditional marginals exactly, and the total incompatibility of such a distribution  $\mu$  is equal to the relative entropy  $D(\mu \parallel D)$ , scaled by the confidence  $\beta$  of the empirical distribution  $D$ .

So, among those distributions  $\mu(X)$  supported on an event  $E \subset \mathcal{V}(X)$ , which minimizes is the relative entropy of  $D(\mu \| D)$ ? It is well known that the conditional distribution  $D | E \propto \delta_E(X)D(X) = \frac{1}{D(E)}\delta_E(X)D(X)$  satisfies this property uniquely (see, for instance, [Halpern \(2017\)](#)). Let  $f = h$  denote the event that  $f$  and  $h$  agree. Then we calculate

$$\begin{aligned}
\left\langle \begin{array}{c} \xrightarrow{D} X \\ \xrightarrow[f]{h} Y \end{array} \right\rangle &= \inf_{\substack{\mu(X) \text{ s.t.} \\ \text{Supp } \mu \subseteq [f=h]}} \beta D(\mu(X) \| D(X)) \\
&= \beta D(D | [f=h] \| D) \\
&= \beta \mathbb{E}_{D|f=h} \log \frac{\delta_{f=h}(X)D(X)}{D(f=h) \cdot D(X)} \\
&= \beta \mathbb{E}_{D|f=h} \log \frac{1}{D(f=h)} \quad \left[ \begin{array}{l} \text{since } \delta_{f=h}(x) = 1 \text{ for all } x \text{ that} \\ \text{contribute to the expectation} \end{array} \right] \\
&= -\beta \log D(f = h) \quad \left[ \begin{array}{l} \text{since } D(f = h) \text{ is a constant} \end{array} \right] \\
&= -\beta \log (\text{accuracy}_{f,D}(h)) \\
&= \beta \mathbb{I}_D[f = h]. \quad \square
\end{aligned}$$

**Proposition 6.16.** Consider a PDG containing two (distinct) conditional Gaussian distributions on a variable  $Y$ , whose parameters can both depend on a variable  $X$ . Its inconsistency takes the form

$$\begin{aligned}
&\left\langle \begin{array}{c} \xrightarrow[D]{(\infty)} X \\ \xrightarrow[f]{s} \begin{array}{c} \mu_1 \\ \sigma_1 \end{array} \\ \xrightarrow[t]{h} \begin{array}{c} \sigma_2 \\ \mu_2 \end{array} \end{array} \right\rangle \xrightarrow[N]{(\beta_1)} Y \\
&= \mathbb{E}_D \left[ (\beta_1 + \beta_2) \log \frac{\text{QM}_{\hat{\beta}}(\sigma_1, \sigma_2)}{\text{GM}_{\hat{\beta}}(\sigma_1, \sigma_2)} + \frac{1}{2} \frac{\beta_1 \beta_2}{\beta_1 + \beta_2} \left( \frac{\mu_1 - \mu_2}{\text{QM}_{\hat{\beta}}(\sigma_1, \sigma_2)} \right)^2 \right] \tag{6.6} \\
&= \frac{1}{2} \mathbb{E}_{x \sim D} \left[ \frac{\beta_1 \beta_2}{2} \frac{(f(x) - h(x))^2}{\beta_2 s(x)^2 + \beta_1 t(x)^2} + \frac{\beta_1 + \beta_2}{2} \log \frac{\beta_2 s(x)^2 + \beta_1 t(x)^2}{\beta_1 + \beta_2} - \beta_2 \log s(x) - \beta_1 \log t(x) \right]
\end{aligned}$$

where  $\hat{\beta} = (\frac{\beta_2}{\beta_1 + \beta_2}, \frac{\beta_1}{\beta_1 + \beta_2})$  represents the normalized and reversed vector of confidences  $\beta = (\beta_1, \beta_2)$  for the two distributions, and  $\mu_1 = f(X)$ ,  $\mu_2 = g(X)$ ,  $\sigma_1 = s(X)$ ,  $\sigma_2 = t(X)$  are random variables over  $X$ .

*Proof.* Let  $m$  denote the PDG in question. Since  $D$  has high confidence, we know any joint distribution  $\mu$  with a finite score must have  $\mu(X) = D(X)$ . Thus,

$$\begin{aligned}\langle\!\langle m \rangle\!\rangle &= \inf_{\mu} \mathbb{E}_{x \sim D} \mathbb{E}_{y \sim \mu|x} \left[ \beta_1 \log \frac{\mu(y|x)}{\mathcal{N}(y|f(x), s(x))} + \beta_2 \log \frac{\mu(y|x)}{\mathcal{N}(y|h(x), t(x))} \right] \\ &= \inf_{\mu} \mathbb{E}_{\substack{x \sim D \\ y \sim \mu|x}} \left[ \beta_1 \log \frac{\mu(y|x)}{\frac{1}{s(x)\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{y-f(x)}{s(x)}\right)^2\right)} + \beta_2 \log \frac{\mu(y|x)}{\frac{1}{t(x)\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{y-h(x)}{t(x)}\right)^2\right)} \right] \\ &= \inf_{\mu} \mathbb{E}_{\substack{x \sim D \\ y \sim \mu|x}} \left[ \log \mu(y|x)^{\beta_1+\beta_2} \begin{matrix} +\frac{\beta_1}{2} \left(\frac{y-f(x)}{s(x)}\right)^2 \\ +\frac{\beta_2}{2} \left(\frac{y-h(x)}{t(x)}\right)^2 \end{matrix} \begin{matrix} \\ +\beta_1 \log(s(x)\sqrt{2\pi}) \\ +\beta_2 \log(t(x)\sqrt{2\pi}) \end{matrix} \right]. \quad (6.7)\end{aligned}$$

At this point, we would like make use of the fact that the sum of two parabolas is itself a parabola, so as to combine the two terms on the top right of the previous equation. Concretely, we claim (Claim 2, whose proof is at the end of the present one), that if we define

$$g(x) := \frac{\beta_1 t(x)^2 f(x) + \beta_2 s(x)^2 h(x)}{\beta_1 t(x)^2 + \beta_2 s(x)^2} \quad \text{and} \quad \tilde{\sigma}(x) := \frac{s(x)t(x)}{\sqrt{\beta_1 t(x)^2 + \beta_2 s(x)^2}},$$

then

$$\frac{\beta_1}{s(x)^2} (y - f)^2 + \frac{\beta_2}{t(x)^2} (y - h)^2 = \left( \frac{y - g}{\tilde{\sigma}} \right)^2 + \frac{\beta_1 \beta_2}{\beta_1 t(x)^2 + \beta_2 s(x)^2} (f - h)^2.$$

Applying this to (6.7) leaves us with:

$$\langle\!\langle m \rangle\!\rangle = \inf_{\mu} \mathbb{E}_{x \sim D} \mathbb{E}_{y \sim \mu|x} \left[ \begin{matrix} \log \mu(y|x)^{\beta_1+\beta_2} + \frac{1}{2\tilde{\sigma}(x)^2} (y - g(x))^2 & +\beta_1 \log(s(x)\sqrt{2\pi}) \\ +\frac{1}{2} \frac{\beta_1 \beta_2}{\beta_1 t(x)^2 + \beta_2 s(x)^2} (f(x) - h(x))^2 & +\beta_2 \log(t(x)\sqrt{2\pi}) \end{matrix} \right]$$

Pulling the term on the bottom left (which does not depend on  $y$ ), out of the expectation, and folding the rest of the terms back inside the logarithm (which in particular means first replacing the top middle term  $\varphi$  by  $-\log(\exp(-\varphi))$ ), we obtain  $\langle\!\langle m \rangle\!\rangle =$

$$\mathbb{E}_{x \sim D} \left[ \inf_{\mu(Y)} \mathbb{E}_{y \sim \mu} \left[ \log \mu(y)^{\beta_1 + \beta_2} - \log \left( \frac{1}{\sqrt{2\pi^{\beta_1 + \beta_2}} s(x)^{\beta_1} t(x)^{\beta_2}} \exp \left\{ -\frac{1}{2} \left( \frac{y - g(x)}{\tilde{\sigma}(x)} \right)^2 \right\} \right) \right] + \frac{1}{2} \frac{\beta_1 \beta_2}{\beta_1 t(x)^2 + \beta_2 s(x)^2} (f(x) - h(x))^2 \right].$$

To simplify the presentation, let  $\psi$  be the term on the top right, and  $\xi$  be the term on the bottom. More explicitly, define

$$\psi(x, y) := \frac{1}{2} \frac{1}{\sqrt{2\pi^{\beta_1 + \beta_2}} s(x)^{\beta_1} t(x)^{\beta_2}} \exp \left\{ -\frac{1}{2} \left( \frac{y - g(x)}{\tilde{\sigma}(x)} \right)^2 \right\},$$

and  $\xi(x) := \frac{\beta_1 \beta_2}{\beta_1 t(x)^2 + \beta_2 s(x)^2} (f(x) - h(x))^2,$

which lets us write the previous expression for  $\langle\!\langle m \rangle\!\rangle$  as

$$\langle\!\langle m \rangle\!\rangle = \mathbb{E}_{x \sim D} \left[ \inf_{\mu(Y)} \mathbb{E}_{y \sim \mu} [\log \mu(y)^{\beta_1 + \beta_2} - \log \psi(x, y)] + \xi(x) \right]. \quad (6.8)$$

Also, let  $\hat{\beta}_1 := \frac{\beta_1}{\beta_1 + \beta_2}$ , and  $\hat{\beta}_2 := \frac{\beta_2}{\beta_1 + \beta_2}$ . For reasons that will soon become clear, we are actually interested in  $\psi^{\frac{1}{\beta_1 + \beta_2}}$ , which we compute as

$$\begin{aligned} \psi(x, y)^{\frac{1}{\beta_1 + \beta_2}} &= (2\pi)^{-\frac{1}{2}} s(x)^{\left(\frac{-\beta_1}{\beta_1 + \beta_2}\right)} t(x)^{\left(\frac{-\beta_2}{\beta_1 + \beta_2}\right)} \exp \left\{ -\frac{1}{2} \left( \frac{y - g(x)}{\tilde{\sigma}(x)} \right)^2 \right\}^{\frac{1}{\beta_1 + \beta_2}} \\ &= \frac{1}{\sqrt{2\pi} s(x)^{\hat{\beta}_1} t(x)^{\hat{\beta}_2}} \exp \left\{ \frac{-1}{2(\beta_1 + \beta_2)} \left( \frac{y - g(x)}{\tilde{\sigma}(x)} \right)^2 \right\}. \end{aligned}$$

Recall that the Gaussian density  $\mathcal{N}(y | g(x), \tilde{\sigma}(x)\sqrt{\beta_1 + \beta_2})$  of mean  $g(x)$  and variance  $\tilde{\sigma}(x)^2(\beta_1 + \beta_2)$  is given by

$$\mathcal{N}(y | g(x), \tilde{\sigma}(x)\sqrt{\beta_1 + \beta_2}) = \frac{1}{\sqrt{2\pi} \tilde{\sigma}(x)\sqrt{\beta_1 + \beta_2}} \exp \left\{ \frac{-1}{2(\beta_1 + \beta_2)} \left( \frac{y - g(x)}{\tilde{\sigma}(x)} \right)^2 \right\},$$

which is quite similar, and has an identical dependence on  $y$ . To facilitate converting one to the other, we explicitly compute the ratio:

$$\frac{\psi(x, y)^{\frac{1}{\beta_1 + \beta_2}}}{\mathcal{N}(y | g(x), \tilde{\sigma}(x)\sqrt{\beta_1 + \beta_2})}$$

$$\begin{aligned}
&= \frac{\tilde{\sigma} \sqrt{2\pi(\beta_1 + \beta_2)}}{\sqrt{2\pi} s(x)^{\hat{\beta}_1} t(x)^{\hat{\beta}_2}} \\
&= \frac{\tilde{\sigma} \sqrt{\beta_1 + \beta_2}}{s(x)^{\hat{\beta}_1} t(x)^{\hat{\beta}_2}} \\
&= \left( \frac{s(x) t(x)}{\sqrt{\beta_1 t(x)^2 + \beta_2 s(x)^2}} \right) \frac{\sqrt{\beta_1 + \beta_2}}{s(x)^{\hat{\beta}_1} t(x)^{\hat{\beta}_2}} && [\text{expand defn of } \tilde{\sigma}(x)] \\
&= s(x)^{1-\hat{\beta}_1} t(x)^{1-\hat{\beta}_2} \sqrt{\frac{\beta_1 + \beta_2}{\beta_1 t(x)^2 + \beta_2 s(x)^2}} \\
&= s(x)^{1-\hat{\beta}_1} t(x)^{1-\hat{\beta}_2} \sqrt{\frac{1}{\hat{\beta}_1 t(x)^2 + \hat{\beta}_2 s(x)^2}} && [\text{defn of } \hat{\beta}_1, \hat{\beta}_2] \\
&= \frac{s(x)^{\hat{\beta}_2} t(x)^{\hat{\beta}_1}}{\sqrt{\hat{\beta}_1 t(x)^2 + \hat{\beta}_2 s(x)^2}} && [\text{since } \hat{\beta}_1 + \hat{\beta}_2 = 1]
\end{aligned}$$

Now, picking up from where we left off in (6.8), we have

$$\begin{aligned}
\langle\langle m \rangle\rangle &= \mathbb{E}_{x \sim D} \left[ \inf_{\mu(Y)} \mathbb{E}_{y \sim \mu} [\log \mu(y)^{\beta_1 + \beta_2} - \log \psi(x, y)] + \xi(x) \right] \\
&= \mathbb{E}_{x \sim D} \left[ \inf_{\mu(Y)} \mathbb{E}_{y \sim \mu} \left[ \log \frac{\mu(y)^{\beta_1 + \beta_2}}{\psi(x, y)^{\frac{\beta_1 + \beta_2}{\beta_1 + \beta_2}}} \right] + \xi(x) \right] \\
&= \mathbb{E}_{x \sim D} \left[ \inf_{\mu(Y)} \mathbb{E}_{y \sim \mu} \left[ (\beta_1 + \beta_2) \log \frac{\mu(y)}{\psi(x, y)^{\frac{1}{\beta_1 + \beta_2}}} \right] + \xi(x) \right] \\
&= \mathbb{E}_{x \sim D} \left[ \inf_{\mu(Y)} \mathbb{E}_{y \sim \mu} \left[ (\beta_1 + \beta_2) \log \frac{\mu(y)}{\mathcal{N}(y | g(x), \tilde{\sigma}(x) \sqrt{\beta_1 + \beta_2}) \frac{s(x)^{\hat{\beta}_2} t(x)^{\hat{\beta}_1}}{\sqrt{\hat{\beta}_1 t(x)^2 + \hat{\beta}_2 s(x)^2}}} \right] + \xi(x) \right] \\
&= \mathbb{E}_{x \sim D} \left[ \inf_{\mu(Y)} \mathbb{E}_{y \sim \mu} \left[ (\beta_1 + \beta_2) \log \frac{\mu(y)}{\mathcal{N}(y | g(x), \tilde{\sigma}(x) \sqrt{\beta_1 + \beta_2})} \right. \right. \\
&\quad \left. \left. + (\beta_1 + \beta_2) \log \frac{\sqrt{\hat{\beta}_1 t(x)^2 + \hat{\beta}_2 s(x)^2}}{s(x)^{\hat{\beta}_2} t(x)^{\hat{\beta}_1}} + \xi(x) \right] \right]
\end{aligned}$$

but now the entire first term is the infimum of a relative entropy, which is non-negative and equal to zero iff  $\mu(y) = \mathcal{N}(y | g(x), \tilde{\sigma}(x) \sqrt{\beta_1 + \beta_2})$ . So the infimum on the left is equal to zero.

$$\langle\langle m \rangle\rangle = \mathbb{E}_{x \sim D} \left[ (\beta_1 + \beta_2) \log \frac{\sqrt{\hat{\beta}_1 t(x)^2 + \hat{\beta}_2 s(x)^2}}{s(x)^{\hat{\beta}_2} t(x)^{\hat{\beta}_1}} + \xi(x) \right] \tag{6.9}$$

$$\begin{aligned}
&= \mathbb{E}_{x \sim D} \left[ (\beta_1 + \beta_2) \log \sqrt{\hat{\beta}_1 t(x)^2 + \hat{\beta}_2 s(x)^2} - (\beta_1 + \beta_2) \log \left( s(x)^{\hat{\beta}_2} t(x)^{\hat{\beta}_1} \right) + \xi(x) \right] \\
&= \mathbb{E}_{x \sim D} \left[ (\beta_1 + \beta_2) \log \sqrt{\hat{\beta}_1 t(x)^2 + \hat{\beta}_2 s(x)^2} \begin{matrix} -\beta_2 \log s(x) \\ -\beta_1 \log t(x) \end{matrix} + \xi(x) \right] \\
&= \mathbb{E}_{x \sim D} \left[ (\beta_1 + \beta_2) \log \sqrt{\frac{\beta_1 t(x)^2 + \beta_2 s(x)^2}{\beta_1 + \beta_2}} \begin{matrix} -\beta_2 \log s(x) \\ -\beta_1 \log t(x) \end{matrix} \right. \\
&\quad \left. + \frac{1}{2} \frac{\beta_1 \beta_2}{\beta_1 t(x)^2 + \beta_2 s(x)^2} (f(x) - h(x))^2 \right] \tag{6.10}
\end{aligned}$$

Whew! Pulling the square root of the logarithm proves complex second half of the proposition. Now, we massage it into into a (slightly) more readable form.

To start, write  $\sigma_1$  (the random variable) in place of  $s(x)$  and  $\sigma_2$  in place of  $t(x)$ . Let  $\hat{\beta}$  without the subscript denote the vector  $(\hat{\beta}_2, \hat{\beta}_1) = (\frac{\beta_2}{\beta_1 + \beta_2}, \frac{\beta_1}{\beta_1 + \beta_2})$ , which we will use for weighted means. The  $\hat{\beta}$ -weighted arithmetic, geometric ( $p = 0$ ), and quadratic ( $p = 2$ ) means of  $\sigma_1$  and  $\sigma_2$  are:

$$\text{GM}_{\hat{\beta}}(\sigma_1, \sigma_2) = (\sigma_1)^{\hat{\beta}_2} (\sigma_2)^{\hat{\beta}_1} \quad \text{and} \quad \text{QM}_{\hat{\beta}}(\sigma_1, \sigma_2) = \sqrt{\hat{\beta}_2 \sigma_1^2 + \hat{\beta}_1 \sigma_2^2}.$$

So, now we can write  $\xi(x)$  as

$$\begin{aligned}
\frac{1}{2} \frac{\beta_1 \beta_2}{\beta_1 t(x)^2 + \beta_2 s(x)^2} (f(x) - h(x))^2 &= \frac{1}{2} \frac{\beta_1 \beta_2}{\beta_1 + \beta_2} \frac{\beta_1 + \beta_2}{\beta_1 t(x)^2 + \beta_2 s(x)^2} (f(x) - h(x))^2 \\
&= \frac{1}{2} \frac{\beta_1 \beta_2}{\beta_1 + \beta_2} \left( \frac{1}{\text{QM}_{\hat{\beta}}(\sigma_1, \sigma_2)} \right)^2 (f(x) - h(x))^2 \\
&= \frac{1}{2} \frac{\beta_1 \beta_2}{\beta_1 + \beta_2} \left( \frac{\mu_1 - \mu_2}{\text{QM}_{\hat{\beta}}(\sigma_1, \sigma_2)} \right)^2;
\end{aligned}$$

in the last step, we have replaced  $f(x)$  and  $g(x)$  with their respective random variables  $\mu_1$  and  $\mu_2$ . As a result, (6.9) can be written as

$$\langle\langle m \rangle\rangle = \mathbb{E}_D \left[ (\beta_1 + \beta_2) \log \frac{\text{QM}_{\hat{\beta}}(\sigma_1, \sigma_2)}{\text{GM}_{\hat{\beta}}(\sigma_1, \sigma_2)} + \frac{1}{2} \frac{\beta_1 \beta_2}{\beta_1 + \beta_2} \left( \frac{\mu_1 - \mu_2}{\text{QM}_{\hat{\beta}}(\sigma_1, \sigma_2)} \right)^2 \right]$$

... which is perhaps more comprehensible, and proves the first half of our proposition.  $\square$

**Claim 2.** *The sum of two functions that are unshifted parabolas as functions of  $y$  (i.e., both functions are of the form  $k(y - a)^2$ ), is itself a (possibly shifted) parabola of  $y$  (and of the form  $k'(y - a') + b'$ ). More concretely, and adapted to our usage above, the following algebraic relation holds:*

$$\frac{\beta_1}{\sigma_1^2}(y - f)^2 + \frac{\beta_2}{\sigma_2^2}(y - h)^2 = \left(\frac{y - g}{\tilde{\sigma}}\right)^2 + \frac{\beta_1\beta_2}{\beta_1\sigma_2^2 + \beta_2\sigma_1^2}(f - h)^2,$$

where

$$g := \frac{\beta_1\sigma_2^2 f + \beta_2\sigma_1^2 h}{\beta_1\sigma_2^2 + \beta_2\sigma_1^2} \quad \text{and} \quad \tilde{\sigma} := \left(\frac{\beta_1}{\sigma_1^2} + \frac{\beta_2}{\sigma_2^2}\right)^{-1/2} = \frac{\sigma_1\sigma_2}{\sqrt{\beta_1\sigma_2^2 + \beta_2\sigma_1^2}}.$$

*Proof.* Expand terms and complete the square. Starting from the left hand side, we have

$$\begin{aligned} & \frac{\beta_1}{\sigma_1^2}(y - f)^2 + \frac{\beta_2}{\sigma_2^2}(y - h)^2 \\ &= \frac{\beta_1}{\sigma_1^2}(y^2 - 2yf + f^2) + \frac{\beta_2}{\sigma_2^2}(y^2 - 2yh + h^2) \\ &= \left(\frac{\beta_1}{\sigma_1^2} + \frac{\beta_2}{\sigma_2^2}\right)y^2 - 2\left(\frac{\beta_1 f}{\sigma_1^2} + \frac{\beta_2 h}{\sigma_2^2}\right)y + \left(\frac{\beta_1 f^2}{\sigma_1^2} + \frac{\beta_2 h^2}{\sigma_2^2}\right) \\ &= \left(\frac{\beta_1}{\sigma_1^2} + \frac{\beta_2}{\sigma_2^2}\right)y^2 - 2\left(\frac{\beta_1 f}{\sigma_1^2} + \frac{\beta_2 h}{\sigma_2^2}\right)y + \left(\frac{\beta_1 f^2}{\sigma_1^2} + \frac{\beta_2 h^2}{\sigma_2^2} - \frac{\beta_1\beta_2(f - h)^2}{\beta_1\sigma_2^2 + \beta_2\sigma_1^2}\right) \\ &\quad + \frac{\beta_1\beta_2(f - h)^2}{\beta_1\sigma_2^2 + \beta_2\sigma_1^2}. \end{aligned} \tag{6.11}$$

In the last step, we added and removed the same term (equal to the “completion of the square”, although at this point it may be unclear that this is the right quantity). The third parenthesized quantity needs the most work. Isolating it and getting a common denominator gives us:

$$\frac{\beta_1 f^2}{\sigma_1^2} + \frac{\beta_2 h^2}{\sigma_2^2} - \frac{\beta_1\beta_2(f - h)^2}{\beta_1\sigma_2^2 + \beta_2\sigma_1^2}$$

$$\begin{aligned}
&= \frac{\beta_1 f^2 (\beta_1 \sigma_2^2 + \beta_2 \sigma_1^2) \sigma_2^2}{\sigma_1^2 (\beta_1 \sigma_2^2 + \beta_2 \sigma_1^2) \sigma_2^2} + \frac{\beta_2 h^2 (\beta_1 \sigma_2^2 + \beta_2 \sigma_1^2) \sigma_1^2}{\sigma_2^2 (\beta_1 \sigma_2^2 + \beta_2 \sigma_1^2) \sigma_1^2} - \frac{\beta_1 \beta_2 (f^2 - 2fh + h^2) (\sigma_1^2 \sigma_2^2)}{(\beta_1 \sigma_2^2 + \beta_2 \sigma_1^2) (\sigma_1^2 \sigma_2^2)} \\
&= \frac{\beta_1^2 \sigma_2^4 f^2 + \cancel{\beta_1 \beta_2 \sigma_2^2 \sigma_1^2 f^2} + \cancel{\beta_1 \beta_2 \sigma_1^2 \sigma_2^2 h^2} + \beta_2^2 \sigma_1^4 h^2 - \cancel{\beta_1 \beta_2 \sigma_1^2 \sigma_2^2 f^2} + 2\beta_1 \beta_2 \sigma_1^2 \sigma_2^2 fh}{(\beta_1 \sigma_2^2 + \beta_2 \sigma_1^2) (\sigma_1^2 \sigma_2^2)} \\
&\quad - \frac{\cancel{\beta_1 \beta_2 \sigma_1^2 \sigma_2^2 h^2}}{(\beta_1 \sigma_2^2 + \beta_2 \sigma_1^2) (\sigma_1^2 \sigma_2^2)} \\
&= \frac{\beta_1^2 \sigma_2^4 f^2 + \beta_2^2 \sigma_1^4 h^2 + 2\beta_1 \beta_2 \sigma_1^2 \sigma_2^2 fh}{(\beta_1 \sigma_2^2 + \beta_2 \sigma_1^2) (\sigma_1^2 \sigma_2^2)}.
\end{aligned}$$

Substituting this expression into the third term of (6.11), while simultaneously computing common denominators for the first and second terms, yields the expression

$$\begin{aligned}
&\left( \frac{\beta_1 \sigma_2^2 + \beta_2 \sigma_1^2}{\sigma_1^2 \sigma_2^2} \right) y^2 - 2 \left( \frac{\beta_1 \sigma_2^2 f + \beta_2 \sigma_1^2 h}{\sigma_1^2 \sigma_2^2} \right) y + \frac{\beta_1^2 \sigma_2^4 f^2 + \beta_2^2 \sigma_1^4 h^2 + 2\beta_1 \beta_2 \sigma_1^2 \sigma_2^2 fh}{(\beta_1 \sigma_2^2 + \beta_2 \sigma_1^2) (\sigma_1^2 \sigma_2^2)} \\
&\quad + \frac{\beta_1 \beta_2 (f - h)^2}{\beta_1 \sigma_2^2 + \beta_2 \sigma_1^2} \tag{6.12}
\end{aligned}$$

for the left hand side On the other hand, using the definitions of  $g$  and  $\tilde{\sigma}$ , we compute:

$$\begin{aligned}
&\left( \frac{y - g}{\tilde{\sigma}} \right)^2 \\
&= \left( \frac{\beta_1 \sigma_2^2 + \beta_2 \sigma_1^2}{\sigma_1^2 \sigma_2^2} \right) \left( y - \frac{\beta_1 \sigma_2^2 f + \beta_2 \sigma_1^2 h}{\beta_1 \sigma_2^2 + \beta_2 \sigma_1^2} \right)^2 \\
&= \left( \frac{\beta_1 \sigma_2^2 + \beta_2 \sigma_1^2}{\sigma_1^2 \sigma_2^2} \right) \left( y^2 - 2y \frac{\beta_1 \sigma_2^2 f + \beta_2 \sigma_1^2 h}{\beta_1 \sigma_2^2 + \beta_2 \sigma_1^2} + \frac{\beta_1^2 \sigma_2^4 f^2 + \beta_2^2 \sigma_1^4 h^2 + 2\beta_1 \beta_2 \sigma_1^2 \sigma_2^2 fh}{(\beta_1 \sigma_2^2 + \beta_2 \sigma_1^2)^2} \right) \\
&= \left( \frac{\beta_1 \sigma_2^2 + \beta_2 \sigma_1^2}{\sigma_1^2 \sigma_2^2} \right) y^2 - 2 \left( \frac{\beta_1 \sigma_2^2 f + \beta_2 \sigma_1^2 h}{\sigma_1^2 \sigma_2^2} \right) y + \frac{\beta_1^2 \sigma_2^4 f^2 + \beta_2^2 \sigma_1^4 h^2 + 2\beta_1 \beta_2 \sigma_1^2 \sigma_2^2 fh}{(\beta_1 \sigma_2^2 + \beta_2 \sigma_1^2) (\sigma_1^2 \sigma_2^2)}
\end{aligned}$$

... which is precisely the first 3 terms of (6.12). Putting it all together, we have shown that

$$\frac{\beta_1}{\sigma_1^2} (y - f)^2 + \frac{\beta_2}{\sigma_2^2} (y - h)^2 = \left( \frac{y - g}{\tilde{\sigma}} \right)^2 + \frac{\beta_1 \beta_2 (f - h)^2}{\beta_1 \sigma_2^2 + \beta_2 \sigma_1^2} \quad \text{as desired.} \quad \square$$

### Proposition 6.7.

$$\left\langle \left. \begin{array}{c} D \\ \xrightarrow{(\infty)} \end{array} \right| X \xrightarrow{f} \mu_f \xrightarrow{\mathcal{N}_1} Y \xrightarrow{h} \mu_h \xrightarrow{\mathcal{N}_1} \right\rangle = \frac{1}{2} \mathbb{E}_D |f(X) - h(X)|^2 =: \text{MSE}_D(f, h),$$

where  $\mathcal{N}_1(Y | \mu)$  is a unit Gaussian on  $Y$  with mean  $\mu$ .

*Proof.* An immediate corollary of Proposition 6.16; simply set  $s(x) = t(x) = \beta_1 = \beta_2 = 1$   $\square$

**Lemma 6.10.** *The inconsistency  $D_{(r,s)}^{\text{PDG}}(p\|q)$  of a PDG comprising  $p(X)$  with confidence  $r$  and  $q(X)$  with confidence  $s$  is given in closed form by*

$$D_{(r,s)}^{\text{PDG}}(p\|q) := \left\langle \left. \frac{p}{(r)} \rightarrow X \leftarrow \frac{q}{(s)} \right. \right\rangle = -(r+s) \log \sum_x \left( p(x)^r q(x)^s \right)^{\frac{1}{r+s}}.$$

*Proof.*

$$\begin{aligned} \left\langle \left. \frac{p}{(\beta:r)} \rightarrow X \leftarrow \frac{q}{(\beta:s)} \right. \right\rangle &= \inf_{\mu} \mathbb{E}_{\mu} \log \frac{\mu(x)^{r+s}}{p(x)^r q(x)^s} \\ &= (r+s) \inf_{\mu} \mathbb{E}_{\mu} \left[ \log \frac{\mu(x)}{(p(x)^r q(x)^s)^{\frac{1}{r+s}}} \cdot \frac{Z}{Z} \right] \\ &= \inf_{\mu} (r+s) D\left(\mu \middle\| \frac{1}{Z} p^{\frac{r}{r+s}} q^{\frac{s}{r+s}}\right) - (r+s) \log Z \end{aligned}$$

where  $Z := (\sum_x p(x)^r q(x)^s)^{\frac{1}{r+s}}$  is the constant required to normalize the denominator as a distribution. The first term is now a relative entropy, and the only usage of  $\mu$ .  $D(\mu \| \dots)$  achieves its minimum of zero when  $\mu$  is the second distribution, so our formula becomes

$$\begin{aligned} &= -(r+s) \log Z \\ &= -(r+s) \log \sum_x \left( p(x)^r q(x)^s \right)^{\frac{1}{r+s}} \quad \text{as promised.} \quad \square \end{aligned}$$

**Proposition 6.8.** *Suppose you have a parameterized model  $p(Y|\Theta)$ , a prior  $q(\Theta)$ , and a trusted distribution  $D(Y)$ . The inconsistency of also believing  $\Theta = \theta$  is the cross*

entropy loss, plus the regularizer:  $\log \frac{1}{q(\theta)}$  times your confidence in  $q$ . That is,

$$\left\langle\left\langle \begin{array}{c} q \\ \xrightarrow{(\beta)} \Theta \xrightarrow{p} Y \\ \xrightarrow{\theta} \end{array} \right\rangle\right\rangle = \mathbb{E}_{y \sim D} \log \frac{1}{p(y|\theta)} + \beta \log \frac{1}{q(\theta)} - H(D) \quad (6.3)$$

*Proof.* This is another case where there's only one joint distribution  $\mu(\Theta, Y)$  that gets a finite score. We must have  $\mu(Y) = D(Y)$  since  $D$  has infinite confidence, which uniquely extends to the distribution  $\mu(\Theta, Y) = D(Y)\delta_\theta(\Theta)$  for which deterministically sets  $\Theta = \theta$ .

The cpds corresponding to the edges labeled  $\theta$  and  $D$ , then, are satisfied by this  $\mu$  and contribute nothing to the score. So the two relevant edges that contribute incompatibility with this distribution are  $p$  and  $q$ . Letting  $\mathcal{M}$  denote the PDG in question, we compute:

$$\begin{aligned} \langle\langle \mathcal{M} \rangle\rangle &= \mathbb{E}_\mu \left[ \log \frac{\mu(Y|\Theta)}{p(Y|\Theta)} + \beta \log \frac{\mu(\Theta)}{q(\Theta)} \right] \\ &= \mathbb{E}_{y \sim D} \left[ \log \frac{D(y)}{p(y|\theta)} + \beta \log \frac{1}{q(\theta)} \right] \\ &= \mathbb{E}_{y \sim D} \left[ \log \frac{1}{p(y|\theta)} + \beta \log \frac{1}{q(\theta)} + \log D(y) \right] \\ &= \mathbb{E}_{y \sim D} \left[ \log \frac{1}{p(y|\theta)} \right] + \beta \log \frac{1}{q(\theta)} - H(D) \quad \text{as desired.} \quad \square \end{aligned}$$

**Proposition 6.11.** *The negative ELBO of  $x$  is the inconsistency of the PDG containing  $p, q$ , and  $X=x$ , with high confidence in  $q$ . That is,*

$$-\text{ELBO}(x) = \left\langle\left\langle \begin{array}{c} p \\ \xrightarrow{(\infty)} Z \xrightarrow{q} X \xleftarrow{x} \end{array} \right\rangle\right\rangle.$$

*Proof.* Every distribution that does marginalize to  $q(Z)$  or places any mass on  $x' \neq x$  will have infinite score. Thus the only distribution that could have a finite

score is  $\mu(X, Z)$ . Thus,

$$\begin{aligned}
\left\langle \xrightarrow[\infty]{q} Z \xleftarrow[p]{X} \xleftarrow{x} \right\rangle &= \inf_{\mu} \left[ \xrightarrow[\infty]{q} Z \xleftarrow[p]{X} \xleftarrow{x} \right] (\mu) \\
&= \left[ \xrightarrow[\infty]{q} Z \xleftarrow[p]{X} \xleftarrow{x} \right] (\delta_x(X) q(Z)) \\
&= \mathbb{E}_{\substack{x' \sim \delta_x \\ z \sim q}} \log \frac{\delta_x(x') q(z)}{p(x', z)} = - \mathbb{E}_{z \sim q} \frac{p(x, z)}{q(z)} = -\text{ELBO}_{p,q}(x). \square
\end{aligned}$$

We prove both [Proposition 6.12](#) and [Proposition 6.17](#) at the same time.

**Proposition 6.12.** *The VAE loss of a sample  $x$  is the inconsistency of the PDG comprising the encoder  $e$  (with high confidence, as it defines the encoding), decoder  $d$ , prior  $p$ , and  $x$ . That is,*

$$-\text{ELBO}_{p,e,d}(x) = \left\langle \xrightarrow[p]{Z} \xleftarrow[e]{X} \xleftarrow{x} \right\rangle.$$

**Proposition 6.17.** *The following analog of [Proposition 6.12](#) for a whole dataset  $\mathcal{D}$  holds:*

$$-\mathbb{E}_{\Pr_{\mathcal{D}}} \text{ELBO}_{p,e,d}(X) = \left\langle \xrightarrow[p]{Z} \xleftarrow[e]{X} \xleftarrow[\Pr_{\mathcal{D}}]{X} \right\rangle + H(\Pr_{\mathcal{D}}).$$

*Proof.* The two proofs are similar. For [Proposition 6.12](#), the optimal distribution must be  $\delta_x(X)e(Z \mid X)$ , and for [Proposition 6.17](#), it must be  $\Pr_{\mathcal{D}}(X)e(Z \mid X)$ , because  $e$  and the data both have infinite confidence, so any other distribution gets an infinite score. At the same time,  $d$  and  $p$  define a joint distribution, so the inconsistency in question becomes

$$D\left(\delta_x(X)e(Z \mid X) \parallel p(Z)d(X \mid Z)\right) = \mathbb{E}_{z \sim e|x} \left[ \log \frac{p(z)d(x \mid z)}{e(z \mid x)} \right] = \text{ELBO}_{p,e,d}(x)$$

in the first, case, and

$$\begin{aligned} D\left(\Pr_{\mathcal{D}}(X)e(Z \mid X) \parallel p(Z)d(X \mid Z)\right) &= \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \mathbb{E}_{z \sim e|x} \left[ \log \frac{p(z)d(x \mid z)}{e(z \mid x)} + \log \frac{1}{\Pr_{\mathcal{D}}(x)} \right] \\ &= \text{ELBO}_{p,e,d}(x) - \text{H}(\Pr_{\mathcal{D}}) \end{aligned}$$

in the second.  $\square$

Now, we formally state and prove the more general result for  $\beta$ -VAEs.

**Proposition 6.18.** *The negative  $\beta$ -ELBO objective for a prior  $p(X)$ , encoder  $e(Z \mid X)$ , decoder  $d(X \mid Z)$ , at a sample  $x$ , is equal to the inconsistency of the corresponding PDG, where the prior has confidence equal to  $\beta$ . That is,*

$$-\beta\text{-ELBO}_{p,e,d}(x) = \left\langle \left. \begin{array}{c} p \xrightarrow{(\beta)} Z \xrightarrow[d]{\curvearrowright} X \xleftarrow{x} \\ e \xleftarrow{(\infty)} \end{array} \right| \right\rangle$$

*Proof.*

$$\begin{aligned} \left\langle \left. \begin{array}{c} p \xrightarrow{(\beta)} Z \xrightarrow[d]{\curvearrowright} X \xleftarrow{x} \\ e \xleftarrow{(\infty)} \end{array} \right| \right\rangle &= \inf_{\mu} \left[ \left. \begin{array}{c} p \xrightarrow{(\beta)} Z \xrightarrow[d]{\curvearrowright} X \xleftarrow{x} \\ e \xleftarrow{(\infty)} \end{array} \right| (\mu) \right] \\ &= \inf_{\mu} \mathbb{E}_{\mu(X,Z)} \left[ \beta \log \frac{\mu(Z)}{p(Z)} + \log \frac{\mu(X, Z)}{\mu(Z)d(X \mid Z)} \right] \end{aligned}$$

As before, the only candidate for a joint distribution with finite score is  $\delta_x(X)e(Z \mid X)$ . Note that the marginal on  $Z$  for this distribution is itself, since  $\int_x \delta_x(X)e(Z \mid X) dx = e(Z \mid x)$ . Thus, our equation becomes

$$\begin{aligned} &= \mathbb{E}_{\delta_x(X)e(Z \mid X)} \left[ \beta \log \frac{e(Z \mid x)}{p(z)} + \log \frac{\delta_x(X)e(Z \mid X)}{e(Z \mid x)d(x \mid Z)} \right] \\ &= \mathbb{E}_{e(Z \mid x)} \left[ \beta \log \frac{e(Z \mid x)}{p(Z)} + \log \frac{1}{d(x \mid Z)} \right] \\ &= D(e(Z \mid x) \parallel p) + \text{Rec}_{e,d}(x) \end{aligned}$$

$$= -\beta\text{-ELBO}_{p,e,d}(x).$$

□

**Proposition 6.14.** *For all weighted factor graphs  $\Psi$ , we have that  $\langle\!\langle \mathbf{m}_\Psi \rangle\!\rangle_1 = -\log Z_\Psi$ .*

*Proof.* In the main text, we defined  $\mathbf{m}_\Psi$  to be the PDG with edges  $\{\overset{J}{\rightarrow} \mathbf{X}_J\}_{\mathcal{J}}$ , cpds  $p_J(\mathbf{X}_J) \propto \phi_J(\mathbf{X}_J)$ , and weights  $\alpha_J, \beta_J := \theta_J$ . Let *the* be a function that extracts the unique element singleton set, so that  $\text{the}(\{x\}) = x$ . It was shown by [Richardson and Halpern \(2021\)](#) (Corollary 4.4.1) that

$$\text{the}[\mathbf{m}_\Psi]_1^* = \Pr_{\Phi,\theta}(\mathbf{x}) = \frac{1}{Z_\Psi} \prod_J \phi_J(\mathbf{x}_J)^{\theta_J}.$$

Recall the statement of Prop 4.6 from [Richardson and Halpern \(2021\)](#):

$$[\mathbf{m}]_\gamma(\mu) = \mathbb{E}_{\mathbf{w} \sim \mu} \left\{ \sum_{X \xrightarrow{a} Y} \left[ \beta_a \log \frac{1}{\mathbb{P}_a(y^{\mathbf{w}} | x^{\mathbf{w}})} + (\gamma \alpha_a - \beta_a) \log \frac{1}{\mu(y^{\mathbf{w}} | x^{\mathbf{w}})} \right] - \gamma \log \frac{1}{\mu(\mathbf{w})} \right\}, \quad (6.13)$$

where  $x^{\mathbf{w}}$  and  $y^{\mathbf{w}}$  are the respective values of the variables  $X$  and  $Y$  in the world  $\mathbf{w}$ . Note that if  $\gamma = 1$ , and  $\alpha, \beta$  are both equal to  $\theta$  in  $\mathbf{m}_\Psi$ , the **middle term (in purple)** is zero. So in our case, since the edges are  $\{\overset{J}{\rightarrow} \mathbf{X}_J\}$  and  $\mathbb{P}_J(\mathbf{X}_J) = \phi_J(\mathbf{X}_J)$ , (6.13) reduces to the standard variational free energy

$$\begin{aligned} VFE_\Psi(\mu) &= \mathbb{E}_\mu \left[ \sum_{J \in \mathcal{J}} \theta_J \log \frac{1}{\phi_J(\mathbf{X}_J)} \right] - H(\mu) \\ &= \mathbb{E}_\mu \langle \boldsymbol{\varphi}, \boldsymbol{\theta} \rangle_{\mathcal{J}} - H(\mu), \quad \text{where } \varphi_J(\mathbf{X}_J) := \log \frac{1}{\phi_J(\mathbf{X}_J)}. \end{aligned} \quad (6.14)$$

By construction,  $\Pr_\Psi$  uniquely minimizes  $VFE$ . The 1-inconsistency,  $\langle\!\langle \mathbf{m}_\Psi \rangle\!\rangle$  is the minimum value attained. We calculate:

$$\langle\!\langle \mathbf{m} \rangle\!\rangle_1 = VFE_\Psi(\Pr_\Psi)$$

$$\begin{aligned}
&= \mathbb{E}_{\mathbf{x} \sim \mu} \left\{ \sum_{J \in \mathcal{J}} \left[ \theta_J \log \frac{1}{\phi_J(\mathbf{x}_J)} \right] - \log \frac{1}{\Pr_{\Phi, \theta}(\mathbf{x})} \right\} && \left[ \text{by (6.14)} \right] \\
&= \mathbb{E}_{\mathbf{x} \sim \mu} \left\{ \sum_{J \in \mathcal{J}} \left[ \theta_J \log \frac{1}{\phi_J(\mathbf{x}_J)} \right] - \log \frac{Z_\Psi}{\prod_{J \in \mathcal{J}} \phi_J(\mathbf{x}_J)^{\theta_j}} \right\} && \left[ \text{definition of } \Pr_\Psi \right] \\
&= \mathbb{E}_{\mathbf{x} \sim \mu} \left\{ \sum_J \left[ \theta_J \log \frac{1}{\phi_J(\mathbf{x}_J)} \right] - \sum_{J \in \mathcal{J}} \left[ \theta_J \log \frac{1}{\phi_J(\mathbf{x}_J)} \right] - \log Z_\Psi \right\} \\
&= \mathbb{E}_{\mathbf{x} \sim \mu} [-\log Z_\Psi] \\
&= -\log Z_\Psi && \left[ Z_\Psi \text{ is constant in } \mathbf{x} \right]
\end{aligned}$$

**Proposition 6.15.**  $\left\langle \left. \xrightarrow[p]{(\infty)} X \xrightarrow{\hat{c}} T \xleftarrow{t} \right\rangle \right\rangle = \mathbb{E}_{x \sim p}[c(x)].$

*Proof.* Since  $p$  has high confidence, and  $T$  is always equal to  $t$ , the only joint distribution on  $(X, T)$  with finite score is  $\mu(X, T) = p(X)\delta_t(T)$ . We compute its score directly:

$$\begin{aligned}
\left\langle \left. \xrightarrow[p]{(\infty)} X \xrightarrow{\hat{c}} T \xleftarrow{t} \right\rangle \right\rangle &= \mathbb{E}_{\mu} \log \frac{\mu(X, T)}{\hat{c}(t | X)} = \mathbb{E}_p \log \frac{1}{\hat{c}(t | X)} = \mathbb{E}_p \log \frac{1}{\exp(-c(X))} \\
&= \mathbb{E}_p \log \exp(c(X)) = \mathbb{E}_p c(X) = \mathbb{E}_{x \sim p} c(x). && \square
\end{aligned}$$

### 6.C.1 Additional Proofs for Inline Claims

#### Details on the Data Processing Inequality Proof

We now provide more details on the proof of the Data Processing Equality that appeared in [Figure 6.2](#) of the main text. We repeat that visual proof here for convenience, augmented with labels for the PDGs ( $m_1, \dots, m_5$ ) and numbered (in)equalities.

$$\begin{array}{ccc}
m_1 & m_2 & m_3 \\
\left\langle \frac{p}{(\beta)} \rightarrow [X] \xleftarrow{q} \right\rangle \stackrel{(1)}{=} \left\langle \frac{p}{(\beta)} \rightarrow [X] \xleftarrow{q} \overset{Y}{\underset{f \uparrow (\beta+\zeta)}{\xleftarrow{}} \right\rangle \stackrel{(2)}{=} \left\langle \frac{p}{(\beta)} \rightarrow [X_1] = [X_2] \xleftarrow{q} \overset{Y}{\underset{f \nearrow (\beta) \quad f \swarrow (\zeta)}{\xleftarrow{}}} \right\rangle \\
& & \stackrel{(3)}{\geq} \left\langle \frac{p}{(\beta)} \rightarrow [X_1] \overset{f \nearrow (\beta)}{\xrightarrow{}} [X_2] \xleftarrow{q} \overset{Y}{\underset{f \swarrow (\zeta)}{\xleftarrow{}}} \right\rangle \stackrel{(4)}{=} \left\langle \frac{f \circ p}{(\beta)} \rightarrow [X] \xleftarrow{f \circ q} \right\rangle \\
& & m_4 \qquad \qquad \qquad m_5
\end{array}$$

We now enumerate the (in)equalities to prove them.

1. Let  $\mu(X)$  denote the (unique) optimal distribution for  $m_1$ . Now, the joint distribution  $\mu(X, Y) := \mu(X)f(Y|X)$  has incompatibility with  $m_2$  equal to

$$\begin{aligned}
& OInc_{m_2}(\mu(X, Y)) \\
&= \beta D(\mu(X) \parallel p(X)) + \zeta D(\mu(X) \parallel q(X)) + (\beta + \zeta) \mathbb{E}_{x \sim \mu} [D(\mu(Y|x) \parallel f(Y|x))] \\
&= OInc_{m_1}(\mu(X)) + (\beta + \zeta) \mathbb{E}_{x \sim \mu} D(\mu(Y|x) \parallel f(Y|x)) \\
&= \langle\!\langle m_1 \rangle\!\rangle \quad \begin{bmatrix} \text{as } \mu(Y|x) = f(Y|x) \text{ wherever } \mu(x) > 0, \\ \text{and } \mu(X) \text{ minimizes } OInc_{m_1} \end{bmatrix}.
\end{aligned}$$

So  $\mu(X, Y)$  witnesses the fact that  $\langle\!\langle m_2 \rangle\!\rangle \leq OInc_{m_2}(\mu(X, Y)) = \langle\!\langle m_1 \rangle\!\rangle$ . Furthermore, every joint distribution  $\nu(X, Y)$  must have at least this incompatibility, as it must have some marginal  $\nu(X)$ , which, even by itself, already gives rise to incompatibility of magnitude  $OInc_{m_1}(\nu(X)) \geq OInc_{m_1}(\mu(X)) = \langle\!\langle m_1 \rangle\!\rangle$ . And since this is true for all  $\nu(X, Y)$ , we have that  $\langle\!\langle m_2 \rangle\!\rangle \geq \langle\!\langle m_1 \rangle\!\rangle$ . So  $\langle\!\langle m_2 \rangle\!\rangle = \langle\!\langle m_1 \rangle\!\rangle$ .

2. The equals sign in  $m_3$  may be equivalently interpreted as a cpd  $eq(X_1|X_2) := x_2 \mapsto \delta_{x_2}(X_1)$ , a cpd  $eq'(X_2|X_1) := x_1 \mapsto \delta_{x_1}(X_2)$ , or both at once; in each case, the effect is that a joint distribution  $\mu$  with support on an outcome for which  $X_1 \neq X_2$  gets an infinite penalty, so a minimizer  $\mu(X_1, X_2, Y)$  of  $OInc_{m_3}$  must be isomorphic to a distribution  $\mu'(X, Y)$ .

Furthermore, it is easy to verify that  $OInc_{m_2}(\mu'(X, Y)) = OInc_{m_3}(\mu(X, X, Y))$ .

More formally, we have:

$$\langle\langle m_3 \rangle\rangle = \inf_{\mu(X_1, X_2, Y)} \mathbb{E}_{\mu} \left[ \begin{array}{ccc} \beta \log \frac{\mu(X_1)}{p(X_1)} & +\zeta \log \frac{\mu(X_2)}{q(X_2)} & + \log \frac{\mu(X_1|X_2)}{eq(X_1, X_2)} \\ +\beta \log \frac{\mu(Y|X_1)}{f(Y|X_1)} & +\zeta \log \frac{\mu(Y|X_2)}{f(Y|X_2)} & \end{array} \right]$$

but if  $X_1$  always equals  $X_2$  (which we call simply  $X$ ), as it must for the optimal distribution  $\mu$ , this becomes

$$\begin{aligned} &= \inf_{\mu(X_1=X_2=X, Y)} \mathbb{E}_{\mu} \left[ \beta \log \frac{\mu(X)}{p(X)} + \zeta \log \frac{\mu(X)}{q(X)} + \beta \log \frac{\mu(Y|X)}{f(Y|X)} + \zeta \log \frac{\mu(Y|X)}{f(Y|X)} \right] \\ &= \inf_{\mu(X, Y)} \mathbb{E}_{\mu} \left[ \beta \log \frac{\mu(X)}{p(X)} + \zeta \log \frac{\mu(X)}{q(X)} + (\beta + \zeta) \log \frac{\mu(Y|X)}{f(Y|X)} \right] \\ &= \inf_{\mu(X, Y)} OInc_{m_2}(\mu) \\ &= \langle\langle m_2 \rangle\rangle. \end{aligned}$$

3. Eliminating the edge or edges enforcing the equality ( $X_1 = X_2$ ) cannot increase inconsistency, by Lemma 6.1.

4. Although this final step of composing the edges with shared confidences looks intuitively like it should be true (and it is!), its proof may not be obvious. We now provide a rigorous proof of this equality.

To ameliorate subscript pains, we henceforth write  $X$  for  $X_1$ , and  $Z$  for  $X_2$ .

We now compute:

$$\begin{aligned} \langle\langle m_4 \rangle\rangle &= \inf_{\mu(X, Z, Y)} \mathbb{E}_{\mu} \left[ \beta \log \frac{\mu(X) \mu(Y|X)}{p(X) f(Y|X)} + \zeta \log \frac{\mu(Z) \mu(Y|Z)}{q(Z) f(Y|Z)} \right] \\ &= \inf_{\mu(X, Z, Y)} \mathbb{E}_{\mu} \left[ \beta \log \frac{\mu(Y) \mu(X|Y)}{p(X) f(Y|X)} + \zeta \log \frac{\mu(Y) \mu(Z|Y)}{q(Z) f(Y|Z)} \right] \end{aligned}$$

by applying Bayes Rule in each numerator. By the chain rule, every distribution  $\mu(X, Z, Y)$  may be specified as  $\mu(Y)\mu(X|Y)\mu(Z|X, Y)$ , so we can rewrite the formula above as  $\langle\langle m_4 \rangle\rangle =$

$$\inf_{\mu(Y)} \inf_{\mu(X|Y)} \inf_{\mu(Z|Y, X)} \mathbb{E}_{y \sim \mu(Y)} \mathbb{E}_{x \sim \mu(X|y)} \mathbb{E}_{z \sim \mu(Z|y, x)} \left[ \beta \log \frac{\mu(y) \mu(x|y)}{p(x) f(y|x)} + \zeta \log \frac{\mu(y) \mu(z|y)}{q(z) f(y|z)} \right],$$

where  $\mu(Z|Y)$  is defined in terms of the primitives  $\mu(X|Y)$  and  $\mu(Z|X, Y)$  as  $\mu(Z|Y) := y \mapsto \mathbb{E}_{x \sim \mu(X|y)} \mu(Z|y, x)$ , and is a valid cpd, since it is a mixture distribution. Since the first term (with  $\beta$ ) does not depend on  $z$ , we can take it out of the expectation, so  $\langle\!\langle m_4 \rangle\!\rangle$  is in fact equal to

$$\inf_{\mu(Y)} \inf_{\mu(X|Y)} \inf_{\mu(Z|Y, X)} \mathbb{E}_{y \sim \mu(Y)} \mathbb{E}_{x \sim \mu(X|y)} \left[ \beta \log \frac{\mu(y) \mu(x|y)}{p(x) f(y|x)} + \zeta \mathbb{E}_{z \sim \mu(Z|y, x)} \left[ \log \frac{\mu(y) \mu(z|y)}{q(z) f(y|z)} \right] \right];$$

we can split up  $\mathbb{E}_{\mu(X|y)}$  by linearity of expectation, to get the expression

$$\inf_{\mu(Y)} \inf_{\mu(X|Y)} \inf_{\mu(Z|Y, X)} \mathbb{E}_{y \sim \mu(Y)} \left[ \beta \mathbb{E}_{x \sim \mu(X|y)} \left[ \log \frac{\mu(y) \mu(x|y)}{p(x) f(y|x)} \right] + \zeta \mathbb{E}_{\substack{x \sim \mu(X|y) \\ z \sim \mu(Z|y, x)}} \left[ \log \frac{\mu(y) \mu(z|y)}{q(z) f(y|z)} \right] \right].$$

Note that the quantity inside the second expectation does not depend on  $x$ . Therefore, the second expectation is just an explicit way of sampling  $z$  from the mixture distribution  $\mathbb{E}_{x \sim \mu(X|y)} \mu(Z|x, y)$ , which is the definition of  $\mu(Z|y)$ . Once we make this replacement, it becomes clear that the only feature of  $\mu(Z|Y, X)$  that matters is the mixture  $\mu(Z|Y)$ . Simplifying the second expectation in this way, and replacing the infimum over  $\mu(Z|X, Y)$  with one over  $\mu(Z|Y)$  yields:

$$\langle\!\langle m_4 \rangle\!\rangle = \inf_{\mu(Y)} \inf_{\mu(X|Y)} \inf_{\mu(Z|Y)} \mathbb{E}_{y \sim \mu(Y)} \left[ \beta \mathbb{E}_{x \sim \mu(X|y)} \left[ \log \frac{\mu(y) \mu(x|y)}{p(x) f(y|x)} \right] + \zeta \mathbb{E}_{z \sim \mu(Z|y)} \left[ \log \frac{\mu(y) \mu(z|y)}{q(z) f(y|z)} \right] \right].$$

Now, a cpd  $\mu(X|Y)$  is just<sup>6</sup> a (possibly different) distribution  $\nu_y(X)$  for every value of  $Y$ . Observe that, inside the expectation over  $\mu(Y)$ , the cpds  $\mu(X|Y)$  and  $\mu(Z|Y)$  are used only for the *present* value of  $y$ , and do not reference, say,  $\mu(X|y')$  for  $y' \neq y$ . Because there is no interaction between the choice of cpd  $\mu(X|y)$  and  $\mu(X|y')$ , it is not necessary to jointly optimize over entire cpds  $\mu(X|Y)$  all at once. Rather, it is equivalent to take the infimum over  $\nu(X)$ , separately for each  $y$ . Symmetrically, we may as well take the infimum over  $\lambda(Z)$  separately for each

---

<sup>6</sup>modulo measurability concerns that do not affect the infimum; see [Section 6.A](#)

$y$ , rather than jointly finding the optimal  $\mu(Z|Y)$  all at once. Operationally, this means we can pull the infima inside the expectation over  $Y$ . And since the first term doesn't depend on  $Z$  and the second doesn't depend on  $X$ , we get:

$$\langle\!\langle m_4 \rangle\!\rangle = \inf_{\mu(Y)} \mathbb{E}_{y \sim \mu(Y)} \left[ \inf_{\nu(X)} \beta \mathbb{E}_{\nu(X)} \left[ \log \frac{\mu(y) \nu(X)}{p(X) f(y|X)} \right] + \inf_{\lambda(Z)} \zeta \mathbb{E}_{\lambda(Z)} \left[ \log \frac{\mu(y) \lambda(Z)}{q(Z) f(y|Z)} \right] \right]$$

Next, we pull the same trick we've used over and over: find constants so that we can regard the dependence as a relative entropy with respect to the quantity being optimized. Grouping the quantities apart from  $\nu(X)$  on the left term and normalizing them (and analogously for  $\lambda(Z)$  on the right), we find that

$$\langle\!\langle m_4 \rangle\!\rangle = \inf_{\mu(Y)} \mathbb{E}_{y \sim \mu(Y)} \left[ \begin{array}{l} \beta \inf_{\nu(X)} D\left(\nu(X) \middle\| \frac{1}{C_1(y)} p(X) \frac{f(y|X)}{\mu(y)}\right) - \beta \log C_1(y) \\ + \zeta \inf_{\lambda(Z)} D\left(\lambda(Z) \middle\| \frac{1}{C_2(y)} q(Z) \frac{f(y|Z)}{\mu(y)}\right) - \zeta \log C_2(y) \end{array} \right],$$

where

$$C_1(y) = \sum_x p(x) \frac{f(y|x)}{\mu(y)} = \frac{1}{\mu(y)} \mathbb{E}_{p(X)} f(y|X)$$

and

$$C_2(y) = \sum_z q(z) \frac{f(y|z)}{\mu(y)} = \frac{1}{\mu(y)} \mathbb{E}_{q(Z)} f(y|Z)$$

are the constants required to normalize the distributions. Both relative entropies are minimized when their arguments match, at which point they contribute zero, so we have

$$\begin{aligned} \langle\!\langle m_4 \rangle\!\rangle &= \inf_{\mu(Y)} \mathbb{E}_{y \sim \mu(Y)} \left[ \beta \log \frac{1}{C_1(y)} + \zeta \log \frac{1}{C_2(y)} \right] \\ &= \inf_{\mu(Y)} \mathbb{E}_{y \sim \mu(Y)} \left[ \beta \log \frac{\mu(y)}{\mathbb{E}_{p(X)} f(y|X)} + \zeta \log \frac{\mu(y)}{\mathbb{E}_{q(Z)} f(y|Z)} \right] \\ &= \inf_{\mu(Y)} \mathbb{E}_{\mu} \left[ \beta D(\mu \| f \circ p) + \zeta D(\mu \| f \circ q) \right] \\ &= \langle\!\langle m_5 \rangle\!\rangle. \end{aligned}$$

## Details for Claims made in Section 6.7

**(Claim 6.7.1)** First, the fact that

$$\mathcal{L}_1 = \lambda_d \mathcal{L}_{\text{dat}} + \lambda_s \mathcal{L}_{\text{sim}} = \left\langle \frac{\lambda}{(\infty)} \xrightarrow{\text{Z}} \boxed{\begin{array}{c} \bullet \\ \text{sim dat} \\ \bullet \end{array}} \xrightarrow[\text{(infty)}]{\begin{array}{l} \text{dat} \mapsto d \\ \text{sim} \mapsto s \end{array}} \boxed{\begin{array}{c} X \\ h \\ Y \end{array}} \right\rangle,$$

where  $\lambda(Z = \text{sim}) = \lambda_s$  and  $\lambda(Z = \text{dat}) = \lambda_d$  is immediate. The two cpds with infinite confidence ensure that the only joint distribution with a finite score is  $\lambda_s s + \lambda_d d$ , and the inconsistency with  $h$  is its surprisal, so the inconsistency of this PDG is

$$\begin{aligned} \mathbb{E}_{\lambda_s s + \lambda_d d} \left[ \log \frac{1}{h(Y|X)} \right] &= -\lambda_s \mathbb{E}_s [\log h(Y|X)] - \lambda_d \mathbb{E}_d [\log h(Y|X)] \\ &= \lambda_d \mathcal{L}_{\text{dat}} + \lambda_s \mathcal{L}_{\text{sim}} \\ &= \mathcal{L}_1, \quad \text{as promised.} \end{aligned}$$

**(Claim 6.7.2)** The second correspondence is the least straightforward. Let  $C : \int_{V(X,Y)} d s \cdot d$  be the normalization constant required to normalize the joint density  $s \cdot d$ . We claim that, for large fixed  $\gamma$ , we have

$$\mathcal{L}_2 \approx C \left\langle \frac{s}{(\alpha:1)_{\beta:\gamma}} \xrightarrow{h} \boxed{\begin{array}{c} X \\ Y \end{array}} \xrightarrow{d} \frac{d}{(\alpha:1)_{\beta:\gamma}} \right\rangle_\gamma + \text{const},$$

where  $\text{const}$  does not depend on  $h$ . To see this, let  $m_2$  be the PDG above, and compute

$$\begin{aligned} \langle m_2 \rangle_\gamma &= \inf_{\mu(X,Y)} \mathbb{E}_\mu \left[ \overbrace{\gamma \log \frac{\mu(XY)}{s(XY)} \frac{\mu(XY)}{d(XY)} + \log \frac{\mu(Y|X)}{h(Y|X)}}^{OInc(\mu)} \right. \\ &\quad \left. + \gamma \log \frac{1}{s(XY)} \frac{1}{d(XY)} - \gamma \log \frac{1}{\mu(XY)} \right] \\ &= \inf_{\mu(X,Y)} \mathbb{E}_\mu \left[ \gamma \log \frac{\mu(XY)}{s(XY)} \frac{\mu(XY)}{d(XY)} \frac{1}{\mu(XY)} \frac{1}{\mu(XY)} \frac{\mu(XY)}{1} + \log \frac{\mu(Y|X)}{h(Y|X)} \right] \end{aligned}$$

$$\begin{aligned}
&= \inf_{\mu(X,Y)} \mathbb{E}_{\mu} \left[ \gamma \log \frac{\mu(XY)}{s(XY)d(XY)} + \log \frac{\mu(Y|X)}{h(Y|X)} \right] \\
&= \inf_{\mu(X,Y)} \mathbb{E}_{\mu} \left[ \gamma \log \frac{\mu(XY)C}{s(XY)d(XY)} - \gamma \log C + \log \frac{\mu(Y|X)}{h(Y|X)} \right] \\
&= \inf_{\mu(X,Y)} \gamma D\left(\mu \middle\| \frac{1}{C}sd\right) + \mathbb{E}_{\mu} \left[ \log \frac{\mu(Y|X)}{h(Y|X)} \right] - \gamma \log C
\end{aligned}$$

$D$  is  $(\gamma m)$ -strongly convex in a region around its minimizer for some  $m > 0$  that depends only on  $s$  and  $d$ . Together with our assumption that  $h$  is positive, we find that when  $\gamma$  becomes large, the first term dominates, and the optimizing  $\mu$  quickly approaches the normalized density  $\nu := \frac{1}{C}s \cdot d$ . Plugging in  $\nu$ , we find that the value of the infimum approaches

$$\begin{aligned}
\langle\!\langle \mathbf{m}_2 \rangle\!\rangle &\approx \mathbb{E}_{\nu} \left[ \log \frac{1}{h(Y|X)} \right] - H_{\nu}(Y|X) - \gamma \log C \\
&= \int_{XY} \frac{1}{C} \log \frac{1}{h(Y|X)} s(X, Y) d(X, Y) - H_{\nu}(Y|X) - \gamma \log C \\
&= \frac{1}{C} \mathbb{E}_s \left[ d(X, Y) \log \frac{1}{h(Y|X)} \right] - H_{\nu}(Y|X) - \gamma \log C \\
&= \frac{1}{C} \mathcal{L}_2 - H_{\nu}(Y|X) - \gamma \log C,
\end{aligned}$$

and therefore  $\mathcal{L}_2 = C \langle\!\langle \mathbf{m}_2 \rangle\!\rangle + C H_{\nu}(Y|X) - \gamma C \log C$

$$= C \langle\!\langle \mathbf{m}_2 \rangle\!\rangle + \text{const.}$$

**(Claim 6.7.3)** Finally, we turn to

$$\mathcal{L}_3 := \left\langle\!\left\langle \frac{s}{(\lambda_s)} \xrightarrow{h} \begin{matrix} X \\ h \downarrow \\ Y \end{matrix} \xleftarrow{d} \frac{d}{(\lambda_d)} \right\rangle\!\right\rangle.$$

To see why the optimal distribution  $\mu^*(XY)$  is the  $\lambda$ -weighted geometric mean of  $s$  and  $d$ , let us first consider the same PDG, except without  $h$ . From Lemma 6.10, we have this loss without  $h$  in closed form, and from the proof of Lemma 6.10, we see that the optimizing distribution in this case is the  $\lambda$ -weighted

geometric distribution  $\mu^* \propto s(XY)^{\lambda_s} d(XY)^{\lambda_d}$ . Now (Lemma 6.1), including  $h$  cannot make the PDG any less inconsistent. In particular, by choosing

$$h^*(Y|X) := \mu^*(Y|X) \propto (Y|X)^{\lambda_s} d(Y|X)^{\lambda_d},$$

to be already compatible with this joint distribution, the inconsistency does not change, while choosing a different  $h$  would cause the inconsistency to increase. Thus, the optimal classifier  $h^*$  by this metric is indeed as we claim. Finally, it is easy to see that this loss is calibrated: if  $s = d$ , then the optimal joint distribution is equal to  $s$  and to  $d$ , and the optimal classifier is  $h(Y|X) = s(Y|X) = d(Y|X)$ . So  $\mathcal{L}_3$  is calibrated.

### Details for Claims made in Section 6.8

**Distortion Due to Inconsistency (Claim 6.8.1).** In the footnote on [Page 209](#), we claimed that if the model confidence  $\beta_p$  were 1 rather than  $\infty$ , we would have obtained an inconsistency of  $-\log \mathbb{E}_{x \sim p} \exp(-c(x))$ , and that the optimal distribution would not have been  $p(X)$ .

$$\begin{aligned} \left\langle \xrightarrow{p} [X] \xrightarrow{\hat{c}} [\mathbf{T}] \xleftarrow{\mathbf{t}} \right\rangle &= \inf_{\mu(X)} \mathbb{E}_{x \sim \mu} \left[ \log \frac{\mu(x)}{p(x)} + \log \frac{\mu(\mathbf{t}|x)}{\hat{c}(\mathbf{t}|x)} \right] \\ &= \inf_{\mu(X)} \mathbb{E}_{x \sim \mu} \left[ \log \frac{\mu(x)}{p(x)} + \log \frac{1}{\hat{c}(\mathbf{t}|x)} \right] \\ &= \inf_{\mu(X)} \mathbb{E}_{x \sim \mu} \left[ \log \frac{\mu(x)}{p(x) \exp(-c(x))} \cdot \frac{Z}{Z} \right] \end{aligned}$$

where  $Z := \sum_x p(x) \exp(-c(x)) = \mathbb{E}_p[\exp(-c(X))]$  is the constant required to normalize the distribution

$$\begin{aligned} &= \inf_{\mu(X)} D\left(\mu \middle\| \frac{1}{Z} p(X) \exp(-c(X))\right) - \log Z \\ &= -\log Z \end{aligned}$$

$$= - \log \mathbb{E}_{x \sim p} \exp(-c(x))$$

as promised. Note also that in the proof, we showed that the optimal distribution is proportional to  $p(X) \exp(-c(X))$  which means that it equals  $p(X)$  if and only if  $c(X)$  is constant in  $X$ .

**Enforcing the Qualitative Picture (Claim 6.8.2).** We also claimed in [Section 6.8](#) that, if  $\alpha_h = \alpha_{\Pr_D} = 1$ , then

$$\lim_{\gamma \rightarrow \infty} \left\langle \begin{array}{c} \Pr_D \\ \downarrow \\ X \end{array} \right\rangle \left\langle \begin{array}{c} Y \\ \hat{\ell} \\ \uparrow \\ T \end{array} \right\rangle = \mathbb{E}_{\substack{(x,y) \sim \Pr_D \\ y' \sim p(Y'|x)}} [\ell(y, y')]$$

Why is this? For such a setting of  $\alpha$ , which intuitively articulates a causal picture where  $X, Y$  is generated from  $\Pr_D$ , and  $Y'$  generated by  $h(Y'|X)$ , the information deficiency  $SDef_S(\mu(X, Y, Y'))$  of a distribution  $\mu$  is

$$\begin{aligned} SDef_S(\mu(X, Y, Y')) &= -H_\mu(X, Y, Y') + H(X, Y) + H(Y'|X) \\ &= H_\mu(Y'|X) - H_\mu(Y'|X, Y) \\ &= I_\mu(Y; Y'|X). \end{aligned}$$

Both equalities of the derivation above standard information-theoretic identities (See, for instance, [MacKay 2003](#)), and the final quantity  $I_\mu(Y; Y'|X)$  is the *conditional mutual information* between  $Y$  and  $Y'$  given  $X$ , and is a non-negative number that equals zero if and only if  $Y$  and  $Y'$  are conditionally independent given  $X$ .

As a result, as  $\gamma \rightarrow \infty$  any distribution that for which  $Y'$  and  $Y$  are not independent given  $X$  will incur infinite cost. Since the confidences in  $h$  and  $\Pr_D$  are also infinite, so will a violation of either cpd. There is only one distribution that has both cpds and also this independence; that distribution is  $\mu(X, Y, Y') :=$

$\Pr_{\mathcal{D}}(X, Y)h(Y'|X)$ . Now the argument of [Proposition 6.15](#) applies: all other cpds must be matched, and the inconsistency is the expected incompatibility of  $\hat{\ell}$ , which equals

$$\begin{aligned} \mathbb{E}_{\substack{(x,y) \sim \Pr_{\mathcal{D}} \\ y' \sim p(Y'|x)}} \log \frac{1}{\hat{\ell}(\mathbf{t}|y, y')} &= \mathbb{E}_{\substack{(x,y) \sim \Pr_{\mathcal{D}} \\ y' \sim p(Y'|x)}} \log \frac{1}{\exp(-\ell(y, y'))} \\ &= \mathbb{E}_{\substack{(x,y) \sim \Pr_{\mathcal{D}} \\ y' \sim p(Y'|x)}} [\log \exp(\ell(y, y'))] = \mathbb{E}_{\substack{(x,y) \sim \Pr_{\mathcal{D}} \\ y' \sim p(Y'|x)}} [\ell(y, y')] = \mathcal{L}. \end{aligned}$$

## CHAPTER 7

### THE LOCAL INCONSISTENCY RESOLUTION (LIR) ALGORITHM

In this chapter, we present a generic algorithm for learning and approximate inference across a broad class of statistical models, that unifies many approaches in the literature. Our algorithm, called local inconsistency resolution (LIR), has an intuitive epistemic interpretation. It is based on the theory of probabilistic dependency graphs (PDGs), an expressive class of graphical models rooted in information theory, which can capture inconsistent beliefs.

#### 7.1 Introduction

What causes a person to change their mind? According to some, it is a response to internal conflict: the result of discovering new information that contradicts our beliefs, or becoming aware of discrepancies between beliefs we already hold ([Festinger 1962](#)). Inconsistencies can be difficult to detect, however ([Selman et al. 1996](#)), and indeed can only be resolved once we are aware of them. Some things are also beyond our control; for example, we might receive conflicting information from two trusted sources and be unable to resolve their disagreement. So in practice, we resolve inconsistencies *locally*—little by little, and looking at only a small part of the picture at a time.

This can have externalities; fixing one inconsistency can easily create others out of view. Furthermore, some inconsistencies are not local in nature, and can only be seen when considering many components at once. Yet despite its imperfections, this process of locally resolving inconsistency can be quite useful. As we shall soon see, it is a powerful recipe for learning and approximate inference. We formalize the process in the language of probability and convex

optimization, and show how that many popular techniques in the literature arise naturally as instances of it.

Our approach leans heavily on the theory of PDGs. (??). As we have seen, there is a natural way to measure how inconsistent a PDG is; we saw in Chapter 6 that many standard loss functions can be viewed as measuring the inconsistency of a PDG that describes the appropriate situation. We introduce an algorithm to operationalize the process of adjusting parameters to resolve this inconsistency.

In general, even just calculating a PDG’s degree of inconsistency is intractable. Much of variational inference can be understood as adopting extra beliefs to minimize an overapproximation of it that is easier to calculate (Richardson 2022). Our approach can capture this, but also enables the opposite: focusing on small parts of the graph at a time to address tractable underapproximations of the global inconsistency. This makes it more suitable for distributed settings, and more amenable to parallelization. The algorithm, which we call *local inconsistency resolution* (LIR), is quite expressive, and naturally reduces to a wide variety of learning and inference algorithms in the literature. This observation suggests a generic approach to learning and inference in models with arbitrary structure.

## 7.2 Parametric PDGs

The main tool of this chapter is an alternate parametric version of a PDG. Just as a neural network can be thought of as a flexible functions (or even cpds) that can be adjusted with internal parameters, so too would we like to model PDGs whose cpds are parametric. For this reason, we will require a parameter space  $\Theta_a$  for each (hyper)arc  $a \in \mathcal{A}$ . For simplicity, assume that each  $\Theta_a$  is a convex

subset of  $\mathbb{R}^n$  (not necessarily of the same dimension). This allows us to formally define a *parametric* variant of a PDG, as promised in [Section 3.2.1](#).

**Definition 7.1.** A *Parametric Probabilistic Dependency Graph* (PPDG)  $\mathbf{m}(\Theta) = (\mathcal{X}, \mathcal{A}, \Theta, \mathbb{P}, \alpha, \beta)$  is a directed hypergraph  $(\mathcal{X}, \mathcal{A})$  whose nodes correspond to variables, each arc  $a \in \mathcal{A}$  of which is associated with:

- a parameter space  $\Theta_a \subseteq \mathbb{R}^n$ , with a default value  $\theta_a^{\text{init}}$ .
- a map  $\mathbb{P}_a : \Theta_a \times \mathcal{V}S_a \rightarrow \Delta \mathcal{V}T_a$  that gives a cpd  $\mathbb{P}_a^\theta(T_a | S_a)$  over  $a$ 's targets given its sources, for every  $\theta \in \Theta_a$ ,
- confidences  $\alpha_a \in \mathbb{R}$  in the functional dependence of  $T_a$  on  $S_a$  expressed by  $a$ , and  $\beta_a \in [0, \infty]$  in the cpd  $\mathbb{P}_a$ .

A PDG is the object obtained by fixing the parameters; thus, a choice of  $\theta \in \Theta := \prod_{a \in \mathcal{A}} \Theta_a$  yields a PDG  $\mathbf{m} = \mathbf{m}(\theta)$ . □

Clearly, a PDG is the special case of a PPDG in which every  $\Theta_a = \{\theta_a^{\text{init}}\}$  is a singleton. Conversely, a PPDG may be viewed as a PDG by adding each  $\Theta_a$  as a variable, as depicted in [Figure 7.1](#).

As argued in [Chapter 6](#), PDG inconsistency is a “universal” loss function, and specializes to standard loss functions in a wide variety of standard situations. It follows that, at an abstract level, much of machine learning can be viewed as inconsistency resolution. We take this idea a few steps further, by (1) operationalizing the resolution process with parametric PDGs, and (2) allowing for a heuristic approach that only resolves inconsistencies *locally*.

### 7.3 Local Inconsistency Resolution (LIR)

**Geometric Preliminaries** To describe how those parameters evolve over time, we will need some additional geometric concepts. A *vector field* over  $\Theta$  is a differentiable map  $X$  assigning to each  $\theta \in \Theta$  a vector  $X_\theta \in \mathbb{R}^n$ . The *gradient* of a twice differentiable map  $f : \Theta \rightarrow \mathbb{R}$ , which we write  $\nabla_\Theta f(\Theta)$ , is a vector field. Given a vector field  $X$  and an initial point  $\theta_0 \in \Theta$ , there is a unique trajectory  $y(t)$  that solves the ODE  $\{\frac{d}{dt}y(t) = X_{y(t)}, y(0) = \theta_0\}$ , and we adopt the notation  $\exp_{\theta_0}(X) := y(1)$  for a compact description of it. At first glance,  $\exp$  only gives us access to  $y(1)$ , but it is easily verified that  $\exp_{\theta_0}(tX) = y(t)$ . So altogether, the map  $t \mapsto \exp_\theta(t\nabla_\Theta f(\Theta))$  is the smooth path beginning at  $\theta$  that follows the gradient of  $f$ . It is known as *gradient flow*.

**Attention and Control.** There are two distinct senses in which inconsistency resolution can be *local*: we can restrict what we can see, or what we can do about it. Correspondingly, there are two “focus” knobs for our algorithm: one that restricts our attention to the inconsistency of a subset of arcs  $A \subseteq \mathcal{A}$ , and another that restricts our control to (only) the parameters of arcs  $C \subseteq \mathcal{A}$  as we resolve that inconsistency. The former makes for an underestimate of the inconsistency that is easier to calculate, while the latter makes for an easier optimization problem. These restrictions are not just cheap approximations, though: they are also appropriate modeling assumptions for actors that cannot see and control everything at once.

Attention and control need not be black or white. A more general approach is to choose an *attention mask*  $\varphi \in \mathbb{R}^{\mathcal{A}}$  and a *control mask*  $\chi \in [0, \infty]^{\mathcal{A}}$ . Large  $\varphi(a)$  makes  $a$  salient, while  $\varphi(a) = 0$  keeps it out of the picture. Similarly, large

$\chi(a)$  gives significant freedom to change  $a$ 's parameters, small  $\chi(a)$  affords only minor adjustments, and  $\chi(a) = 0$  prevents change altogether. Either mask can then be applied to a tensor that has an axis corresponding to  $\mathcal{A}$ , via pointwise multiplication ( $\odot$ ).

**The Algorithm.** LIR modifies the parameters  $\theta$  of a PPDG  $m(\Theta)$  so as to make it more consistent with its context. It proceeds as follows. First, receive context in the form of a PDG  $Ctx$ , and initialize mutable memory  $m(\Theta)$ . In each iteration, choose  $\gamma$  (which can be viewed as attention to structure), an attention mask  $\varphi$  over the arcs of  $m(\Theta) + Ctx$ , and a control mask  $\chi$  over the arcs of  $m(\Theta)$ . Calculate  $\langle\!\langle \varphi \odot (m(\theta) + Ctx) \rangle\!\rangle_\gamma$ , the inconsistency of the combined context and memory, weighted by attention. (For discrete PDGs, this can be done with the methods of Chapter 8.) Then mitigate this local inconsistency by updating mutable memory  $\theta$  via (an approximation to) gradient flow, changing  $a$ 's parameters in proportion to control  $\chi(a)$ . The procedure is fully formalized in [Algorithm 1](#).

---

#### Algorithm 1 Local Inconsistency Resolution (LIR)

---

```

Input: context  $Ctx$ , mutable memory  $m(\Theta)$ .
Initialize  $\theta^{(0)} \leftarrow \theta^{\text{init}}$ ;
for  $t = 0, 1, 2, \dots$  do
     $Ctx \leftarrow \text{REFRESH}(Ctx)$ ;                                //optional
     $\varphi, \chi, \gamma \leftarrow \text{REFOCUS}()$ ;
     $\theta^{(t+1)} \leftarrow \exp_{\theta^{(t)}} \left\{ -\chi \odot \nabla_\Theta \langle\!\langle \varphi \odot (Ctx + m(\Theta)) \rangle\!\rangle_\gamma \right\}$ ;

```

---

In order to execute this procedure, we must say something about how the choice of  $(\varphi, \chi, \gamma)$  is made. Thus, we must supply an additional procedure REFOCUS to select attention and control masks. We focus mostly on the case where  $\gamma$  is fixed, and REFOCUS chooses non-deterministically from a fixed set of attention/control mask pairs  $(\varphi, \chi) \in \mathbf{F}$ , which we call *foci*. [Algorithm 1](#) also

allows us to select a second procedure, REFRESH, which makes it easier to model receiving new information in online settings.

The ODE on the last line of [Algorithm 1](#), which is an instance of gradient flow, may be approximated with an inner loop running an iterative gradient-based optimization algorithm. Alternatively, if REFOCUS produces small  $\chi$ , then it is well-approximated by a single gradient descent step of size  $\chi$ . At the other extreme: if  $\chi$  is infinite in every component, then, the final line reduces to

$$\theta^{(t+1)} \leftarrow \arg \min_{\theta} \langle\!\langle \varphi \odot (\mathcal{C}tx + \mathbf{m}(\theta)) \rangle\!\rangle_{\gamma},$$

at least in the typical case in which the parameterizations  $\mathbb{P}$  unconditional and log-concave. This is because of the following result.

**Proposition 7.1.** *If  $\mathbf{m}(\Theta)$  is a PPDG whose parameterizations  $\mathbb{P}$  are either constant or unconditional and log-concave, then for small enough  $\gamma$ , the map  $\theta \mapsto \langle\!\langle \varphi \odot (\mathcal{C}tx + \mathbf{m}(\theta)) \rangle\!\rangle_{\gamma}$  is convex.*

In the remaining sections, we give a sample of some historically important algorithms that are instances of LIR.

## 7.4 LIR in the Classification Setting

Consider a parametric classifier  $p_{\theta}(Y|X)$ , perhaps arising from a neural network whose final layer is a softmax. Suppose  $\mathcal{V}Y$  is a finite set of classes. If  $\mathcal{V}X$  is itself a manifold (such as the space of images), we can regard a value  $x \in \mathcal{V}X$  as parameterizing a deterministic cpd, written  $\xrightarrow{x} \boxed{X}$ . Together with a labeled sample  $(x, y)$ , we get a PPDG  $\mathbf{m}(\theta) := \xrightarrow{x} \boxed{X} \xrightarrow{p_{\theta}} \boxed{Y} \xleftarrow{y}$  whose observational

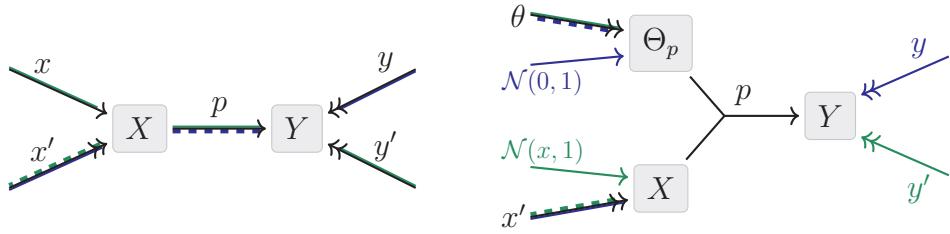
inconsistency is  $\langle\!\langle \mathcal{M} \rangle\!\rangle_0 = -\log p_\theta(y|x)$ , the standard training objective for such a classifier (Richardson 2022). Each cpd plays major role in this inconsistency.

What happens when we resolve this inconsistency by modifying the parameters associated to different arcs?

- Adjusting  $\theta$  amounts to training the network in the standard way—that is, changing the parameters little by little to minimize cross entropy. In this case, the value  $\chi$  of the control mask corresponds roughly to the product of the learning rate and the number of optimization iterations.
- Adjusting  $y$  is like a forward pass, in that it changes the label  $y$  to match the distribution  $p_\theta(Y|x)$ . that the model predicts over labels, given the input  $x$ .
- Adjusting  $x$  creates an adversarial example. That is, it makes incremental changes to the input  $x$  until the (fixed) network assigns it label  $y$ .

**Stochastic Gradient Descent (SGD).** Take the mutable state to be the classifier  $p$  as before. Define REFRESH so that it draws a batch of samples  $\{(x_i, y_i)\}_{i=1}^m$ , and returns a PDG with a single arc describing their emperical distribution  $d(X, Y)$ ; let REFOCUS be such that  $\varphi(d) = \infty$  (reflecting high confidence in the data). If  $\eta := \chi(p)\varphi(p)$  is small, then LIR is SGD with batch size  $m$  and learning rate  $\eta$ .

**Adversarial training.** Suppose we want to slightly alter  $x$  to obtain  $x'$  that is classified as  $y'$  instead of  $y$ . By adding arcs corresponding to  $x'$  and  $y'$  to  $\mathcal{M}$ , and relaxing the cpd  $\mathbb{P}_x$  associated with  $x$  to be a Gaussian centered  $x$  rather than a point mass, we get the PPDG on the left of Figure 7.1. An iteration of LIR whose focus is the edges marked in green (with control over the dashed green edge) is then an adversarial attack with Euclidean distance (Biggio et al. 2013). The blue focus, by contrast, “patches” the adversarial example by adjusting the model



*Figure 7.1:* Two illustrations of adversarial training. Left: the PPDG obtained by including a perturbed input  $x'$  and target  $y'$  to the classification setting. Right: the PDG obtained by making the parameters for  $p$  explicit, together with a Gaussian prior  $\Theta_p \sim \mathcal{N}(0, 1)$  over them. Both are colored with two foci: the blue focus trains the network, and the green one creates adversarial examples. Dashes indicate control.

parameters to again classify it correctly. Thus, LIR that alternates between the two foci, in which REFRESH selects a fresh  $(x, y, x' = x)$  from the dataset and target label  $y'$ , is adversarial training, a standard defense to adversarial attacks (Goodfellow et al. 2014).

The ML community's focus on adversarial examples may appear to be a cultural phenomenon, but mathematically, it is no accident. At this level of abstraction, there is no difference between model parameters and inputs. Indeed, if we make the parameterization of  $p$  explicit and add L2 regularization (i.e., a Gaussian prior over  $\Theta_p$ ), the symmetry becomes striking (Figure 7.1, right). This may help explain why, even outside of adversarial contexts, it can be just as sensible to train an input, as a model (Kishore et al. 2021).

## 7.5 The EM Algorithm as LIR

Suppose we have a generative model  $p(Z, X|\Theta)$  describing the probability over an observable variable  $X$  and a latent one  $Z$ . Given an observation  $X=x$ , the standard approach for trying to learn the parameters despite the missing data is

called the EM algorithm. It iteratively computes

$$\theta_{\text{EM}}^{(t+1)} = \arg \max_{\theta} \mathbb{E}_{z \sim p(Z|x, \theta_{\text{EM}}^{(t)})} [\log p(x, z|\theta)].$$

**Proposition 7.2.**  $LIR\left(\xrightarrow{x} X, \begin{array}{c} p \\ \swarrow \downarrow \searrow \\ X \quad Z \end{array} \xleftarrow[\infty]{q}\right)$  in which REFOCUS fixes  $\varphi = 1$  and alternates between full control of  $p$  and  $q$  implements EM, in that  $\theta_{\text{EM}}^{(t)} = \theta_{LIR}^{(2t)}$ .

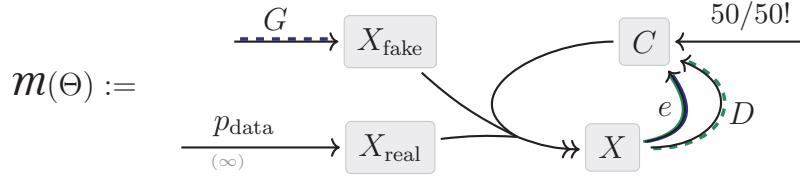
This result is closely related to one due to [Neal and Hinton \(1998\)](#), who view it as an intuitive explanation of why the EM algorithm works. Indeed, it is obvious in this form that every adjustment reduces the overall inconsistency. The result can also be readily adapted to an entire dataset by replacing  $x$  with a high confidence empirical distribution, or batched with the same technique in [Section 7.4](#). It also captures fractional EM when  $\chi < \infty$ .

This form of the EM algorithm is closely related to variational inference. Indeed, analogous choices applied to the analysis of [Richardson \(2022\)](#) yields the usual training algorithm for variational autoencoders (VAEs).

## 7.6 Generative Adversarial Training as LIR

LIR also subsumes more complex training procedures such as the one used to train GANs ([Goodfellow et al. 2020](#)). The goal is to train a network  $G$  to generate images that cannot be distinguished from real ones. More precisely, define  $X$  to be either an image  $X_{\text{fake}} \sim G$  or from a dataset  $X_{\text{real}} \sim p_{\text{data}}$ , based on a fair coin  $C$ . A discriminator  $D$  then predicts  $C$  from  $X$ . The generator also has a belief that, even given  $X$ , the coin is equally likely heads as tails (call this  $e$ ). This state

of affairs is summarized below.



The GAN objective is typically written as a 2-player minimax game:  
 $\min_G \min_D \mathcal{L}^{\text{GAN}}(G, D)$ , where

$$\mathcal{L}^{\text{GAN}}(G, D) = \mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{x' \sim G} [\log(1 - D(x'))],$$

and  $D(x)$  stands for the conditional probability  $D(C=1 \mid X = x)$ .

**The Discriminator's Focus.** The discriminator has full control over  $D$ , and attends to everything but  $e$ . That inconsistency of this PDG is what might be called the discriminator's objective: the expected KL divergence from  $D$  to the optimal discriminator. If  $D$  also disbelieves that any image is equally likely to be fake as real (by choosing  $\varphi(e) = -1$ ), then the inconsistency becomes  $-\mathcal{L}^{\text{GAN}}$ .

**The Generator's Focus.** The generator has control over  $G$ . If it ignores  $D$  attends only to  $e$ , the inconsistency is the Jensen-Shannon Divergence between  $G$  and  $p_{\text{data}}$ . If the generator also disbelieves the discriminator  $D$  (i.e.,  $\varphi(D) = -1$ ), then the inconsistency becomes  $+\mathcal{L}^{\text{GAN}}$ .

Standard practice is to use small  $\chi(G)$  and large  $\chi(D)$ , so that the discriminator is well-adapted to the generator.

## 7.7 Message Passing Algorithms as LIR

Nearly every standard graphical model can be viewed as a factor graph, and correspondingly admits an (approximate) inference procedure known variously as

(loopy) belief propagation ([Koller and Friedman 2009](#)), the generalized distributive law ([Aji and McEliece 2000](#)), and the sum-product algorithm ([Kschischang et al. 2001](#)). It also turns out to be the special case of LIR specialized to factor graphs.

Recall (from [Section 2.5.2](#)) that a *factor graph* over a set of variables  $\mathcal{X}$  is a set of factors  $\Phi = \{\phi_a : \mathbf{X}_a \rightarrow \mathbb{R}_{\geq 0}\}_{a \in \mathcal{A}}$ , where each  $\mathbf{X}_a \subseteq \mathcal{X}$  is called the *scope* of  $a$ . Conversely, for  $X \in \mathcal{X}$ , let  $\partial X := \{a \in \mathcal{A} : X \in \mathbf{X}_a\}$  be the set of factors with  $X$  in scope.  $\Phi$  specifies a distribution  $\Pr_{\Phi}(\mathcal{X}) \propto \prod_a \phi_a(\mathbf{X}_a)$ , and corresponds to a PDG

$$\mathbf{m}_{\Phi} = \left\{ \xrightarrow[\alpha, \beta=1]{\infty \phi_a} \boxed{\mathbf{X}_a} \right\}_{a \in \mathcal{A}}$$

that specifies the same joint distribution  $\Pr_{\Phi}$ , when observation and structure are weighted equally (i.e.,  $\gamma = 1$ ).

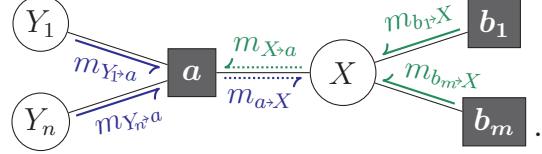
Sum-product belief propagation ([Kschischang et al. 2001](#)) aims to approximate marginals of  $\Pr_{\Theta}$  with only local computations: messages sent between factors and the variables they have in scope. Its state consists of pairs of “messages”  $\{m_{X \rightarrow a}, m_{a \rightarrow X}\}$ , both (unnormalized) distributions over  $X$ , for each pair  $(a, X)$  with  $a \in \partial X$ , which together form a PDG  $\mathbf{msg}$  in the same way as the original factor graph. After initialization, belief propagation repeatedly recomputes:

$$m_{X \rightarrow a}(x) := \prod_{b \in \partial X \setminus a} m_{b \rightarrow X}(x) \tag{7.1}$$

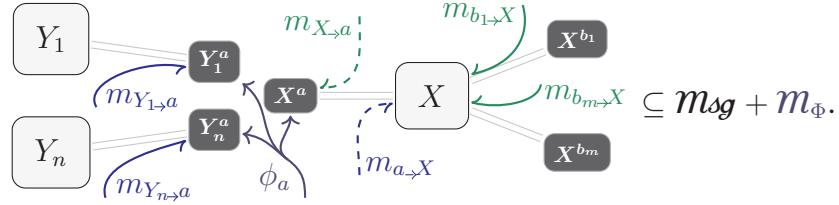
$$m_{a \rightarrow X}(x) := \sum_{\mathbf{y} \in \mathcal{V}(\mathbf{X}_a \setminus X)} \phi_a(x, \mathbf{y}) \prod_{Y \in \mathbf{X}_a \setminus X} m_{Y \rightarrow a}(Y(\mathbf{y})), \tag{7.2}$$

where  $Y(\mathbf{y})$  is the value of  $Y$  in the joint setting  $\mathbf{y}$ . Finally, variable marginals  $\{b_X\}_{X \in \mathcal{X}}$ , which we regard as another PDG,  $\mathcal{B}$ , are computed from the messages according to  $b_X(x) \propto \prod_{a \in \partial X} m_{a \rightarrow X}(x)$ . Observe that every calculation is a (marginal of) a product of factors, and thus amounts to inference in some “local”

factor graph. The traditional depiction of messages moving between variables and factors (Kschischang et al. 2001) looks something like this:



This is only a schematic, but the PDG  $M_{\delta g}$  can be made to look similar to it. Adding a variable  $X^a$  for every pair  $(X, a)$  with  $X \in \mathbf{X}_a$  along with edges asserting that  $X^a = X$ , we obtain an equivalent PDG that does not look so different:



Indeed, it can be shown that (7.1,7.2) minimize inconsistency of the dotted components in their appropriate contexts (shown in green and blue above).

This means that LIR with the appropriate views computes belief propagation. To show this formally, we need a few definitions. Equation (7.1) adjusts the parameters of  $C_{X \rightarrow a} := \{m_{X \rightarrow a}\}$  so as to minimize 1-inconsistency in context  $A_{X \rightarrow a} := \{m_{b \rightarrow X}\}_{b \in \partial X \setminus a} \cup \{m_{X \rightarrow a}\}$ , while (7.2) adjusts  $C_{a \rightarrow X} := \{m_{a \rightarrow X}\}$  so as to minimize the 1-inconsistency in context  $A_{a \rightarrow X} := \{\phi_a, m_{a \rightarrow X}\} \cup \{m_{Y \rightarrow a}\}_{Y \in \mathbf{X}_a \setminus X}$ . The only wrinkle is that we do not want to attend to the structural aspect of a message  $e$  that we are updating—that is, we must select  $\varphi$  so as to ignore its causal weight  $\alpha_e$ . Intuitively: when we are updating some message  $e$ , we are interested in summarizing information in the other messages (both observational and causal information), purely with an observation.

More precisely, the foci

$$\mathbf{F} := \left\{ (\varphi_j, \chi_j) : j \in \bigcup_{\substack{a \in \mathcal{A} \\ X \in \mathbf{X}_a}} \left\{ a \rightarrow X, X \rightarrow a, X \right\}, \right\}$$

are indexed by messages and variables, and defined as follows. The attention mask  $\varphi_j$  is given by:

$$\varphi_j(a) := \begin{cases} \begin{pmatrix} 1 \\ 1 \end{pmatrix} & \text{if } a \in A_j \setminus C_j \\ \begin{pmatrix} 1 \\ 0 \end{pmatrix} & \text{if } a \in C_j \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix} & \text{otherwise} \end{cases},$$

where  $\begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}$  scales  $\beta_a$  by  $\phi_1$  and  $\alpha_a$  by  $\phi_2$ . Finally, full control over  $C_j$  means defining  $\chi_j(a) := \infty \cdot \mathbb{1}[a \in C_j]$ . With these choices in place, it's not hard to show that [Algorithm 1](#) amounts to belief propagation.

**Proposition 7.3.** *If REFOCUS selects a focus non-deterministically from  $\{a \rightarrow X, X \rightarrow a, X\}_{X \in \mathcal{X}, a \in \partial X}$ , then the possible runs of LIR( $\mathbf{m}_\Phi, \mathbf{msg} + \mathcal{B}$ ) are precisely those of BP for different message schedules.*

There are many established variants of this algorithm. Some of them are generated different by clustering factors together—in the language of [Koller and Friedman \(2009\)](#), that is to say choosing something other than the Bethe cluster graph as the basis for message passing. Our analysis immediately applies to these other cluster graphs.

[Minka \(2005\)](#) offers a different perspective, in which a broader class of message passing algorithms can be viewed as iteratively adjusting some local context to minimize an  $\alpha$ -divergence. We suspect that LIR generalizes these procedure as well—not only because it is similar in spirit, but also because these divergences can be viewed as the degree of inconsistency of a PDG containing two distributions ([Section 6.4](#)).

## 7.8 Discussion and Future Work

These examples are only the beginning. Our initial investigations suggest that opinion dynamics models, the training process for diffusion models, and much more, are all naturally captured by LIR. The surprising generality of LIR begs some theoretical questions. What assumptions are needed to prove that it reduces overall inconsistency, as is often the case? What are the simplest choices we could make to produce an efficient non-standard algorithm? How expressive is this mode of computation?

It also suggests a novel approach to structured generative modeling: haphazardly assemble a PDG with many variables, existing models, priors, constraints, and data of all shapes and sizes. Then, train new models to predict variables from one another, using LIR (with random refocusing, say). Is this effective? We are excited to find out!

## APPENDICES FOR CHAPTER 7

### 7.A Proofs

First, some extra details for [Proposition 7.1](#). By parameteriations  $\mathbb{P}$  log-concave, we mean that, for every  $a \in \mathcal{A}$ , and  $(s, t) \in \mathcal{V}(S_a, T_a)$ , the function

$$\theta \mapsto -\log \mathbb{P}_a^\theta(T_a = t \mid S_a = a) : \Theta_a \rightarrow [0, \infty]$$

is convex. This is true for many families of distributions of interest. For example, if  $S_a, T_a$  is discrete, and the cpd is parameterized by stochastic matrices  $\mathbf{P} = [p_{s,t}] \in [0, 1]^{\mathcal{V}(S_a, T_a)}$ , then

$$-\log \mathbb{P}_a^\mathbf{P}(T_a = t \mid S_a = s) = -\log(p_{s,t})$$

which is clearly convex in  $\mathbf{P}$ .

To take another example: if  $\mathbb{P}_a$  is linear Gaussian, i.e.,  $\mathbb{P}_a(T|S) = \mathcal{N}(T|\mathbf{A}s + b, \sigma^2)$ , parameterized by  $(\mathbf{A}, b, 1/\sigma^2)$ , then

$$-\log \mathbb{P}_a^{(\mathbf{A}, b, \sigma^2)}(t|s) = -\frac{1}{2} \log \frac{2\pi}{\sigma^2} + \frac{1}{2} \left( \frac{t - \mathbf{A}s + b}{\sigma} \right)^2$$

which is convex in  $(\mathbf{A}, b, \frac{1}{\sigma^2})$ . Now, for the proof.

**Proposition 7.1.** *If  $\mathbf{m}(\Theta)$  is a PPDG whose parameterizations  $\mathbb{P}$  are either constant or unconditional and log-concave, then for small enough  $\gamma$ , the map  $\theta \mapsto \langle\!\langle \varphi \odot (\mathcal{C}tx + \mathbf{m}(\theta)) \rangle\!\rangle_\gamma$  is convex.*

*Proof.* By definition,

$$\langle\!\langle \varphi \odot (\mathcal{C}tx + \mathbf{m}(\theta)) \rangle\!\rangle_\gamma = \inf_\mu \left\{ OInc_{\mathcal{C}tx}(\mu) + \gamma SDef_{\mathcal{C}tx}(\mu) + \gamma SDef_{\mathbf{m}(\theta)}(\mu) + OInc_{\mathbf{m}(\theta)}(\mu) \right\}.$$

Only the final term actually depends on  $\theta$ , though—recall that  $SDef_{m(\theta)}$  depends only on the structure of the hypergraph (and the weights  $\alpha$ ), and not the parameters of the cpds. Thus, we can write  $F(\mu)$  for the first three terms.

For all of our examples, and indeed, if  $\gamma$  is chosen small enough, we have seen that the sum of the two terms is convex in  $\mu$ . Then we have

$$\begin{aligned} & \left\langle \varphi \odot (\mathcal{C}tx + m(\theta)) \right\rangle_\gamma \\ &= \inf_\mu \left( F(\mu) + \mathbb{E}_\mu \left[ \sum_{S \xrightarrow{a} T} \beta_a \log \frac{\mu(T|S)}{\mathbb{P}_a^\theta(T|S)} \right] \right) \\ &= \inf_\mu \left( F(\mu) + \mathbb{E}_\mu \left[ \sum_{S \xrightarrow{a} T} \beta_a \log \frac{\mu(T|S)}{\lambda(T|S)} \right] + \underbrace{\mathbb{E}_\mu \left[ \sum_{S \xrightarrow{a} T} \beta_a \log \frac{\lambda(T|S)}{\mathbb{P}_a^\theta(T|S)} \right]}_{\text{third term}} \right) \end{aligned}$$

The second term is then entropy (relative to the base distribution), which is convex in  $\mu$ . The first term,  $F(\mu)$ , is convex in  $\mu$  as well, and neither depend on  $\theta$ . The final term is linear in  $\mu$ . Since  $\mathbb{P}$  is log-convex in  $\theta$ , we know that  $(\log \frac{\lambda(t|s)}{\mathbb{P}_a^\theta(t|s)})$  is convex in  $\theta$ . It follows that the third term is a positive linear combination of expectations that are all convex in  $\theta$ , and hence itself convex in  $\theta$ . Because the first two terms do not depend on  $\theta$  and are convex in  $\mu$ , they are jointly convex in  $(\mu, \theta)$ . And, as we have seen, the third term is linear in  $\mu$  and convex in  $\theta$ , so it is also jointly convex in  $(\mu, \theta)$ . Thus, the sum of all three terms in the infimum is jointly convex in  $(\theta, \mu)$ . Taking an infimum over  $\mu$  pointwise, the result is still convex in  $\theta$  (Boyd and Vandenberghe 2004).  $\square$

**Proposition 7.3.** *If REFOCUS selects a view non-deterministically from  $\{a \rightarrow X, X \rightarrow a, X\}_{X \in \mathcal{X}, a \in \partial X}$  with  $\varphi, \chi$  as above, and  $\gamma = 1$ , then the possible runs of LIR( $m_\Phi, m_{\mathcal{M}} + \mathcal{B}$ ) are precisely those of BP for different message schedules.*

*Proof.* When  $\gamma = 1$ , and  $\alpha, \beta = 1$  for all of the input factors, then the optimal

distribution  $\mu^*$  that realizes the infimum is just the product of factors. It follows that any distribution that has those marginals will minimize the observational inconsistency.

The different orders that the (7.1), and (7.2) can be ordered for different adjacent pairs  $(a, X)$  correspond to both the message passing schedules, and to the possible view selections of LIR.  $\square$

## **Part III**

# **Algorithms, Logic, and Complexity**

## CHAPTER 8

### INFERENCE FOR PDGS, VIA EXPONENTIAL CONIC PROGRAMMING

So far, we have seen that PDGs are a flexible modeling tool that plays two seemingly very different roles. On one hand, PDGs specify a joint distribution; in this capacity, they are an especially modular and interpretable generalization of standard probabilistic graphical models. On the other hand, PDGs have a natural inconsistency measure; in this capacity, they are a principled model-based way to specify a loss function. Yet our discussion so far leaves a burning question: how can we compute with PDGs? Is it even possible? So far, we have seen no reason that there should be any practical way to compute with PDGs in either role: no inference algorithm, and no (provably correct) way to calculate the degree of inconsistency. In this chapter, we solve both problems.

#### 8.1 Introduction

As we have seen, PDGs (Definition 3.1) form a very general class of probabilistic graphical models, that includes not only Bayesian Networks (BNs) and Factor Graphs (FGs), but also more recent statistical models built out of neural networks, such as Variational AutoEncoders (VAEs) (Kingma and Welling 2014). PDGs can also capture inconsistent beliefs, and provide a useful way to measure the degree of this inconsistency; for a VAE, this is the loss function used in training (Section 6.5.1). But up to now, there has been no practical way to do inference for PDGs—that is, to answer questions of the form “what is the probability of  $Y$  given  $X$ ?”. This chapter presents the first algorithm to do so.

Before discussing our algorithm, we must discuss what it even means to do inference for a PDG. A BN or FG represents a unique joint distribution. Thus,

for example, when we ask “what is the probability of  $Y$  given that  $X=x$ ?” in a BN, we mean “what is  $\mu(Y|X=x)$ ?” for the probability measure  $\mu$  that the BN represents. But a PDG might, in general, represent more than just one distribution.

Like a BN, a PDG encodes two types of information: “structural” information about the independence of causal mechanisms, and “observational” information about conditional probabilities. Unlike in a BN, the two can conflict in a PDG. Corresponding to these two types of information, a PDG has two loss functions, which quantify how far a distribution  $\mu$  is from modeling the information of each type. Given a number  $\hat{\gamma} \in [0, 1]$  indicating the importance of structure relative to observation, the  $\hat{\gamma}$ -semantics of a PDG is the set of distributions that minimize the appropriate convex combination of losses. We also consider the  $0^+$ -semantics: the limiting case that arises as  $\hat{\gamma}$  goes to zero (which focuses on observation, using structure only to break ties). This set can be shown to contain precisely one distribution for PDGs satisfying a mild regularity condition, which we call *proper*. Thus, we have a parameterized family of inference notions: to do  $\hat{\gamma}$ -inference, for  $\hat{\gamma} \in [0, 1] \cup \{0^+\}$ , is to answer queries in a way that is true of all distributions in the  $\hat{\gamma}$ -semantics.

If there are distributions fully consistent with both the observational and the structural information in a PDG  $\mathcal{M}$ , then for  $\hat{\gamma} \in (0, 1) \cup \{0^+\}$ , all notions of  $\hat{\gamma}$ -inference coincide. If  $\mathcal{M}$  is also proper, this means there is a single distribution  $\mu_{\mathcal{M}}$  that minimizes both loss functions, in which case we want to answer queries with respect to  $\mu_{\mathcal{M}}$  no matter how we weight observational and structural information. Moreover, if  $\mathcal{M}$  represents a BN, then  $\mu_{\mathcal{M}}$  is the distribution represented by the BN. However, if there is no distribution that is consistent with both types of information, then the choice of  $\hat{\gamma}$  matters.

Since PDGs subsume BNs, and inference for BNs is already NP-hard, the same must be true of PDGs. At a high level, the best we could hope for would be tractability on the restricted class of models on which inference has traditionally been tractable—that is, a polynomial algorithm for models whose underlying structure has *bounded treewidth* (see [Section 8.2](#) for formal definitions). That is indeed what we have. More precisely, we show that  $0^+$ -inference and  $\hat{\gamma}$ -inference for small  $\hat{\gamma}$  can be done for discrete PDGs of bounded treewidth containing  $N$  variables in  $\tilde{O}(N^4)$  time.

Our algorithm is based on a line of recent work in convex programming that establishes polynomial-time for a class of optimization problems called *exponential conic programs* ([Badenbroek and Dahl 2021](#); [Skajaa and Ye 2015](#); [Nesterov et al. 1999](#)). Our contribution is to show that the problem of inference in a PDG of bounded treewidth can be efficiently converted to a (sequence of) exponential conic program(s), at which point it can be solved with a commercial solver (e.g., [ApS \(2022\)](#)) in polynomial time. The direct appeal to a solver allows us to benefit from the speed and reliability of such highly optimized solvers, and also from future improvements in exponential conic optimization. Thus, our result is not only a theoretical one, but practical as well.

Beyond its role as a probabilistic model, a PDG is also of interest for its degree of inconsistency—that is, the minimum value of its loss function. As shown in [Chapter 6](#) many loss functions and statistical divergences can be viewed as measuring the inconsistency of a PDG that models the context appropriately. This makes calculating this minimum value of interest—but up to now, there has been no way to do so. There is a deep connection between this problem and PDG inference (which we develop in [Chapter 9](#)); for now, we remark that this number is a byproduct of our techniques.

**Chapter Contributions.** We provide the first algorithm for inference in a PDG; in addition, it calculates a PDG’s degree of inconsistency. We prove that our algorithm is correct, and also fixed-parameter tractable: for PDGs of bounded treewidth, it runs in polynomial time. We also prove that PDG inference and inconsistency calculation are equivalent problems. Our algorithm reduces inference in PDGs to exponential conic programming in a way that can be offloaded to powerful existing solvers. We provide an implementation of this reduction in a standard convex optimization framework, giving users an interface between such solvers and the standard PDG Python library. Finally, we evaluate our implementation. The results suggest our method is faster and significantly more reliable than simple baseline approaches.

## 8.2 Preliminaries & Related Work

**Treewidth.** Recall that an undirected hypergraph  $(V, \mathcal{E})$  is a set  $V$  of vertices and a set  $\mathcal{E}$  of subsets of  $V$ . Thus, a directed hypergraph  $(N, \{S_a \xrightarrow{a} T_a\}_{a \in \mathcal{A}})$  can be viewed as a hypergraph by joining the source and target set of each hyperarc (i.e., taking  $\mathcal{E} = \{S_a \cup T_a : a \in \mathcal{A}\}$ ), thereby “forgetting” the direction of the arrow. Thus, notions defined for undirected hypergraphs (like that of treewidth, which we now review), can be applied to directed hypergraphs as well.

Many problems that are intractable for general graphs are tractable for trees, and some graphs are closer to being trees than others. A tree decomposition of a (hyper)graph  $G = (V, \mathcal{E})$  is a tree  $(\mathcal{C}, \mathcal{T})$  whose vertices  $C \in \mathcal{C}$ , called *clusters*, are subsets of  $V$  such that:

1. every vertex  $v \in V$  and every hyperedge  $E \in \mathcal{E}$  is contained in at least one

cluster, and

2. every cluster  $D$  along the unique path from  $C_1$  to  $C_2$  in  $\mathcal{T}$ , contains  $C_1 \cap C_2$ .

The *width* of a tree decomposition is one less than the size of its largest cluster, and the *treewidth* of a (hyper)graph  $G$  is the smallest possible width of any tree decomposition of  $G$ . It is NP-hard to determine the tree-width of a graph, but if the tree-width is known to be bounded above, a tree decomposition may be constructed in linear time (Bodlaender 1993). For graphs of bounded tree-width, many problems (indeed, any problem expressible in a certain second-order logic (Courcelle 1990)) can be solved in linear time. This is also true of inference in standard graphical models.

**Inference for Traditional Graphical Models.** Semantically, a traditional graphical model  $\mathcal{M}$  (such as a Bayesian Network or a factor graph) typically represents a joint probability distribution  $\Pr_{\mathcal{M}} \in \Delta^{\mathcal{V}\mathcal{X}}$  over its variables. Inference for  $\mathcal{M}$  is then the ability to calculate conditional probabilities  $\Pr_{\mathcal{M}}(Y|X=x)$ , where  $X, Y \subset \mathcal{X}$  and  $x \in \mathcal{V}X$ .

There are several different approaches to inference in graphical models. Many of them (such as belief propagation; see Chapter 7), when applied to tree-like graphical models, run in linear time and are provably correct. If the same algorithms are naïvely applied to graphs with cycles (as in loopy belief propagation), then they may not converge, and even if they do, may give incorrect (or even inconsistent) answers (Wainwright et al. 2008). Nearly all *exact* inference algorithms (including variable elimination (Bertele and Brioschi 1972), message-passing with (Lauritzen and Spiegelhalter 1988) and without division (Shafer and Shenoy 1990), among others (Wainwright et al. 2003)) effectively construct a tree decomposition, and can be viewed as running on a tree (Koller and Friedman

[2009](#), §9-11). Indeed, under widely believed assumptions, every class of graphical models for which (exact) inference is *not* NP-hard has bounded treewidth ([Chandrasekaran et al. 2012](#)).

Given a tree decomposition  $(\mathcal{C}, \mathcal{T})$  of the underlying model structure, many of these algorithms use a standard data structure that we will call a *tree marginal*, which is a collection  $\mu = \{\mu_C(C)\}_{C \in \mathcal{C}}$  of probabilities over the clusters ([Koller and Friedman 2009](#), §10). A tree marginal  $\mu$  is said to be *calibrated* if neighboring clusters' distributions agree on the variables they share. In this case,  $\mu$  determines a joint distribution

$$\Pr_{\mu}(\mathcal{X}) = \prod_{C \in \mathcal{C}} \mu_C(C) / \prod_{(C \cap D) \in \mathcal{T}} \mu_C(C \cap D), \quad (8.1)$$

which has the property that  $\Pr_{\mu}(C) = \mu_C$  for all  $C \in \mathcal{C}$ . A calibrated tree marginal summarizes the answers to many queries about  $\Pr_{\mu}$  (see [Koller and Friedman 2009](#), §10.3.3).<sup>1</sup> Therefore, to answer probabilistic queries with respect to a distribution  $\mu$ , it suffices to find a calibrated tree marginal  $\mu$  that represents  $\mu$ , and appeal to standard algorithms.

For the reader's convenience, we now repeat a (more compact) version of the definition of a PDG. A PDG  $m = (\mathcal{X}, \mathcal{A}, \mathbb{P}, \alpha, \beta)$  is a directed hypergraph  $(\mathcal{X}, \mathcal{A})$  whose nodes are variables, together with probabilities  $\mathbb{P}$  and confidence vectors  $\alpha = [\alpha_a]_{a \in \mathcal{A}}$  and  $\beta = [\beta_a]_{a \in \mathcal{A}}$ , so that each  $S \xrightarrow{a} T \in \mathcal{A}$  is associated with:

- a conditional probability distribution  $\mathbb{P}_a(T|S)$  on the target variables given values of the source variables,

---

<sup>1</sup>To see why in a simple case, note that for an unconditional query about  $Y$  contained within a single cluster  $C$ , we have  $\Pr_{\mu}(Y) = \mu_C(Y)$ . With some care, the general idea can be extended to arbitrary queries (see [Koller and Friedman 2009](#), §10.3.3); those conditional on evidence  $X=x$  can be handled by conditioning the clusters that contain  $X$ , and then recalibrating  $\mu$  with a standard algorithm like belief propagation.

- a weight  $\beta_a \in \bar{\mathbb{R}}$  indicating the modeler's confidence in the cpd  $\mathbb{P}_a(T|S)$ , and
- a weight  $\alpha_a \in \mathbb{R}$  indicating the modeler's confidence in the functional dependence of  $T$  on  $S$  expressed by  $a$ .

If  $\beta \geq 0$  and  $\alpha_a > 0$  implies  $\beta_a > 0$ , we write  $\beta \gg \alpha$  and call  $\mathcal{M}$  *proper*. Note that  $\beta \gg \alpha$  if  $\beta > 0$ . Recall that PDG contains two types of information: “structural” information, in the hypergraph  $\mathcal{A}$  and weights  $\alpha$ , and “observational” data, in the cpds  $\mathbb{P}$  and weights  $\beta$ . PDG semantics are based on two scoring functions that quantify discrepancy between each type of information and a distribution  $\mu \in \Delta \mathcal{VX}$  over its variables. (See [Section 3.3.2](#) for details.) Given a value of  $\gamma > 0$ , a trade-off parameter that controls the strength of the structural information, recall that the overall scoring function is given by

$$\begin{aligned} \llbracket \mathcal{M} \rrbracket_\gamma(\mu) &:= OInc_{\mathcal{M}}(\mu) + \gamma SDef_{\mathcal{M}}(\mu) \\ &= \mathbb{E}_\mu \left[ \sum_{S \xrightarrow{a} T \in \mathcal{A}} \log \frac{\mu(T|S)^{\beta_a - \gamma \alpha_a}}{\mathbb{P}_a(T|S)^{\beta_a}} \right] - \gamma H(\mu). \end{aligned} \tag{3.4}$$

Recall that  $\llbracket \mathcal{M} \rrbracket_\gamma^* = \arg \min_\mu \llbracket \mathcal{M} \rrbracket_\gamma(\mu)$  denote the set of optimal distributions at a particular value  $\gamma$ . One natural conception of inference in PDGs is then parameterized by  $\hat{\gamma}$ : to do  $\hat{\gamma}$ -inference in  $\mathcal{M}$  is to respond to probabilistic queries in a way that is sound with respect to every  $\mu \in \llbracket \mathcal{M} \rrbracket_\gamma^*$ . It is not too difficult to see that when  $\beta \geq \gamma \alpha$ , (3.4) is strictly convex, which ensures that  $\llbracket \mathcal{M} \rrbracket_\gamma^*$  is a singleton. This chapter demonstrates that  $\hat{\gamma}$ -inference is tractable under this condition.

The limiting behavior of the  $\hat{\gamma}$ -semantics as  $\hat{\gamma} \rightarrow 0$ , which we denote  $\llbracket \mathcal{M} \rrbracket_{0+}^*$  and call the  *$0^+$ -semantics*, has some special properties (e.g., [Theorem 3.3](#)). This distribution intuitively reflects an extreme empirical view: observational data trumps causal structure. One should be careful to distinguish  $\llbracket \mathcal{M} \rrbracket_{0+}^*$  from  $\llbracket \mathcal{M} \rrbracket_0^*$ , the set of distributions that minimize  $OInc_{\mathcal{M}}$ ; the latter set includes  $\llbracket \mathcal{M} \rrbracket_{0+}^*$  ([Proposition 3.4](#)) but may also contain other distributions. This chapter also shows how

to efficiently answer queries with respect to the unique distribution in  $\llbracket m \rrbracket_{0^+}^*$ , which we call  $0^+$ -*inference*.

**Interior-Point Methods and Convex Optimization.** Interior-point methods provide an iterative way of approximately solving linear programs in polynomial time (Karmarkar 1984). With the theory of “symmetric cones”, these methods were extended in the 1990s to handle second-order cone programs (SOCPs) and semidefinite programs (SDPs), which allow more expressive constraints. But the constraints that these methods can handle are insufficient for our purposes. We need what have been called *exponential cone constraints*. The *exponential cone* is the convex set

$$\begin{aligned} K_{\text{exp}} := & \left\{ (x_1, x_2, x_3) : x_1 \geq x_2 e^{x_3/x_2}, x_2 > 0 \right\} \\ & \cup \left\{ (x_1, 0, x_3) : x_1 \geq 0, x_3 \leq 0 \right\} \subset \bar{\mathbb{R}}^3. \end{aligned}$$

It is also sometimes called the “relative entropy cone”, for reasons we will see in Section 8.3.1. Suppose  $K = K_{\text{exp}}^p \times [0, \infty]^q \subset \bar{\mathbb{R}}^n$  is a product of  $p$  exponential cones and  $q = n - 3k$  non-negative orthants. An *exponential conic program* is then an optimization problem of the form

$$\underset{\mathbf{x}}{\text{minimize}} \quad \mathbf{c}^\top \mathbf{x} \quad \text{subject to } A\mathbf{x} = \mathbf{b}, \quad \mathbf{x} \in K, \tag{8.2}$$

where  $\mathbf{c} \in \bar{\mathbb{R}}^n$  is some cost vector, the function  $\mathbf{x} \mapsto \mathbf{c}^\top \mathbf{x}$  is called the *objective*, and  $\mathbf{b} \in \bar{\mathbb{R}}^m$ ,  $A \in \bar{\mathbb{R}}^{m \times n}$  encode linear constraints. Nesterov, Todd, and Ye (1999) first established that such problems can be solved in polynomial time, but incur double the memory and eight times the time, compared to the symmetric counterparts. These drawbacks were eliminated in Skajaa and Ye (2015). The algorithm that seems to display the best empirical performance (Dahl and Andersen 2022), however, was only recently shown to run in polynomial time (Badenbroek and Dahl 2021).

Disciplined Convex Programming ([Grant 2004](#)) is a compositional approach to convex optimization that imposes certain restrictions on how problems can be specified. A problem conforming to those rules is said to be *dcp*, and can be efficiently compiled to a standard form ([Agrawal et al. 2018](#)), which in our case is an exponential conic program. Only two rules are relevant to us: a constraint of the form  $(x, y, z) \in K_{\text{exp}}$  is dcp iff  $x, y$ , and  $z$  are affine transformations of the optimization variables, and a linear program augmented with dcp constraints is dcp. Because all the optimization problems that we give are of this form, we can easily compile them to exponential conic programs even if they do not exactly conform to [\(8.2\)](#).

### 8.3 Inference as a Convex Program

Here is an obvious, if inefficient, way of calculating  $\Pr_{\mathcal{M}}(Y|X=x)$  in a probabilistic model  $\mathcal{M}$ . First compute an explicit representation of the joint distribution  $\Pr_{\mathcal{M}} \in \Delta \mathcal{V}\mathcal{X}$ , then marginalize to  $\Pr_{\mathcal{M}}(X, Y)$  and condition on  $X=x$ . For a factor graph or BN, each step is straightforward; the problem is the exponential time and space required to represent  $\Pr_{\mathcal{M}}(\mathcal{X})$  explicitly. A key feature of inference algorithms for BNs and FGs is that they do not represent joint distributions in this way. For PDGs, though, it is not obvious that we can calculate the  $\hat{\gamma}$ -semantics, even if we know it is unique, and we ignore the space required to represent it (as we do in this section). Note that  $\hat{\gamma}$ -inference is already an optimization problem by definition:

$$\underset{\mu}{\text{minimize}} \quad [\![\boldsymbol{m}]\!]_{\gamma}(\mu) \quad \text{subject to} \quad \mu \in \Delta \mathcal{V}\mathcal{X}.$$

For small enough  $\gamma$ , it is even convex. But can we solve it efficiently? With exponential cone constraints, the answer is yes, as we show in [Section 8.3.2](#).

Moreover, we can compute the  $0^+$ -semantics with a sequence of two exponential conic programs (Section 8.3.3). To give a flavor of our constructions and ease into the more complicated ones, we begin by minimizing  $OInc$ , the simpler of the two scoring functions.

### 8.3.1 Minimizing Incompatibility ( $\gamma = 0$ )

When  $\gamma = 0$ , we want to find minimizers of  $OInc$ , which is a weighted sum of conditional relative entropies. There is a straightforward connection between the exponential cone and relative entropy: if  $\mathbf{m}, \mathbf{p} \in \Delta\{1, \dots, n\} \subset \mathbb{R}^n$  are points on a probability simplex, then  $(-\mathbf{u}, \mathbf{m}, \mathbf{p}) \in K_{\exp}^n$  if and only if  $\mathbf{u}$  is an upper bound on  $\mathbf{m} \log \frac{\mathbf{m}}{\mathbf{p}}$ , the pointwise contribution to relative entropy at each outcome. Thus, perhaps unsurprisingly, we can use an exponential conic program to find minimizers of  $OInc$ . If all beliefs are unconditional and over the same space, the construction is standard; we review it here, so that we can build upon it.

**Warm-up.** Consider a PDG with only one variable  $X$  with  $\mathcal{V}X = \{1, \dots, n\}$ . Suppose further that every arc  $j \in \mathcal{A} = \{1, \dots, k\}$  has  $T_j = \{X\}$  and  $S_j = \emptyset$ . Then each  $\mathbb{P}_j(X)$  can be identified with a vector  $\mathbf{p}_j \in [0, 1]^n$ , and all  $k$  of them can conjoined to form a matrix  $\mathbf{P} = [p_{ij}] \in [0, 1]^{n \times k}$ . Similarly, a candidate distribution  $\mu$  can be identified with  $\mathbf{m} \in [0, 1]^n$ . Now consider a matrix  $\mathbf{U} = [u_{i,j}] \in \bar{\mathbb{R}}^{n \times k}$  that, intuitively, gives an upper bound on the contribution to  $OInc$

due to each edge and value of  $X$ . Observe that

$$\begin{aligned}
& (-\mathbf{U}, [\mathbf{m}, \dots, \mathbf{m}], \mathbf{P}) \in K_{\exp}^{n \times k} \\
\iff & \forall i, j. \ u_{ij} \geq m_i \log(m_i/p_{ij}) \\
\implies & \forall j. \sum_i u_{ij} \geq D(\mu \| p_j) \\
\implies & \sum_{i,j} \beta_j u_{ij} \geq \sum_j \beta_j D(\mu \| p_j) \\
\iff & \mathbf{1}^\top \mathbf{U} \boldsymbol{\beta} \geq OInc(\mu).
\end{aligned} \tag{8.3}$$

So now, if  $(\mathbf{U}, \mathbf{m})$  is a solution to the convex program

$$\begin{aligned}
\underset{\mathbf{m}, \mathbf{U}}{\text{minimize}} \quad & \mathbf{1}^\top \mathbf{U} \boldsymbol{\beta} \quad \text{subject to} \quad \mathbf{1}^\top \mathbf{m} = 1, \\
& (-\mathbf{U}, [\mathbf{m}, \dots, \mathbf{m}], \mathbf{P}) \in K_{\exp}^{n \times k},
\end{aligned}$$

then (a) the objective value  $\mathbf{1}^\top \mathbf{U} \boldsymbol{\beta}$  equals the inconsistency  $\langle\!\langle \mathbf{m} \rangle\!\rangle_0$ , and (b)  $\mu \in \llbracket \mathbf{m} \rrbracket_0^*$ , meaning  $\mu$  minimizes  $OInc_{\mathbf{m}}$ .

**The General Case.** We now show how the same construction can be used to find a distribution  $\mu \in \llbracket \mathbf{m} \rrbracket_0^*$  for an arbitrary PDG  $\mathbf{m} = (\mathcal{X}, \mathcal{A}, \mathcal{P}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ . To further simplify the presentation, for each arc  $a \in \mathcal{A}$ , let  $\mathcal{V}a := \mathcal{V}(S_a, T_a)$  denote all joint settings of  $a$ 's source and target variables, and write  $\mathcal{V}\mathcal{A} := \sqcup_{a \in \mathcal{A}} \mathcal{V}a = \{(a, s, t) : a \in \mathcal{A}, (s, t) \in \mathcal{V}(S_a, T_a)\}$  for the set of all choices of an arc together with values of its source and target. For each  $a \in \mathcal{A}$ , we can regard  $\mu(T_a, S_a)$  and  $\mu(S_a) \mathbb{P}_a(T_a | S_a)$ , both distributions over  $\{S_a, T_a\}$ , as vectors of shape  $\mathcal{V}a$ . As before, we introduce an optimization variable  $\mathbf{u}$  that packages together all of the relevant pointwise upper bounds. To that end, consider a vector  $\mathbf{u} = [u_{a,s,t}] \in \bar{\mathbb{R}}^{\mathcal{V}\mathcal{A}}$  in the

optimization problem

$$\begin{aligned} & \underset{\mu, \mathbf{u}}{\text{minimize}} \quad \sum_{(a,s,t) \in \mathcal{V}\mathcal{A}} \beta_a u_{a,s,t} \\ & \text{subject to} \quad \mu \in \Delta \mathcal{V}\mathcal{X}, \\ & \forall a \in \mathcal{A}. \left( -\mathbf{u}_a, \mu(T_a, S_a), \mathbb{P}_a(T_a | S_a) \mu(S_a) \right) \in K_{\text{exp}}^{\mathcal{V}a}, \end{aligned} \tag{8.4}$$

where  $\mathbf{u}_a = [u_{a,s,t}]_{(s,t) \in \mathcal{V}a}$  consists of those components of  $\mathbf{u}$  associated with arc  $a$ . Note that the marginals  $\mu(S_a, T_a)$  and  $\mu(S_a)$  are affine transformations of  $\mu$ , so (8.4) is dcp. A straightforward generalization of the logic in (8.3) gives us:

**Proposition 8.1.** *If  $(\mu, \mathbf{u})$  is a solution to (8.4), then  $\mu \in [\![\mathbf{m}]\!]_0^*$ , and* [link to proof]

$$\sum_{(a,s,t) \in \mathcal{V}\mathcal{A}} \beta_a u_{a,s,t} = \langle\!\langle \mathbf{m} \rangle\!\rangle_0.$$

Thus, a solution to (8.4) encodes a distribution that minimizes  $OInc$ , and the (0-)inconsistency of  $\mathbf{m}$ . This is a start, but to do  $0^+$ -inference, among the minimizers of  $OInc$  we must find the unique distribution in  $[\![\mathbf{m}]\!]_{0+}^*$ , while for  $\hat{\gamma}$ -inference ( $\hat{\gamma} > 0$ ), we need to find the optimizers of  $[\![\mathbf{m}]\!]_{\hat{\gamma}}^*$ . Either way, we must consider  $SDef$  in addition to  $OInc$ .

### 8.3.2 $\gamma$ -Inference for small $\gamma > 0$

When  $\gamma > 0$  is small enough, the scoring function (3.4) is not only convex, but admits a straightforward representation as an exponential conic program. To see this, note that (3.4) can be rewritten Proposition 3.9 as:

$$\begin{aligned} [\![\mathbf{m}]\!]_{\gamma}(\mu) &= -\gamma H(\mu) - \sum_{a \in \mathcal{A}} \beta_a \mathbb{E}_{\mu} \log \mathbb{P}_a(T_a | S_a) \\ &\quad + \sum_{a \in \mathcal{A}} (\gamma \alpha_a - \beta_a) H_{\mu}(T_a | S_a). \end{aligned} \tag{8.5}$$

The first term,  $-\gamma H(\mu)$ , is strictly convex and has a well-known translation into an exponential cone constraint; the second one linear in  $\mu$ . If  $0 < \gamma \leq \min_a \frac{\beta_a}{\alpha_a}$ , then every summand of the last term is a negative conditional entropy, and can be captured by an exponential cone constraint. The only wrinkle is that it is possible for a user to specify that some  $\mathbb{P}_a(t | s) = 0$ , in which case the linear term is undefined. The result is a requirement that  $\mu(s, t) = 0$  at such points, which we can instead encode directly with linear constraints. To do this formally, divide  $\mathcal{VA}$  into two parts:  $\mathcal{VA}^+ := \{(a, s, t) \in \mathcal{VA} : \mathbb{P}_a(t|s) > 0\}$  and  $\mathcal{VA}^0 := \{(a, s, t) \in \mathcal{VA} : \mathbb{P}_a(t|s) = 0\}$ . Armed with this notation, consider upper bound vectors  $\mathbf{u} = [u_{a,s,t}]_{(a,s,t) \in \mathcal{VA}}$  and  $\mathbf{v} = [v_w]_{w \in \mathcal{VX}}$ , in the following optimization problem:

$$\begin{aligned} \underset{\mu, \mathbf{u}, \mathbf{v}}{\text{minimize}} \quad & \sum_{(a,s,t) \in \mathcal{VA}} (\beta_a - \alpha_a \gamma) u_{a,s,t} + \gamma \sum_{w \in \mathcal{VX}} v_w \\ & - \sum_{(a,s,t) \in \mathcal{VA}^+} \alpha_a \gamma \mu(S_a=s, T_a=t) \log \mathbb{P}_a(t|s) \\ \text{subject to} \quad & \mu \in \Delta \mathcal{VX}, \quad (-\mathbf{v}, \mu, \mathbf{1}) \in K_{\text{exp}}^{\mathcal{VX}}, \\ & \forall a \in \mathcal{A}. \left( -\mathbf{u}_a, \mu(T_a, S_a), \mathbb{P}_a(T_a|S_a) \mu(S_a) \right) \in K_{\text{exp}}^{\mathcal{VX}}, \\ & \forall (a, s, t) \in \mathcal{VA}^0. \mu(S_a=s, T_a=t) = 0. \end{aligned} \tag{8.6}$$

This optimization problem may look complex, but it falls out of (8.5) fairly directly, and gives us what we wanted.

**Proposition 8.2.** *If  $(\mu, \mathbf{u}, \mathbf{v})$  is a solution to (8.6), and  $\beta \geq \gamma \alpha$ , then  $\mu$  is the unique element of  $[\![m]\!]_{\gamma}^*$ , and  $\langle\!\langle m \rangle\!\rangle_{\gamma}$  equals the objective of (8.6) evaluated at  $(\mu, \mathbf{u}, \mathbf{v})$ .*

[ link to proof ]

### 8.3.3 Calculating the $0^+$ -semantics ( $\gamma \rightarrow 0$ )

Section 8.3.1 shows how to find a distribution  $\nu$  that minimizes  $OInc$ —but to do  $0^+$ -inference, we need to find the minimizer that, uniquely among them, best minimizes  $SDef$ . It turns out this can be done by using  $\nu$  to construct a second optimization problem. The justification requires two more results; we start by characterizing the minimizers of  $OInc$ .

**Proposition 8.3.** *If  $\mathbf{m}$  has arcs  $\mathcal{A}$  and  $\beta \geq 0$ , the minimizers of  $OInc_m$  all have the same conditional marginals along  $\mathcal{A}$ . That is, for all  $\mu_1, \mu_2 \in [\![\mathbf{m}]\!]_0^*$  and all  $S \xrightarrow{a} T \in \mathcal{A}$  with  $\beta_a > 0$ , we have  $\mu_1(T, S)\mu_2(S) = \mu_2(T, S)\mu_1(S)$ .<sup>2</sup>*

[ link to proof ]

As a result, once we find one minimizer  $\nu$  of  $OInc_m$  (e.g., via (8.4)), it suffices to optimize  $SDef$  among distributions that have the same conditional marginals along  $\mathcal{A}$  that  $\nu$  does. This presents another problem:  $SDef$  is typically not convex. Fortunately, if we constrain to distributions that minimize  $OInc$ , then it is. Moreover, on this restricted domain, it can be represented with dcp exponential cone constraints.

**Proposition 8.4.** *If  $\mu \in [\![\mathbf{m}]\!]_0^*$ , then*

[ link to proof ]

$$SDef_m(\mu) = \sum_{\omega \in \mathcal{V}\mathcal{X}} \mu(\omega) \log \left( \frac{\mu(\omega)}{\prod_{a \in \mathcal{A}} \nu(T_a(\omega) | S_a(\omega))^{\alpha_a}} \right), \quad (8.7)$$

where  $\{\nu(T_a | S_a)\}_{a \in \mathcal{A}}$  are the cpds along the arcs  $\mathcal{A}$  shared by all distributions in  $[\![\mathbf{m}]\!]_0^*$  (per Proposition 8.3), and  $S_a(\omega), T_a(\omega)$  are the respective values of variables  $S_a$  and  $T_a$  in the joint setting  $\omega \in \mathcal{V}\mathcal{X}$ .

---

<sup>2</sup>Intuitively, this asserts  $\mu_1(T_a | S_a) = \mu_2(T_a | S_a)$ , but also handles cases where some  $\mu_1(S_a = s)$  or  $\mu_2(S_a = s)$  equals zero.

If we already know a distribution  $\nu \in \llbracket m \rrbracket_0^*$ , perhaps by solving (8.4), then the denominator of (8.7) does not depend on  $\mu$  and so is constant in our search for minimizers of  $SDef$ . For ease of exposition, aggregate these values into a vector

$$\mathbf{k} := \left[ \prod_{a \in \mathcal{A}} \nu(T_a(w)|S_a(w))^{\alpha_a} \right]_{w \in \mathcal{VX}}. \quad (8.8)$$

We can now capture  $\llbracket m \rrbracket_{0+}^*$  with a convex program.

**Proposition 8.5.** *If  $\nu \in \llbracket m \rrbracket_0^*$  and  $(\mu, \mathbf{u})$  solves the problem*

[ link to  
proof ]

$$\begin{aligned} & \underset{\mu, \mathbf{u}}{\text{minimize}} \quad \mathbf{1}^\top \mathbf{u} \\ & \text{subject to} \quad (-\mathbf{u}, \mu, \mathbf{k}) \in K_{\text{exp}}^{\mathcal{VX}}, \quad \mu \in \Delta \mathcal{VX}, \\ & \quad \forall S \xrightarrow{a} T \in \mathcal{A}. \quad \mu(S, T) \nu(S) = \mu(S) \nu(S, T), \end{aligned} \quad (8.9)$$

then  $\llbracket m \rrbracket_{0+}^* = \{\mu\}$  and  $\mathbf{1}^\top \mathbf{u} = SDef_m(\mu)$ .

Running (8.9) through a convex solver gives rise to the first algorithm that can reliably find  $\llbracket m \rrbracket_{0+}^*$ .

## 8.4 Polynomial-Time Inference Under Bounded Treewidth

We have now seen how  $\hat{\gamma}$ -inference (for small  $\hat{\gamma}$ ) can be reduced to convex optimization over joint distributions  $\mu$ —but  $\mu$  grows exponentially with the number of variables in the PDG, so we do not yet have a tractable inference algorithm. We now show how  $\mu$  can be replaced with a tree marginal over the PDG’s structure. What makes this possible is a key independence property of traditional graphical models, which we now prove holds for PDGs as well.

[ link to  
proof ]

**Theorem 8.6** (Markov Property for PDGs). *If  $\mathbf{m}_1$  and  $\mathbf{m}_2$  are PDGs over sets  $\mathcal{X}_1$  and  $\mathcal{X}_2$  of variables, respectively, then for all  $\gamma > 0$  and  $\gamma = 0^+$ ,*

$$[\![\mathbf{m}_1 + \mathbf{m}_2]\!]_{\gamma}^* \models \mathcal{X}_1 \perp\!\!\!\perp \mathcal{X}_2 \mid \mathcal{X}_1 \cap \mathcal{X}_2.$$

*That is: for every distribution  $\mu \in [\![\mathbf{m}_1 + \mathbf{m}_2]\!]_{\gamma}^*$ , the variables of  $\mathbf{m}_1$  and of  $\mathbf{m}_2$  are conditionally independent given the variables they have in common.*

For the remainder of this section, fix a PDG  $\mathbf{m}$  and a tree decomposition  $(\mathcal{C}, \mathcal{T})$  of  $\mathbf{m}$ 's hypergraph. One significant consequence of [Theorem 8.6](#) is that, in the search for optimizers of (3.4), we need consider only distributions that satisfy those independencies, all of which can be represented as a tree marginal  $\boldsymbol{\mu} = \{\mu_C \in \Delta \mathcal{V}(C)\}_{C \in \mathcal{C}}$  over  $(\mathcal{C}, \mathcal{T})$ .

**Corollary 8.6.1.** *If  $\mathbf{m}$  is a PDG with arcs  $\mathcal{A}$ ,  $(\mathcal{C}, \mathcal{T})$  is a tree decomposition of  $\mathcal{A}$ ,  $\gamma > 0$ , and  $\mu \in [\![\mathbf{m}]\!]_{\gamma}^*$ , then there exists a tree marginal  $\boldsymbol{\mu}$  over  $(\mathcal{C}, \mathcal{T})$  such that  $\Pr_{\boldsymbol{\mu}} = \mu$ .*

[ link to  
proof ]

For convenience, let  $\mathcal{VC} := \{(C, c) : C \in \mathcal{C}, c \in \mathcal{V}(C)\}$  be the set of all choices of a cluster together with a setting of its variables. Like before, we start by optimizing  $OInc$ , this time over calibrated tree marginals  $\boldsymbol{\mu}$ , which we identify with vectors  $\boldsymbol{\mu} \cong [\mu_C(C=c)]_{(C,c) \in \mathcal{VC}}$ . We need the conditional marginals  $\Pr_{\boldsymbol{\mu}}(T_a | S_a)$  of  $\boldsymbol{\mu}$  along every arc  $a$  in order to calculate  $OIncm(\Pr_{\boldsymbol{\mu}})$ ; fortunately, they are readily available. Since  $(\mathcal{C}, \mathcal{T})$  is a tree decomposition, we know  $S_a$  and  $T_a$  lie entirely within some cluster  $C_a \in \mathcal{C}$ , and  $\Pr_{\boldsymbol{\mu}}(T_a | S_a) = \mu_{C_a}(T_a | S_a)$  if  $\boldsymbol{\mu}$  is

calibrated. For  $\mathbf{u} \in \bar{\mathbb{R}}^{\mathcal{V}\mathcal{A}}$ , consider the problem

$$\begin{aligned} & \underset{\boldsymbol{\mu}, \mathbf{u}}{\text{minimize}} \quad \sum_{(a,s,t) \in \mathcal{V}\mathcal{A}} \beta_a u_{a,s,t} \\ & \text{subject to} \quad \forall C \in \mathcal{C}. \mu_C \in \Delta\mathcal{V}(C), \\ & \quad \forall a \in \mathcal{A}. (-\mathbf{u}_a, \mu_{Ca}(S_a, T_a), \mu_{Ca}(S_a) \mathbb{P}_a(T_a | S_a)) \in K_{\text{exp}}^{\mathcal{V}a} \\ & \quad \forall (C, D) \in \mathcal{T}. \mu_C(C \cap D) = \mu_D(C \cap D), \end{aligned} \tag{8.10}$$

where again  $\mathbf{u}_a$  is the restriction of  $\mathbf{u}$  to components associated with  $a$ . Problem (8.10) is similar to (8.4), except that it requires local marginal constraints to restrict our search to calibrated tree marginals. It is analogous to problem CTREE-OPTIMIZE-KL of [Koller and Friedman \(2009, pg. 384\)](#).

**Proposition 8.7.** *If  $(\boldsymbol{\mu}, \mathbf{u})$  is a solution to (8.10), then*

link to  
proof

- (a)  $\boldsymbol{\mu}$  is a calibrated, with  $\Pr_{\boldsymbol{\mu}} \in [\![m]\!]_0^*$ , and
- (b) the objective of (8.10) evaluated at  $\mathbf{u}$  equals  $\langle\!\langle m \rangle\! \rangle_0$ .

We can now find a minimizer of  $OInc$  and compute  $\langle\!\langle m \rangle\! \rangle_0$  without storing a joint distribution. But to do anything else, we must deviate from the template laid out in [Section 8.3](#).

**Dealing with Joint Entropy.** In the construction of (8.10), we rely heavily on the fact that each term of  $OInc_m$  depends only on local marginal distributions  $\mu_{Ca}(T_a, S_a)$  and  $\mu_{Ca}(S_a)$ . The same is not true of  $SDef$ , which depends on the joint entropy  $H(\Pr_{\boldsymbol{\mu}})$  of the entire distribution. At this point we should point out an important reason to restrict our focus to trees: it allows the joint entropy to be expressed in terms of the cluster marginals ([Wainwright et al. 2008](#)), by

$$-H(\Pr_{\boldsymbol{\mu}}) = -\sum_{C \in \mathcal{C}} H(\mu_C) + \sum_{(C,D) \in \mathcal{T}} H_{\boldsymbol{\mu}}(C \cap D). \tag{8.11}$$

Even so, it is not obvious that (8.11) can be captured with dcp exponential cone constraints. (Exponential conic programs can minimize negative entropy, but not positive entropy, which is concave.) We now describe how this can be done.

Choose a root node  $C_0$  of the tree decomposition, and orient each edge of  $\mathcal{T}$  so that it points away from  $C_0$ . Each cluster  $C \in \mathcal{C}$ , except for  $C_0$ , then has a parent cluster  $\text{Par}(C)$ ; define  $\text{Par}(C_0) := \emptyset$  to be an empty cluster, since  $C_0$  has no parent. Finally, for each  $C \in \mathcal{C}$ , let  $VCP_C := C \cap \text{Par}(C)$  denote the set of variables that cluster  $C$  has in common with its parent cluster.<sup>3</sup> As  $\mathcal{T}$  is now a directed tree, this definition allow us to express (8.11) in a more useful form:

$$\begin{aligned} -H(\Pr_{\boldsymbol{\mu}}) &= -H(\mu_{C_0}) - \sum_{(C \rightarrow D) \in \mathcal{T}} H_{\Pr_{\boldsymbol{\mu}}}(D | C) \\ &= \sum_{C \in \mathcal{C}} \sum_{c \in V(C)} \mu_C(C=c) \log \frac{\mu_C(C=c)}{\mu_C(VCP_C(c))}, \end{aligned} \quad (8.12)$$

where  $VCP_C(c)$  is the restriction of the joint value  $c \in \mathcal{V}(C)$  to the variables  $VCP_C \subseteq C$ . Crucially, the denominator of (8.12) is an affine transformation of  $\mu_C$ . The upshot: we have rewritten the joint entropy as a sum of functions of the clusters, each of which can be captured with a dcp exponential cone constraint. This gives us analogues of the problems in Sections 8.3.2 and 8.3.3 that operate on tree marginals.

**Finding tree marginals for  $\gamma$ -inference.** The ability to decompose the joint entropy as in (8.12) allows us to adapt (8.6) to operate on calibrated tree marginals, rather than joint distributions. Beyond the changes already present in (8.10), the key is to replace the exponential cone constraint  $(-\mathbf{v}, \boldsymbol{\mu}, \mathbf{1}) \in K_{\text{exp}}^{\mathcal{V}\mathcal{X}}$ , which captures the entropy of  $\boldsymbol{\mu}$ , with

$$(-\mathbf{v}, \boldsymbol{\mu}, [\mu_C(VCP_C(c))]_{(C,c) \in \mathcal{V}\mathcal{C}}) \in K_{\text{exp}}^{\mathcal{V}\mathcal{C}},$$

which captures the entropy of  $\boldsymbol{\mu}$ , by (8.12). Over vectors  $\mathbf{v}, \boldsymbol{\mu} \in \bar{\mathbb{R}}^{\mathcal{V}\mathcal{C}}$  and  $\mathbf{u} \in \bar{\mathbb{R}}^{\mathcal{V}\mathcal{A}}$ ,

the problem becomes:

$$\begin{aligned}
& \underset{\boldsymbol{\mu}, \mathbf{u}, \mathbf{v}}{\text{minimize}} \quad \sum_{(a,s,t) \in \mathcal{VA}} (\beta_a - \alpha_a \gamma) u_{a,s,t} + \gamma \sum_{(C,c) \in \mathcal{VC}} v_{C,c} \\
& \quad - \sum_{(a,s,t) \in \mathcal{VA}^+} \alpha_a \gamma \mu_{C_a}(S_a = s, T_a = t) \log \mathbb{P}_a(T_a = t \mid s) \\
& \text{subject to} \quad \forall C \in \mathcal{C}. \mu_C \in \Delta \mathcal{V}(C), \\
& \forall a \in \mathcal{A}. (-\mathbf{u}_a, \mu_{C_a}(S_a, T_a), \mu_{C_a}(S_a) \mathbb{P}_a(T_a | S_a)) \in K_{\text{exp}}^{\mathcal{V}a}, \\
& \forall (a, s, t) \in \mathcal{VA}^0. \mu_{C_a}(S_a = s, T_a = t) = 0, \\
& \forall (C, D) \in \mathcal{T}. \mu_C(C \cap D) = \mu_D(C \cap D), \\
& (-\mathbf{v}, \boldsymbol{\mu}, [\mu_C(VCP_C(c))]_{(C,c) \in \mathcal{VC}}) \in K_{\text{exp}}^{\mathcal{VC}}.
\end{aligned} \tag{8.13}$$

**Proposition 8.8.** If  $(\boldsymbol{\mu}, \mathbf{u}, \mathbf{v})$  is a solution to (8.13) and  $\beta \geq \gamma \alpha$ , then  $\Pr_{\boldsymbol{\mu}}$  is the unique element of  $[\![\mathbf{m}]\!]_{\gamma}^*$ , and the objective of (8.13) at  $(\boldsymbol{\mu}, \mathbf{u}, \mathbf{v})$  equals  $\langle\!\langle \mathbf{m} \rangle\!\rangle_{\gamma}$ .

link to  
proof

A related use of (8.12) is to enable an analogue of (8.9) that searches over tree marginals (rather than joint distributions), to find a compact representation of  $[\![\mathbf{m}]\!]_{0+}^*$ . We begin with a straightforward adaptation of the relevant machinery in Section 8.3.3. Suppose that  $\boldsymbol{\nu} = \{\nu_C : C \in \mathcal{C}\}$  is a calibrated tree marginal over the tree decomposition  $(\mathcal{C}, \mathcal{T})$  representing a distribution  $\Pr_{\boldsymbol{\nu}} \in [\![\mathbf{m}]\!]_0^*$ , say obtained by solving (8.10). For  $C \in \mathcal{C}$ , let  $\mathcal{A}_C := \{a \in \mathcal{A} : C_a = C\}$  be the set of arcs assigned to cluster  $C$ , and let

$$\mathbf{k} := \left[ \prod_{a \in \mathcal{A}_C} \nu_C(T_a(c) | S_a(c))^{\alpha_a} \right]_{(C,c) \in \mathcal{VC}} \in \bar{\mathbb{R}}^{\mathcal{VC}}$$

be the analogue of (8.8) for a cluster tree. Once again, consider  $\mathbf{u} := [u_{(C,c)}]_{(C,c) \in \mathcal{VC}}$

---

<sup>3</sup>Different choices of  $C_0$  yield different definitions of  $VCP$ , and ultimately optimization problems of different sizes; the optimal choice can be found with Edmund's Algorithm (Chu 1965), which computes a directed analogue of the minimum spanning tree.

in the optimization problem

$$\underset{\boldsymbol{\mu}, \mathbf{u}}{\text{minimize}} \quad \mathbf{1}^T \mathbf{u} \quad (8.14)$$

$$\text{subject to } \forall C \in \mathcal{C}. \mu_C \in \Delta \mathcal{V}(C),$$

$$(-\mathbf{u}, \boldsymbol{\mu}, \mathbf{k} \odot [\mu_C(VCP_C(c))]_{(C,c) \in \mathcal{VC}}) \in K_{\text{exp}}^{\mathcal{VC}},$$

$$\forall a \in \mathcal{A}. \mu_{Ca}(S_a, T_a) \nu_{Ca}(S_a) = \mu_{Ca}(S_a) \nu_{Ca}(S_a, T_a)$$

$$\forall (C, D) \in \mathcal{T}. \mu_C(C \cap D) = \mu_D(C \cap D).$$

The biggest change is in the second constraint: the upper bounds  $[\mu_{(C,c)}]_{c \in \mathcal{VC}}$  for cluster  $C$  now account only for the additional entropy not already modeled by  $C$ 's ancestors.

**Proposition 8.9.** *If  $(\boldsymbol{\mu}, \mathbf{u})$  is a solution to (8.14), then  $\boldsymbol{\mu}$  is a calibrated tree marginal*

[ link to  
proof ]

$$\text{and } \llbracket \mathbf{m} \rrbracket_{0^+}^* = \{\Pr_{\boldsymbol{\mu}}\}.$$

At this point, standard algorithms can use  $\boldsymbol{\mu}$  to answer probabilistic queries about  $\Pr_{\boldsymbol{\mu}}$  in polynomial time (Koller and Friedman 2009, §10.3.3).<sup>4</sup> From Propositions 8.8 and 8.9, it follows that  $\hat{\gamma}$ -inference (for small  $\hat{\gamma}$ , and for  $0^+$ ) can be reduced to a (pair of) convex optimization problem(s) with a polynomial number of variables and constraints. All that remains is to show that such a problem can be solved in polynomial time. For this, we turn to interior-point methods. As (8.13) and (8.14) are dcp, they can be transformed via established methods (Agrawal et al. 2018) into a standard form that can be solved in polynomial time by commercial solvers (ApS 2022; Domahidi et al. 2013). Threading the details of our constructions through the analyses of Dahl and Andersen (2022) and Nesterov et al. (1999) results in our main theorem.

---

<sup>4</sup>Concretely: marginal probabilities can essentially be read off of a calibrated a tree marginal, and evidence  $X=x$  may be incorporated by setting  $\mu_C(c) := 0$  for every  $C=c$  that conflicts with  $X=x$  and recalibrating the tree marginal (e.g., with belief propagation).

[ link to  
proof ]

**Theorem 8.10.** Let  $\mathbf{m} = (\mathcal{X}, \mathcal{A}, \mathbb{P}, \boldsymbol{\alpha}, \boldsymbol{\beta})$  be a proper discrete PDG with  $N = |\mathcal{X}|$  variables each taking at most  $V$  values and  $A = |\mathcal{A}|$  arcs, in which each component of  $\boldsymbol{\beta} \in \mathbb{R}^{\mathcal{A}}$  and  $\mathbb{P} \in \mathbb{R}^{V\mathcal{A}}$  is specified in binary with at most  $k$  bits. Suppose that  $\gamma \in \{0^+\} \cup (0, \min_{a \in \mathcal{A}} \frac{\beta_a}{\alpha_a}]$ . If  $(\mathcal{C}, \mathcal{T})$  is a tree decomposition of  $(\mathcal{X}, \mathcal{A})$  of width  $T$  and  $\boldsymbol{\mu}^* \in \mathbb{R}^{V\mathcal{C}}$  is the unique calibrated tree marginal over  $(\mathcal{C}, \mathcal{T})$  that represents the  $\hat{\gamma}$ -semantics of  $\mathbf{m}$ , then

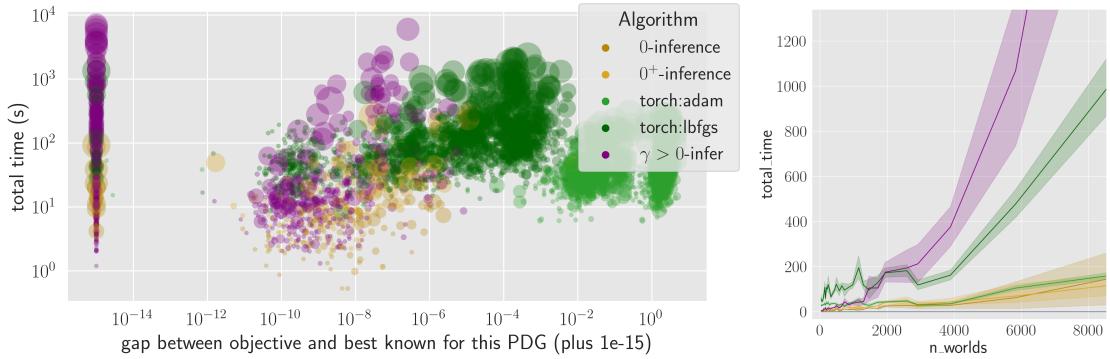
- (a) Given  $\mathbf{m}$ ,  $\gamma$ , and  $\epsilon > 0$ , we can find a calibrated tree marginal  $\epsilon$  close in  $\ell_2$  norm to  $\boldsymbol{\mu}^*$  in time<sup>1</sup>

$$\begin{aligned} O\left(|V\mathcal{A} + V\mathcal{C}|^4 \left(\log |V\mathcal{A} + V\mathcal{C}| + \log \frac{1}{\epsilon}\right) k^2 \log k\right) \\ \subseteq \tilde{O}\left(k^2 |V\mathcal{A} + V\mathcal{C}|^4 \log^{1/\epsilon}\right) \\ \subseteq \tilde{O}\left(k^2 (N + A)^4 V^{4(T+1)} \log^{1/\epsilon}\right). \end{aligned}$$

- (b) The unique tree marginal closest to  $\boldsymbol{\mu}^*$  in which every component is represented with a  $k$ -bit binary number, can be calculated in time<sup>1</sup>

$$\tilde{O}\left(k^2 |V\mathcal{A} + V\mathcal{C}|^4\right) \subseteq \tilde{O}\left(k^2 (N + A)^4 V^{4(T+1)}\right).$$

Observe that the dependence on the precision is  $\log(1/\epsilon)$ , which is optimal in the sense that, in general, it takes time  $\Omega(\log 1/\epsilon)$  to write down the binary representation of any number within  $\epsilon$  of a given value.<sup>5</sup> In practice, this procedure can be used as if it were an exact algorithm, with no more overhead than that incurred by floating point arithmetic.



*Figure 8.1:* Accuracy and resource costs for the methods in Section 8.3. Left: a scatter plot of several algorithms on random PDGs of  $\approx 10$  variables. The x-axis is the difference in scores  $\llbracket m \rrbracket_\gamma(\mu) - \llbracket m \rrbracket_\gamma(\mu^*) + 10^{-15}$ , where  $\mu$  is the method’s output, and  $\mu^*$  achieves best (smallest) known value of  $\llbracket m \rrbracket_\gamma$ . (Thus, the best solutions lie on the far left.) The y axis is the time required to compute  $\mu$ . Our methods are in gold ( $0^+$ -inference) and violet ( $\hat{\gamma}$ -inference, for  $\hat{\gamma} > 0$ ); the baselines (black-box optimizers applied directly to (3.4)) are in green. The area of each circle is proportional to the size of the optimization problem, as measured by  $n\_worlds := |\mathcal{V}\mathcal{X}|$ . Right: how the same methods scale in run time, as  $|\mathcal{V}\mathcal{X}|$  increases.

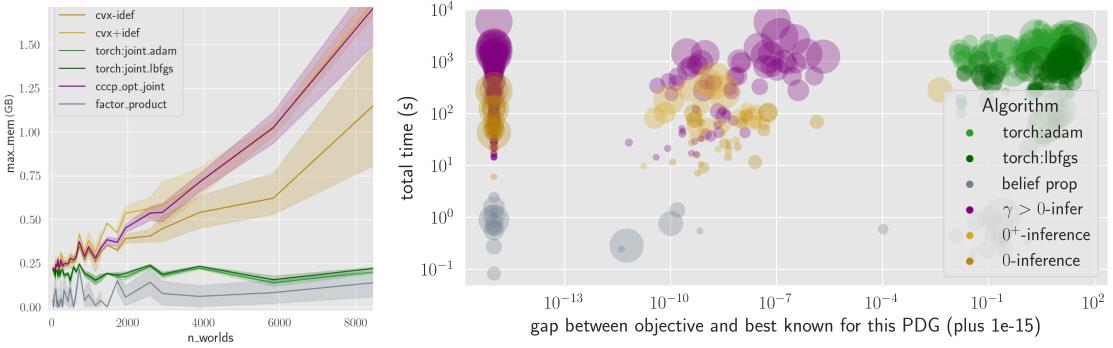
## 8.5 Experiments

We have given the first algorithm to provably do inference in polynomial time, but that does not mean that it is the best way of answering queries in practice; it also makes sense to use black-box optimization tools such as Adam (Kingma and Ba 2014) or L-BFGS (Fletcher 2013) to find minimizers of  $\llbracket m \rrbracket_\gamma$ . Indeed, this scoring function has several properties that make it highly amenable to such methods: it is infinitely differentiable,  $\gamma$ -strongly convex, and its derivatives have simple closed-form expressions. So it may seem surprising that  $\llbracket m \rrbracket_\gamma$  poses a challenge to standard optimization tools—but it does, even when we optimize directly over joint distributions.

**Synthetic Experiment 1** (over joint distributions). Repeatedly do the follow-

---

<sup>5</sup>More precisely: if a value  $x$  is chosen uniformly from  $[0, 1]$ , then with probability  $1 - \sqrt{\epsilon}$  the binary representation of every  $y \in [x - \epsilon, x + \epsilon]$  has at least  $\lfloor \frac{1}{2} \log_2 1/\epsilon \rfloor - 1$  bits.



*Figure 8.2:* Left: Memory footprint. The convex solver (violet, gold) requires more memory than baselines (green). Right: Analogue of Figure 8.1 for the cluster setting. Here there is even more separation between exponential conic optimization (gold, violet) and black-box optimization (greens). The grey points represent belief propagation, which is fastest and most accurate—but only applies in the special case when  $\beta = \gamma\alpha$ .

ing. First, randomly generate a small PDG  $m$  containing at most 10 variables and 15 arcs. Then for various values of  $\gamma \in \{0, 0^+, 10^{-8}, \dots, \min_a \frac{\beta_a}{\alpha_a}\}$ , optimize  $[m]_\gamma(\mu)$  over joint distributions  $\mu$ , in one of two ways.

- (a) Use cvxpy (Diamond and Boyd 2016) to feed one of problems (8.4,8.6,8.9) to the MOSEK solver (ApS 2022), or
- (b) Choose a learning rate and a representation of  $\mu$  in terms of optimization variables  $\theta \in \mathbb{R}^n$ . Then run a standard optimizer (Adam or L-BFGS) built into pytorch (Paszke et al. 2019) to optimize  $\theta$  until  $\mu_\theta$  converges to a minimizer of  $[m]_\gamma$  (or a time limit is reached). Keep only the best result across all learning rates.

The results are shown in Figure 8.1. Observe that the convex solver (gold, violet) is significantly more accurate than the baselines, and also much faster for small PDGs. Our implementation of  $0^+$ -inference (gold) also appears to scale better than L-BFGS in this regime, although that of  $\hat{\gamma}$ -inference (purple) seems to scale much worse. We suspect that the difference comes from cvxpy's compilation process, because the two use similar amounts of memory (Figure 8.2),

and so are problems of similar sizes.

**Synthetic Experiment 2** (over tree marginals). For PDGs of bounded treewidth, Corollary 8.6.1 allows us to express these optimization problems compactly not just for the convex solver, but for the black-box baseline approaches as well. We adapt the previous experiment for tree marginals as follows. First randomly sample a maximal graph  $G$  of tree-width  $k$ , called a  $k$ -tree (Patil 1986); then generate a PDG  $\mathcal{m}$  whose hyperarcs lie within cliques of  $G$ . This ensures that the maximal cliques of  $G$  form a tree-decomposition  $(\mathcal{C}, \mathcal{T})$  of  $\mathcal{m}$ 's underlying hypergraph. We can now proceed as before: either encode (8.10, 8.13, 8.14) as disciplined convex programs in `cvxpy`, or use `torch` to directly minimize  $\llbracket \mathcal{m} \rrbracket_\gamma(\Pr_{\boldsymbol{\mu}})$  amongst tree marginals  $\boldsymbol{\mu}$  over  $(\mathcal{C}, \mathcal{T})$ .

In the latter case, however, there is now an additional difficulty: it is not easy to strictly enforce the calibration constraints with the black-box methods. Common practice is to instead add extra loss terms to “encourage” calibration—but it can still be worthwhile for the optimizer to simply incur that loss in order to violate the constraints. Thus, for fairness, we must recalibrate the tree marginals returned by all methods before evaluation. The result is an even more significant advantage for the convex solver; see Figure 8.2.

**Evaluation on BNs.** We also applied the procedure of the Synthetic Experiment 2 to the smaller BNs in the `bnlearn` repository, and found similar results (but with fewer examples; see Section 8.C.3). But for a PDG that happens to also be a BN, it is possible to use belief propagation, which is much faster and at least as accurate.

Explicit details about all of our experiments, and many more figures, can be found in Section 8.C.

## 8.6 Discussion and Conclusion

In this chapter, we have provided the first practical algorithm for inference in PDGs. In more detail, we have defined a parametric family of PDG inference notions, given a fixed-parameter tractable inference algorithm for a subset of these parameters, proven our algorithm correct, implemented it, and shown our code to empirically outperform baselines. Yet many questions about PDG inference remain open.

Asymptotically, there may be a lot of room for improvement. Our implementation runs in time  $\tilde{O}(N^4)$ , and our analysis suggests one of time  $\tilde{O}(N^{2.872})$ . But assuming bounded tree-width, most graph problems, including inference inference for BNs and FGs, can be solved in time  $O(N)$ .

Furthermore, we have shown how to do inference for only a subset of possible parameter values, specifically, when either  $\beta \geq \gamma\alpha$  or  $\beta \gg \alpha$ . The remaining cases are also of interest, and likely require different techniques. When  $\beta = 0$  and  $(\mathcal{A}, \alpha)$  encodes the structure of a BN, for instance, inference is about characterizing the BN’s independencies. While we do not know how to tackle the inference problem in the general setting, our methods can be augmented with the convex-concave procedure (Yuille and Rangarajan 2003) to obtain an inference algorithm that applies slightly more broadly; see [Section 8.B](#). We imagine that this extension could also be useful for computing with PDGs beyond the specific inference problem considered in this chapter.

Our analysis does not resolve these problems, but it does shed light on some of them. The 0-semantics, for instance, is characterized by [Propositions 8.3](#) and [8.7](#). Also, when  $\llbracket m \rrbracket_\gamma$  is not convex, we can still find an optimal distribution with

the concave-convex procedure [Yuille and Rangarajan \(2003\)](#), which we do in [Section 8.B](#)—but this only suffices for inference if we already know there's a unique optimal distribution. In some cases, this might actually allow us to do inference—say, if we happen to know for external reasons that  $\llbracket m \rrbracket_{\gamma}^*$  is pseudo-convex (although we loose polynomial time guarantees and have no ability to automatically recognize such situations). In any case, we have implemented this, and describe it in [Section 8.B](#).

Given the long history of improvements to our understanding of inference for Bayesian networks, we are optimistic that faster and more general inference algorithms for PDGs are possible.

## APPENDICES FOR CHAPTER 8

### 8.A Proofs

Our results fall broadly into three categories:

1. Foundational results about PDGs that we needed to prove to get an inference procedure, but which are likely to be generally useful for anyone working with PDGs ([Section 8.A.1](#));
2. Correctness and efficiency results, showing that the optimization problems we present in the body of the chapter give the correct answers, and that they can be formulated and solved in polynomial time; ([Section 8.A.2](#))
3. Hardness results, i.e., [Theorem 9.4](#) and the constructions and lemmas needed to support it ([Section 9.4.1](#)).

#### 8.A.1 Properties of PDG Semantics Needed for Inference

**Proposition 8.3.** *If  $\mathcal{M}$  has arcs  $\mathcal{A}$  and  $\beta \geq 0$ , the minimizers of  $OInc_{\mathcal{M}}$  all have the same conditional marginals along  $\mathcal{A}$ . That is, for all  $\mu_1, \mu_2 \in [\![\mathcal{M}]\!]_0^*$  and all  $S \xrightarrow{a} T \in \mathcal{A}$  with  $\beta_a > 0$ , we have  $\mu_1(T, S)\mu_2(S) = \mu_2(T, S)\mu_1(S)$ .*

*Proof.* For contradiction, suppose that  $\mu_1, \mu_2 \in [\![\mathcal{M}]\!]_0^*$ , but there is some  $(\hat{a}, \hat{s}, \hat{t}) \in \mathcal{V}\mathcal{A}$  such that  $\beta_{\hat{a}} > 0$  and

$$\mu_1(T_{\hat{a}}=\hat{t}, S_{\hat{a}}=\hat{s})\mu_2(S_{\hat{a}}=\hat{s}) \neq \mu_2(T_{\hat{a}}=\hat{t}, S_{\hat{a}}=\hat{s})\mu_1(S_{\hat{a}}=\hat{s}).$$

For  $t \in [0, 1]$ , let  $\mu_t := (1 - t)\mu_0 + t\mu_1$  as before. Then define

$$F(t) := D\left(\mu_t(S_a, T_a) \parallel \mu_t(S_a)\mathbb{P}_a(T_a|S_a)\right).$$

Since  $\mu_0(S_a, T_a)$  and  $\mu_1(S_a, T_a)$  are joint distributions over two variables, with different conditional marginals, as above, Lemma 8.12 applies, and so  $F(t)$  is strictly convex.

Let

$$OInc_{m \setminus \hat{a}} := \sum_{a \neq \hat{a}} \beta_a D(\mu(T_a, S_a) \parallel \mathbb{P}_a(T_a|S_a)\mu(S_a))$$

be the observational incompatibility loss, but without the term corresponding to edge  $\hat{a}$ . Since  $OInc_{m \setminus \hat{a}}$  is convex in its argument, it is in particular convex along the segment from  $\mu_0$  to  $\mu_1$ ; that is, for  $t \in [0, 1]$ , the function  $t \mapsto OInc_{m \setminus \hat{a}}(\mu_t)$  is convex. Therefore, we know that the function

$$G(t) := OInc_m(\mu_t) = OInc_{m \setminus \hat{a}}(\mu_t) + \beta_{\hat{a}} F(t),$$

is *strictly* convex. But then this means  $\mu_{1/2}$  satisfies

$$OInc_m(\mu_{1/2}) < OInc_m(\mu_0),$$

contradicting the premise that  $\mu_0$  minimizes  $OInc_m$  (i.e.,  $\mu_0 \in [\![m]\!]_0^*$ ). Therefore, it must be the case that all distributions in  $[\![m]\!]_0^*$  have the same conditional marginals, as promised.  $\square$

**Proposition 8.4.** If  $\mu \in [\![\mathbf{m}]\!]_0^*$ , then

$$SDef_m(\mu) = \sum_{\omega \in \mathcal{V}\mathcal{X}} \mu(\omega) \log \left( \frac{\mu(\omega)}{\prod_{a \in \mathcal{A}} \nu(T_a(\omega) | S_a(\omega))^{\alpha_a}} \right), \quad (8.7)$$

where  $\{\nu(T_a | S_a)\}_{a \in \mathcal{A}}$  are the cpds along the arcs  $\mathcal{A}$  shared by all distributions in  $[\![\mathbf{m}]\!]_0^*$  (per Proposition 8.3), and  $S_a(\omega), T_a(\omega)$  are the respective values of variables  $S_a$  and  $T_a$  in the joint setting  $\omega \in \mathcal{V}\mathcal{X}$ .

*Proof.* This is mostly a simple algebraic manipulation. By definition:

$$\begin{aligned} SDef_m(\mu) &= -H(\mu) + \sum_{a \in \mathcal{A}} \alpha_a H_\mu(T_a | S_a) \\ &= \mathbb{E}_\mu \left[ -\log \frac{1}{\mu} + \sum_{a \in \mathcal{A}} \alpha_a \log \frac{1}{\mu(T_a | S_a)} \right] \\ &= \sum_{w \in \mathcal{V}\mathcal{X}} \mu(w) \left[ \log \mu(w) + \sum_{a \in \mathcal{A}} \log \frac{1}{\mu(T_a(w) | S_a(w))^{\alpha_a}} \right] \\ &= \sum_{w \in \mathcal{V}\mathcal{X}} \mu(w) \log \left( \frac{\mu(w)}{\prod_{a \in \mathcal{A}} \mu(T_a(w) | S_a(w))^{\alpha_a}} \right) \end{aligned}$$

But, by Proposition 8.3, if we restrict  $\mu \in [\![\mathbf{m}]\!]_0^*$ , then the conditional marginals in the denominator do not depend on the particular choice of  $\mu$ ; they're shared among all  $\nu \in [\![\mathbf{m}]\!]_0^*$ .  $\square$

**Theorem 8.6.** If  $\mathbf{m}_1$  and  $\mathbf{m}_2$  are PDGs over sets  $\mathcal{X}_1$  and  $\mathcal{X}_2$  of variables, respectively, then for all  $\gamma > 0$  and  $\gamma = 0^+$ ,

$$[\![\mathbf{m}_1 + \mathbf{m}_2]\!]_\gamma^* \models \mathcal{X}_1 \perp\!\!\!\perp \mathcal{X}_2 \mid \mathcal{X}_1 \cap \mathcal{X}_2.$$

That is: for every distribution  $\mu \in [\![\mathbf{m}_1 + \mathbf{m}_2]\!]_\gamma^*$ , the variables of  $\mathbf{m}_1$  and of  $\mathbf{m}_2$  are conditionally independent given the variables they have in common.

Or symbolically:  $\mathbf{m}_1 + \mathbf{m}_2 \models \mathcal{X}_1 \perp\!\!\!\perp \mathcal{X}_2 \mid \mathcal{X}_1 \cap \mathcal{X}_2$ .

*Proof.* Note that, save for the joint entropy, every summand the scoring function  $\llbracket \mathbf{m}_1 + \mathbf{m}_2 \rrbracket_\gamma : \Delta(\mathcal{V}\mathcal{X}_1 \times \mathcal{V}\mathcal{X}_2)$ , is a function of the conditional marginal of  $\mu$  along some edge. In particular, those terms that correspond to edges of  $\mathbf{m}_1$  can be computed from the marginal  $\mu(\mathcal{X}_1)$ , while those that correspond to edges of  $\mathbf{m}_2$  can be computed from the marginal  $\mu(\mathcal{X}_2)$ . Therefore, there are functions  $f$  and  $g$  such that:

$$\llbracket \mathbf{m}_1 + \mathbf{m}_2 \rrbracket_\gamma(\mu) = f(\mu(\mathcal{X}_1)) + g(\mu(\mathcal{X}_2)) - \gamma H(\mu).$$

To make this next step extra clear, let  $\mathbf{X} := \mathcal{X}_1 \setminus \mathcal{X}_2$  and  $\mathbf{Z} := \mathcal{X}_2 \setminus \mathcal{X}_1$ , be the variables unique to each PDG, and  $\mathbf{S} := \mathcal{X}_1 \cap \mathcal{X}_2$  be the set of variables they have in common, so that  $(\mathbf{X}, \mathbf{S}, \mathbf{Z})$  is a partition of all variables  $\mathbf{X}_1 \cup \mathbf{X}_2$ . Now define a new distribution  $\mu' \in \Delta(\mathcal{V}\mathcal{X}_1 \times \mathcal{V}\mathcal{X}_2)$  by

$$\mu'(\mathbf{X}, \mathbf{S}, \mathbf{Z}) := \mu(\mathbf{S})\mu(\mathbf{Z} | \mathbf{S})\mu(\mathbf{X} | \mathbf{S}) \quad \left( = \mu(\mathbf{X}, \mathbf{S})\mu(\mathbf{Z} | \mathbf{S}) = \mu(\mathbf{Z}, \mathbf{S})\mu(\mathbf{X} | \mathbf{S}) \right).$$

One can easily verify that  $\mathbf{X}$  and  $\mathbf{Z}$  are independent given  $\mathbf{S}$  in  $\mu'$  (by construction), and the alternate forms on the right make it easy to see that  $\mu(\mathcal{X}_1) = \mu'(\mathcal{X}_1)$  and  $\mu(\mathcal{X}_2) = \mu'(\mathcal{X}_2)$ . Furthermore, for any  $\nu'(\mathbf{X}, \mathbf{S}, \mathbf{Z})$ , we can write

$$\begin{aligned} H(\nu) &= H_\nu(\mathbf{X}, \mathbf{S}, \mathbf{Z}) = H_\nu(\mathbf{X}, \mathbf{S}) + H_\nu(\mathbf{Z} | \mathbf{X}, \mathbf{S}) \\ &= H_\nu(\mathbf{X}, \mathbf{S}) + H_\nu(\mathbf{Z} | \mathbf{X}, \mathbf{S}) - H_\nu(\mathbf{Z} | \mathbf{S}) + H_\nu(\mathbf{Z} | \mathbf{S}) \\ &= H_\nu(\mathbf{X}, \mathbf{S}) + H_\nu(\mathbf{Z} | \mathbf{S}) - I_\nu(\mathbf{Z}; \mathbf{X} | \mathbf{S}), \end{aligned}$$

where  $I_\nu(\mathbf{X}; \mathbf{Z} | \mathbf{S})$ , the conditional mutual information between  $\mathbf{X}$  and  $\mathbf{Z}$  given  $\mathbf{S}$  (in  $\nu$ ), is non-negative, and equal to zero if and only if  $\mathbf{X}$  and  $\mathbf{Z}$  are conditionally independent given  $\mathbf{S}$  (see, for instance, MacKay 2003, §1). So  $I_{\mu'}(\mathbf{X}; \mathbf{Z} | \mathbf{S}) = 0$ , and  $H_{\mu'} = H_{\mu'}(\mathbf{X}, \mathbf{S}) + H_{\mu'}(\mathbf{Z} | \mathbf{S})$ . Because  $\mu$  and  $\mu'$  share marginals on  $\mathcal{X}_1$  and  $\mathcal{X}_2$ , while the terms  $H(\mathbf{X}, \mathbf{S})$  and  $H(\mathbf{Z} | \mathbf{S})$  depend only on these marginals, respectively, we

also know that  $H_\mu(\mathbf{X}, \mathbf{S}) = H_{\mu'}(\mathbf{X}, \mathbf{S})$  and  $H_\mu(\mathbf{Z}|\mathbf{S}) = H_{\mu'}(\mathbf{Z}|\mathbf{S})$ ; thus we have

$$\begin{aligned} H(\mu) &= H_\mu(\mathbf{X}, \mathbf{S}) + H_\mu(\mathbf{Z} | \mathbf{S}) - I_\mu(\mathbf{Z}; \mathbf{X} | \mathbf{S}) \\ &= H(\mu') - I_\mu(\mathbf{Z}; \mathbf{X} | \mathbf{S}). \end{aligned}$$

Therefore,

$$\begin{aligned} \llbracket \mathbf{m}_1 + \mathbf{m}_2 \rrbracket_\gamma(\mu) &= f(\mu(\mathcal{X}_1)) + g(\mu(\mathcal{X}_2)) - \gamma H(\mu) \\ &= f(\mu'(\mathcal{X}_1)) + g(\mu'(\mathcal{X}_2)) - \gamma H(\mu') + \gamma I_\mu(\mathbf{Z}; \mathbf{X} | \mathbf{S}) \\ &= \llbracket \mathbf{m}_1 + \mathbf{m}_2 \rrbracket_\gamma(\mu') + \gamma I_\mu(\mathbf{Z}; \mathbf{X} | \mathbf{S}). \end{aligned}$$

But conditional mutual information is non-negative, and by assumption,  $\llbracket \mathbf{m}_1 + \mathbf{m}_2 \rrbracket_\gamma(\mu)$  is minimal. Therefore, it must be the case that

$$I_\mu(\mathbf{Z}; \mathbf{X} | \mathbf{S}) = I_\mu(\mathcal{X}_1; \mathcal{X}_2 | \mathcal{X}_1 \cap \mathcal{X}_2) = 0,$$

showing that  $\mathcal{X}_1$  and  $\mathcal{X}_2$  are conditionally independent given the variables that they have in common.

(The fact that  $I_\mu(\mathbf{Z}; \mathbf{X} | \mathbf{S}) = I_\mu(\mathcal{X}_1; \mathcal{X}_2 | \mathcal{X}_1 \cap \mathcal{X}_2)$  is both easy to show and an instance of a well-known identity; see CIRV2 in Theorem 4.4.4 of [Halpern \(2017\)](#), for instance.) □

**Corollary 8.6.1.** *If  $\mathbf{m}$  is a PDG with arcs  $\mathcal{A}$ ,  $(\mathcal{C}, \mathcal{T})$  is a tree decomposition of  $\mathcal{A}$ ,  $\gamma > 0$ , and  $\mu \in \llbracket \mathbf{m} \rrbracket_\gamma^*$ , then there exists a tree marginal  $\boldsymbol{\mu}$  over  $(\mathcal{C}, \mathcal{T})$  such that  $\Pr_{\boldsymbol{\mu}} = \mu$ .*

*Proof.* The set of distributions that can be represented by a calibrated tree marginal over  $(\mathcal{C}, \mathcal{T})$  is the same as the set of distributions that can be represented by a factor graph for which  $(\mathcal{C}, \mathcal{T})$  is a tree decomposition. One direction holds because any such product of factors “calibrated”, via message passing algorithms

such as belief propagation, to form a tree marginal. The other direction holds because  $\Pr_\mu$  itself is a product of factors that decomposes over  $(\mathcal{C}, \mathcal{T})$ .

Alternatively, this same set of distributions that satisfy the independencies of the Markov Network obtained by connecting every pair of variables that share a cluster. More formally, this network is the graph  $G := (\mathcal{X}, E := \{(X-Y) : \exists C \in \mathcal{C}. \{X, Y\} \subseteq C\})$ . Also,  $G$  happens to chordal as well, which we prove at the end.

Using only the PDG Markov property (Theorem 8.6), we now show that every independence described by  $G$  also holds in every distribution  $\mu \in \llbracket m \rrbracket_\gamma^*$ . Suppose that, for sets of variables  $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathcal{X}$ ,  $I(\mathbf{X}; \mathbf{Y} | \mathbf{Z})$  is an independence described by  $G$ . This means (Koller and Friedman 2009, Defn 4.8) that if  $X \in \mathbf{X}, Y \in \mathbf{Y}$ , and  $\pi$  is a path in  $G$  between them, then some node along  $\pi$  lies in  $\mathbf{Z}$ .

Let  $\mathcal{T}'$  be the graph that results from removing each edge  $(C-D) \in \mathcal{T}$  that satisfies  $C \cap D \subseteq \mathbf{Z}$ , which is a disjoint union  $\mathcal{T}' = \mathcal{T}_1 \sqcup \dots \sqcup \mathcal{T}_n$  of subtrees that have no clusters in common. To parallel this notation, let  $\mathcal{C}_1, \dots, \mathcal{C}_n$  be their respective vertex sets. Note that for every edge  $e = (C-D) \in \mathcal{T}'$ , there must by definition be some variable  $U_e \in (C \cap D) \setminus \mathbf{Z}$ .

We claim that no subtree  $\mathcal{T}_i$  can have both a cluster  $D_X$  containing a variable  $X \in \mathbf{X} \setminus \mathbf{Z}$  and also a cluster  $D_Y$  containing a variable  $Y \in \mathbf{Y} \setminus \mathbf{Z}$ . Suppose that it did. Then the (unique) path in  $\mathcal{T}$  between  $D_X$  and  $D_Y$ , which we label

$$D_X = D_0 \xrightarrow{e_1} D_1 \xrightarrow{e_2} \dots \xrightarrow{e_{m-1}} D_{m-1} \xrightarrow{e_m} D_m = D_Y ,$$

would lie entirely within  $\mathcal{T}_i \subseteq \mathcal{T}'$ . This gives rise to a corresponding path in  $G$ :

$$\begin{array}{ccccccccccc} X & \xlongequal{\quad} & U_{e_1} & \xlongequal{\quad} & U_{e_2} & \xlongequal{\quad} & \cdots & \xlongequal{\quad} & U_{e_{n-1}} & \xlongequal{\quad} & U_{e_n} & \xlongequal{\quad} & Y \\ \cap & & \cap & & \cap & & & & \cap & & \cap & & \cap & , \\ D_0 & & D_0 \cap D_1 & & D_1 \cap D_2 & & & & D_{n-2} \cap D_{n-1} & & D_{n-1} \cap D_n & & D_n \end{array}$$

and moreover, this path is disjoint from  $\mathbf{Z}$ . This contradicts our assumption that every path in  $G$  between a member of  $\mathbf{X}$  and a member of  $\mathbf{Y}$  must intersect with  $\mathbf{Z}$ , and so no subtree can have both a cluster containing a variable  $X \in \mathbf{X} \setminus \mathbf{Z}$  and also one containing  $Y \in \mathbf{Y} \setminus \mathbf{Z}$ .

We can now partition the clusters as  $\mathcal{C} = \mathcal{C}_{\mathbf{X}} \sqcup \mathcal{C}_{\mathbf{Y}}^+$ , where  $\mathcal{C}_{\mathbf{X}}$  is the set of the clusters that belong to subtrees  $\mathcal{T}_i$  with a cluster containing some  $X \in \mathbf{X} \setminus \mathbf{Z}$ , and its  $\mathcal{C}_{\mathbf{Y}}^+$  is its complement, which in particular contains those subtrees have some  $Y \in \mathbf{Y} \setminus \mathbf{Z}$ . Or, more formally, we define

$$\mathcal{C}_{\mathbf{X}} := \bigcup_{\substack{i \in \{1, \dots, n\} \\ (\cup \mathcal{C}_i) \cap (\mathbf{X} \setminus \mathbf{Z}) \neq \emptyset}} \mathcal{C}_i \quad \text{and} \quad \mathcal{C}_{\mathbf{Y}}^+ := \bigcup_{\substack{i \in \{1, \dots, n\} \\ (\cup \mathcal{C}_i) \cap (\mathbf{X} \setminus \mathbf{Z}) = \emptyset}} \mathcal{C}_i .$$

Let  $\mathcal{X}_{\mathbf{X}} := \cup \mathcal{C}_{\mathbf{X}}$  set of all variables appearing in the clusters  $\mathcal{C}_{\mathbf{X}}$ ; symmetrically, define  $\mathcal{X}_{\mathbf{Y}}^+ := \cup \mathcal{C}_{\mathbf{Y}}^+$ .

We claim that  $\mathcal{X}_{\mathbf{X}} \cap \mathcal{X}_{\mathbf{Y}}^+ \subset \mathbf{Z}$ . Choose any variable  $U \in \mathcal{X}_{\mathbf{X}} \cap \mathcal{X}_{\mathbf{Y}}^+$ . From the definitions of  $\mathcal{X}_{\mathbf{X}}$  and  $\mathcal{X}_{\mathbf{Y}}^+$ , this means  $U$  is a member of some cluster  $C \in \mathcal{C}_{\mathbf{X}}$ , and also a member of a cluster  $D \in \mathcal{C}_{\mathbf{Y}}^+$ . Recall that the clusters of each disjoint subtree  $\mathcal{T}_i$  either fall entirely within  $\mathcal{C}_{\mathbf{X}}$  or entirely within  $\mathcal{C}_{\mathbf{Y}}^+$  by construction. This means that  $C$  and  $D$ , which are on opposite sides of the partition, must have come from distinct subtrees. So, some edge  $e = (C' - D') \in \mathcal{T}$  along the (unique) path from  $C$  to  $D$  must have been removed when forming  $\mathcal{T}'$ , which by the definition of  $\mathcal{T}'$ , means that  $(C' \cap D') \subset Z$ . But by the running intersection property (tree marginal property 2), every cluster along the path from  $C$  to  $D$  must contain  $C \cap D$ —in particular, this must be true of both  $C'$  and  $D'$ . Therefore,

$$U \in C \cap D \subset C' \cap D' \subset \mathbf{Z}.$$

So  $\mathcal{X}_{\mathbf{X}} \cap \mathcal{X}_{\mathbf{Y}}^+ \subset \mathbf{Z}$ , as promised. We will rather use it in the equivalent form  $(\mathcal{X}_{\mathbf{X}} \cap \mathcal{X}_{\mathbf{Y}}^+) \cup \mathbf{Z} = \mathbf{Z}$ .

Next, since  $(\mathcal{C}, \mathcal{T})$  is a tree decomposition of  $\mathcal{A}$ , each hyperarc  $a \in \mathcal{A}$  can be assigned to some cluster  $C_a$  that contains all of its variables; this allows us to lift the cluster partition  $\mathcal{C} = \mathcal{C}_X \sqcup \mathcal{C}_Y^+$  to a partition  $\mathcal{A} = \mathcal{A}_X \sqcup \mathcal{A}_Y^+$  of edges, and consequently, a partition of PDGs  $m = m_X + m_Y^+$ . Concretely: let  $m_X$  be the sub-PDG of  $m$  induced by restricting to the variables  $X \subseteq \mathcal{X}$  arcs  $\mathcal{A}_X = \{a \in \mathcal{A} : C_a \in \mathcal{C}_X\} \subseteq \mathcal{A}$ ; define  $m_Y^+$  symmetrically. (To be explicit: the other data of  $m_X$  and  $m_Y^+$  are given by restricting each of  $\{\mathbb{P}, \alpha, \beta\}$  to  $\mathcal{A}_X$  and  $\mathcal{A}_Y^+$ , respectively.)

This partition of  $m$  allows us to use the PDG Markov property. Suppose that for some  $\gamma > 0$  that  $\mu \in \llbracket m \rrbracket_\gamma^* = \llbracket m_X + m_Y^+ \rrbracket_\gamma^*$ . We can then apply [Theorem 8.6](#), to find that  $X$  and  $Y$  are independent given  $X \cap Y$ . We use standard properties of random variable independence (CIRV1-5 of [Halpern 2017](#), Theorem 4.4.4) to find that  $\mu$  must satisfy:

$$\begin{aligned}
& X \perp\!\!\!\perp Y^+ \mid X \cap Y^+ \\
\implies & (X \setminus Z) \perp\!\!\!\perp (Y^+ \setminus Z) \mid (X \cap Y^+) \cup Z & [\text{CIRV3}] \\
\implies & (X \setminus Z) \perp\!\!\!\perp (Y \setminus Z) \mid (X \cap Y^+) \cup Z & \left[ \begin{array}{l} \text{by CIRV2, as} \\ X \subseteq X \cap Y^+ \end{array} \right] \\
\implies & (X \setminus Z) \perp\!\!\!\perp (Y \setminus Z) \mid Z & \left[ \text{since } (X \cap Y^+) \cup Z = Z \right] \\
\iff & X \perp\!\!\!\perp Y \mid Z & [\text{standard; e.g., Exercise 4.18 of } \text{Halpern (2017)}]
\end{aligned}$$

Using only the PDG Markov property, we have now shown that every independence modeled by the Markov Network  $G$  also holds in every distribution  $\mu \in \llbracket m \rrbracket_\gamma^*$ . Moreover,  $G$  is chordal (as we will prove momentarily), and it is well-known that distributions that have the independencies of a chordal graph can be represented by tree marginals ([Koller and Friedman 2009](#), Theorem 4.12).

Therefore, there is a tree marginal  $\mu$  representing every  $\mu \in \llbracket m \rrbracket_{\gamma}^*$ .

**Claim 8.10.1.** *G is chordal.*

*Proof.* Suppose that  $G$  contains a loop  $X - Y - Z - W - X$ . Suppose further, for contradiction, that neither  $X$  and  $Z$  nor  $Y$  and  $W$  share a cluster. Given a variable  $V$ , it is easy to see that property (2) of the tree decomposition ensures that the subtree  $\mathcal{T}(V) \subseteq \mathcal{T}$  induced by the clusters  $C \in \mathcal{C}$  that contain  $V$ , is connected. By assumption,  $\mathcal{T}(Y)$  and  $\mathcal{T}(W)$  must be disjoint. There is an edge between  $Y$  and  $Z$ , so some cluster must contain both variables, meaning  $\mathcal{T}(Y) \cap \mathcal{T}(Z)$  is non-empty. Similarly,  $\mathcal{T}(Z) \cap \mathcal{T}(W)$  is non-empty because of the edge between  $Z$  and  $W$ . This creates an (indirect) connection in  $\mathcal{T}$  between  $\mathcal{T}(Y)$  and  $\mathcal{T}(W)$ . Because  $\mathcal{T}$  is a tree, and  $\mathcal{T}(Y) \cap \mathcal{T}(W) = \emptyset$ , every path from a cluster  $C_1 \in \mathcal{T}(Y)$  to a cluster  $C_2 \in \mathcal{T}(W)$  must pass through  $\mathcal{T}(Z)$ , which is not part of  $\mathcal{T}(Y)$  or  $\mathcal{T}(W)$ .  $\mathcal{T}(X)$  and  $\mathcal{T}(Y)$  intersect as well, meaning that, for any  $C \in \mathcal{T}(X)$ , there is a (unique) path from  $C$  to that point of intersection, then across edges of  $\mathcal{T}(Y)$ , then edges of  $\mathcal{T}(Z)$ , and finally connects to the clusters of  $\mathcal{T}(W)$ . And also, since  $\mathcal{T}$  is a tree, that path must be unique. The problem is that there is also an edge between  $X$  and  $W$ , so there's some cluster that contains  $X$  and  $W$ ; let's call it  $C_0$ . It's distinct from the cluster  $D_0$  that contains  $Z$  and  $W$ , since no cluster contains both  $X$  and  $Z$  by assumption. The unique path from  $C_0$  to  $D_0$  intersects with  $\mathcal{T}(Y)$ . But now  $W \in C_0 \cap D_0$ , and by the running intersection property, every node along this unique path must contain  $W$  as well. But this contradicts our assumption that  $W$  is disjoint from  $Y$ ! So  $G$  is chordal.  $\square$

Having proved the subclaim [Claim 8.10.1](#), we have now finished the proof of [Corollary 8.6.1](#).  $\square$

### 8.A.2 Correctness and Complexity Analysis for PDG Inference via Exponential Conic Programming

**Proposition 8.1.** If  $(\mu, \mathbf{u})$  is a solution to (8.4), then  $\mu \in \llbracket \mathcal{M} \rrbracket_0^*$ , and

$$\sum_{(a,s,t) \in \mathcal{VA}} \beta_a u_{a,s,t} = \langle\!\langle \mathcal{M} \rangle\!\rangle_0.$$

*Proof.* Suppose that  $(\mu, \mathbf{u})$  is a solution to (8.4). The exponential cone constraints ensure that, for every  $(a, s, t) \in \mathcal{VA}$ ,

$$u_{a,s,t} \geq \mu(s, t) \log \frac{\mu(s, t)}{\mathbb{P}_a(t|s)\mu(s)},$$

where  $\mu(s, t)$  and  $\mu(s)$ , as usual, are shorthand for  $\mu(S_a=s, T_a=t)$  and  $\mu(S_a=s)$ , respectively. Suppose, for contradiction, that one of these inequalities is strict at some index  $(a', s', t') \in \mathcal{VA}$  for which  $\beta_{a'} > 0$ . Explicitly, this means

$$u_{a',s',t'} > \mu(s_0, t_0) \log \frac{\mu(s', t')}{\mathbb{P}_{a'}(t'|s')\mu(s')}.$$

In that case, we can define a vector  $\mathbf{u}' = [u'_{a,s,t}]_{(a,s,t) \in \mathcal{VA}}$  which is identical to  $\mathbf{u}$ , except that at  $(a', s', t')$ , it is halfway between the two quantities described as different above. More precisely:

$$u'_{a',s',t'} = \frac{1}{2}u_{a',s',t'} + \frac{1}{2}\log\mu(s', t') \log \frac{\mu(s', t')}{\mathbb{P}_a(t'|s')\mu(s')}.$$

Note that  $u'_{a',s',t'} < u_{a',s',t'}$ , and also that, by construction,  $(\mu, \mathbf{u}')$  also satisfies the constraints of (8.4). In more detail: at the index  $(a', s', t')$ ,  $\mathbf{u}'$  does not violate the associated exponential cone constraint

$$\left( \text{because } u'_{a',s',t'} = \frac{1}{2}u_{a',s',t'} + \frac{1}{2}\log\mu(s', t') \log \frac{\mu(s', t')}{\mathbb{P}_{a'}(t'|s')\mu(s')} > \mu(s', t') \log \frac{\mu(s', t')}{\mathbb{P}_{a'}(t'|s')\mu(s')} \right),$$

and  $\mathbf{u}'$  equals  $\mathbf{u}$  at other indices, and therefore satisfies the constraint everywhere else as well. But now, because  $u'_{a',s',t'} < u_{a',s',t'}$ , and  $\beta_{a'} > 0$ , we also have

$$\sum_{(a,s,t) \in \mathcal{VA}} \beta_a u'_{a,s,t} > \sum_{(a,s,t) \in \mathcal{VA}} \beta_a u'_{a,s,t}.$$

Thus the objective value at  $(\mu, \mathbf{u}')$  is strictly smaller than the one at  $(\mu, \mathbf{u})$ , both of which are feasible points. This contradicts the assumption that  $(\mu, \mathbf{u})$  is optimal. We therefore conclude that none of these inequalities can be strict at points where  $\beta_a > 0$ . This can be compactly written as:

$$\begin{aligned} \forall (a, s, t) \in \mathcal{VA}. \quad \beta_a u_{a,s,t} &= \beta_a \mu(s, t) \log \frac{\mu(s, t)}{\mathbb{P}_a(t|s)\mu(s)} \\ \implies \sum_{(a,s,t) \in \mathcal{VA}} \beta_a u_{a,s,t} &= \sum_{(a,s,t) \in \mathcal{VA}} \beta_a \mu(s, t) \log \frac{\mu(s, t)}{\mathbb{P}_a(t|s)\mu(s)} = OInc_m(\mu). \end{aligned}$$

In other words, the objective of problem (8.4) at  $(\mu, \mathbf{u})$  is equal to the observational incompatibility  $OInc_m(\mu)$  of  $\mu$  with  $\mathbf{m}$ . And, because  $(\mu, \mathbf{u})$  minimizes this value among all joint distributions,  $\mu$  must be a minimum of  $OInc_m$ .

More formally: assume for contradiction that  $\mu$  is not a minimizer of  $OInc_m$ . Then there would be some other distribution  $\mu'$  for which  $OInc_m(\mu') < OInc_m(\mu)$ . Let  $\mathbf{u}'' := [\mu'(s, t) \log \frac{\mu'(s, t)}{\mathbb{P}_a(t|s)\mu'(s)}]_{(a,s,t) \in \mathcal{VA}}$ . Clearly  $(\mu', \mathbf{u}'')$  satisfies the constraints of the problem, and moreover,

$$\sum_{(a,s,t) \in \mathcal{VA}} \beta_a u_{a,s,t} = OInc_m(\mu) > OInc_m(\mu') = \sum_{(a,s,t) \in \mathcal{VA}} \beta_a u'_{a,s,t},$$

contradicting the assumption that the  $(\mu, \mathbf{u})$  is optimal for problem (8.4). Thus,  $\mu$  is a minimizer of  $OInc_m$ , and the objective value is  $\inf_\mu OInc_m(\mu) = \langle\!\langle \mathbf{m} \rangle\!\rangle_0$ , as desired.  $\square$

**Proposition 8.2.** *If  $(\mu, \mathbf{u}, \mathbf{v})$  is a solution to (8.6), and  $\beta \geq \gamma\alpha$ , then  $\mu$  is the unique element of  $[\mathbf{m}]_\gamma^*$ , and  $\langle\!\langle \mathbf{m} \rangle\!\rangle_\gamma$  equals the objective of (8.6) evaluated at  $(\mu, \mathbf{u}, \mathbf{v})$ .*

For convenience, we repeat problem (8.6) (left) and an equivalent variant of it that we implement (right) below.

$$\underset{\mu, \mathbf{u}, \mathbf{v}}{\text{minimize}} \quad \sum_{(a,s,t) \in \mathcal{VA}} (\beta_a - \alpha_a \gamma) u_{a,s,t} + \gamma \sum_{w \in \mathcal{VX}} v_w \quad (8.6)$$

$$- \sum_{(a,s,t) \in \mathcal{VA}^+} \alpha_a \gamma \mu(S_a=s, T_a=t) \log \mathbb{P}_a(t|s)$$

subject to  $\mu \in \Delta \mathcal{VX}$ ,  $(-\mathbf{v}, \mu, \mathbf{1}) \in K_{\text{exp}}^{\mathcal{VX}}$ ,

$$\forall a \in \mathcal{A}. (-\mathbf{u}_a, \mu(T_a, S_a), \mathbb{P}_a(T_a|S_a)\mu(S_a)) \in K_{\text{exp}}^{\mathcal{V}a},$$

$$\forall (a, s, t) \in \mathcal{VA}^0. \mu(S_a=s, T_a=t) = 0;$$

$$\underset{\mu, \mathbf{u}, \mathbf{v}}{\text{minimize}} \quad \sum_{(a,s,t) \in \mathcal{VA}} (\beta_a - \alpha_a \gamma) u_{a,s,t} + \gamma \sum_{w \in \mathcal{VX}} v_w \quad (8.6b)$$

$$- \sum_{(a,s,t) \in \mathcal{VA}^+} \beta_a \mu(S_a=s, T_a=t) \log \mathbb{P}_a(t|s)$$

subject to  $\mu \in \Delta \mathcal{VX}$ ,  $(-\mathbf{v}, \mu, \mathbf{1}) \in K_{\text{exp}}^{\mathcal{VX}}$ ,

$$\forall a \in \mathcal{A}. (-\mathbf{u}_a, \mu(T_a, S_a), [\mu(S_a=s)]_{(s,t) \in \mathcal{V}a}) \in K_{\text{exp}}^{\mathcal{V}a},$$

$$\forall (a, s, t) \in \mathcal{VA}^0. \mu(S_a=s, T_a=t) = 0.$$

*Proof.* We start with the problem on the left, which is (8.6) from the main text.

Suppose that  $(\mu, \mathbf{u}, \mathbf{v})$  is a solution to (8.6). The exponential constraints ensure that

$$\forall (a, s, t) \in \mathcal{VA}. u_{a,s,t} \geq \mu(s, t) \log \frac{\mu(t|s)}{\mathbb{P}_a(t|s)} \quad \text{and} \quad \forall w \in \mathcal{VX}. v_w \geq \mu(w) \log \mu(w).$$

As in the previous proof, we claim that these must hold with equality (except possibly for  $u_{a,s,t}$  at indices satisfying  $\beta_a = \gamma \alpha_a$ , when it doesn't matter). This is because otherwise one could reduce the value of a component of  $u$  or  $v$  while still satisfying all of the constraints, to obtain a strictly smaller objective, contradicting the assumption that  $(\mu, \mathbf{u}, \mathbf{v})$  minimizes it.

Thus,  $\mathbf{v}$  is a function of  $\mu$ , as is every value of  $\mathbf{u}$  that affects the objective value

of (8.6), meaning that this objective value can be written as a function of  $\mu$  alone:

$$\begin{aligned}
& \sum_{(a,s,t) \in \mathcal{VA}} (\beta_a - \alpha_a \gamma) u_{a,s,t} + \gamma \sum_{w \in \mathcal{VX}} v_w - \sum_{(a,s,t) \in \mathcal{VA}^+} \alpha_a \gamma \mu(s, t) \log \mathbb{P}_a(t|s) \\
&= \sum_{(a,s,t) \in \mathcal{VA}} (\beta_a - \alpha_a \gamma) \left[ \mu(s, t) \log \frac{\mu(t|s)}{\mathbb{P}_a(t|s)} \right] + \gamma \sum_{w \in \mathcal{VX}} \mu(w) \log \mu(w) - \sum_{(a,s,t) \in \mathcal{VA}^+} \alpha_a \gamma \mu(s, t) \log \mathbb{P}_a(t|s) \\
&= \sum_{a \in \mathcal{A}} (\beta_a - \alpha_a \gamma) \sum_{(s,t) \in \mathcal{VA}} \mu(s, t) \log \frac{\mu(t|s)}{\mathbb{P}_a(t|s)} - \gamma H(\mu) - \sum_{a \in \mathcal{A}} \alpha_a \gamma \sum_{(s,t) \in \mathcal{VA}} \mu(s, t) \log \mathbb{P}_a(t|s) \\
&= \sum_{a \in \mathcal{A}} (\beta_a - \alpha_a \gamma) \sum_{(s,t) \in \mathcal{VA}} \mu(s, t) \left[ \log \frac{1}{\mathbb{P}_a(t|s)} - \log \frac{1}{\mu(t|s)} \right] - \gamma H(\mu) - \sum_{\substack{S \xrightarrow{a} T \in \mathcal{A}}} \alpha_a \gamma \mathbb{E}_{\mu} \log \mathbb{P}_a(T|S) \\
&= \sum_{a \in \mathcal{A}} (\beta_a - \alpha_a \gamma) \mathbb{E}_{\mu} [-\log \mathbb{P}_a(T_a|S_a)] - \sum_{a \in \mathcal{A}} (\beta_a - \alpha_a \gamma) H_{\mu}(T_a|S_a) \\
&\quad - \gamma H(\mu) - \sum_{a \in \mathcal{A}} \alpha_a \gamma \mathbb{E}_{\mu} [\log \mathbb{P}_a(T_a|S_a)] \\
&= \sum_{a \in \mathcal{A}} \left( -\alpha_a \gamma - (\beta_a - \alpha_a \gamma) \right) \mathbb{E}_{\mu} [\log \mathbb{P}_a(T_a|S_a)] + \sum_{a \in \mathcal{A}} (\alpha_a \gamma - \beta_a) H_{\mu}(T_a|S_a) - \gamma H(\mu) \\
&= - \sum_{a \in \mathcal{A}} \beta_a \mathbb{E}_{\mu} [\log \mathbb{P}_a(T_a|S_a)] + \sum_{a \in \mathcal{A}} (\alpha_a \gamma - \beta_a) H_{\mu}(T_a|S_a) - \gamma H(\mu).
\end{aligned}$$

( In the third step, we were able to convert  $\mathcal{VA}^+$  to  $\mathcal{VA}$  because, as usual in when dealing with information-theoretic quantities, we take  $0 \log \frac{1}{0}$  to equal zero, which is its limit. )

The algebra for the right side variant (8.6b) is slightly simpler. In this case the middle conic constraint is almost the same, except for that  $\mathbb{P}_a(t|s)$  has been replaced with 1, and so it ensures that  $u_{a,s,t} = \mu(s, t) \log \mu(t | s)$  (i.e., the same as

before, but without the probability in the denominator). So,

$$\begin{aligned}
& \sum_{(a,s,t) \in \mathcal{VA}} (\beta_a - \alpha_a \gamma) u_{a,s,t} + \gamma \sum_{w \in \mathcal{VX}} v_w - \sum_{(a,s,t) \in \mathcal{VA}^+} \beta_a \mu(s, t) \log \mathbb{P}_a(t|s) \\
&= \sum_{(a,s,t) \in \mathcal{VA}} (\beta_a - \alpha_a \gamma) \mu(s, t) \log \mu(t|s) + \gamma \sum_{w \in \mathcal{VX}} \mu(w) \log \mu(w) - \sum_{(a,s,t) \in \mathcal{VA}^+} \beta_a \mu(s, t) \log \mathbb{P}_a(t|s) \\
&= \sum_{a \in \mathcal{A}} (\beta_a - \alpha_a \gamma) \sum_{(s,t) \in \mathcal{VA}} \mu(s, t) \log \mu(t|s) - \gamma H(\mu) - \sum_{a \in \mathcal{A}} \beta_a \sum_{(s,t) \in \mathcal{VA}} \mu(s, t) \log \mathbb{P}_a(t|s) \\
&= \sum_{a \in \mathcal{A}} (\alpha_a \gamma - \beta_a) H_\mu(T_a | S_a) - \gamma H(\mu) - \sum_{a \in \mathcal{A}} \beta_a \mathbb{E}_\mu [\log \mathbb{P}_a(T_a | S_a)].
\end{aligned}$$

In either case, the objective value is equal to  $\llbracket m \rrbracket_\gamma(\mu)$ , by (8.5). Because  $(\mu, \mathbf{u}, \mathbf{v})$  is optimal for this problem, we know that  $\mu$  is a minimizer of  $\llbracket m \rrbracket_\gamma(\mu)$ , and that the objective value equals  $\langle\!\langle m \rangle\!\rangle_\gamma$ .  $\square$

**Lemma 8.11.** *The gradient and Hessian of conditional relative entropy are given by*

$$\begin{aligned}
\left[ \nabla_\mu D(\mu(X, Y) \| \mu(X)p(Y|X)) \right]_u &= \log \frac{\mu(Yu|Xu)}{p(Yu|Xu)} \\
\left[ \nabla_\mu^2 D(\mu(X, Y) \| \mu(X)p(Y|X)) \right]_{u,v} &= \frac{\mathbb{1}[Xu=Xv \wedge Yu=Yv]}{\mu(Yu, Xu)} - \frac{\mathbb{1}[Xv=Xu]}{\mu(Xu)},
\end{aligned}$$

where  $Xu = X(u)$  it the value of the variable  $X$  in the joint setting  $u \in \mathcal{VX}$  of all variables.

*Proof.* Represent  $\mu$  as a vector  $[\mu_w]_{w \in \mathcal{VX}}$ . We will make repeated use of the following facts:

$$\begin{aligned}
\frac{\partial}{\partial \mu_u} [\mu(X=x)] &= \frac{\partial}{\partial \mu_u} [\mu(x)] = \sum_w \frac{\partial}{\partial \mu_u} [\mu_w] \mathbb{1}[Xw=x] = \mathbb{1}[Xu=x]; \quad \text{and} \\
\frac{\partial}{\partial \mu_u} [\mu(y|x)] &= \frac{\partial}{\partial \mu_u} \left[ \frac{\mu(x,y)}{\mu(x)} \right] \\
&= \mu(x,y) \frac{\partial}{\partial \mu_u} \left[ \frac{1}{\mu(x)} \right] + \frac{1}{\mu(x)} \frac{\partial}{\partial \mu_u} [\mu(x,y)] \\
&= -\mu(x,y) \frac{\mathbb{1}[Xu=x]}{\mu(x)^2} + \frac{1}{\mu(x)} \mathbb{1}[Xu=x \wedge Yu=y]
\end{aligned}$$

$$= \frac{\mathbb{1}[Xu = x]}{\mu(x)} \left( \mathbb{1}[Yu = y] - \mu(y|x) \right).$$

We now apply this to the (conditional) relative entropy:

$$\begin{aligned}
& \frac{\partial}{\partial \mu_u} \left[ D(\mu(X, Y) \parallel \mu(X)p(Y|X)) \right] \\
&= \frac{\partial}{\partial \mu_u} \left[ \sum_w \mu_w \log \frac{\mu(Yw|Xw)}{p(Yw|Xw)} \right] \\
&= \sum_w \mathbb{1}[u=w] \log \frac{\mu(Yw|Xw)}{p(Yw|Xw)} + \sum_w \mu_w \frac{\partial}{\partial \mu_u} \left[ \log \frac{\mu(Yw|Xw)}{p(Yw|Xw)} \right] \\
&= \log \frac{\mu(Yu|Xu)}{p(Yu|Xu)} + \sum_w \mu_w \frac{p(Yw|Xw)}{\mu(Yw|Xw)} \frac{\partial}{\partial \mu_u} \left[ \log \frac{\mu(Yw|Xw)}{p(Yw|Xw)} \right] \\
&= \log \frac{\mu(Yu|Xu)}{p(Yu|Xu)} + \sum_w \mu_w \frac{1}{\mu(Yw|Xw)} \frac{\partial}{\partial \mu_u} [\mu(Yw|Xw)] \\
&= \log \frac{\mu(Yu|Xu)}{p(Yu|Xu)} + \sum_w \mu_w \frac{1}{\mu(Yw|Xw)} \frac{\mathbb{1}[Xu = Xw]}{\mu(Xw)} \left( \mathbb{1}[Yu = Yw] - \mu(Yw|Xw) \right) \\
&= \log \frac{\mu(Yu|Xu)}{p(Yu|Xu)} + \sum_w \mu_w \frac{\mathbb{1}[Xu = Xw \wedge Yu = Yw]}{\mu(Xw, Yw)} - \sum_w \mu_w \frac{\mathbb{1}[Xu = Xw]}{\mu(Xw)} \\
&= \log \frac{\mu(Yu|Xu)}{p(Yu|Xu)} + \frac{1}{\mu(Xu, Yu)} \sum_w \mu_w \mathbb{1}[Xu = Xw \wedge Yu = Yw] - \frac{1}{\mu(Xu)} \sum_w \mu_w \mathbb{1}[Xu = Xw] \\
&= \log \frac{\mu(Yu|Xu)}{p(Yu|Xu)} + \frac{\mu(Xu, Yu)}{\mu(Xu, Yu)} - \frac{\mu(Xu)}{\mu(Xu)} \\
&= \log \frac{\mu(Yu|Xu)}{p(Yu|Xu)} + 1 - 1 \\
&= \log \frac{\mu(Yu|Xu)}{p(Yu|Xu)}.
\end{aligned}$$

This allows us to compute the Hessian of the conditional relative entropy, whose components are

$$\begin{aligned}
\frac{\partial^2}{\partial \mu_u \partial \mu_v} \left[ D(\mu(XY) \parallel \mu(X)p(Y|X)) \right] &= \frac{\partial}{\partial \mu_v} \left[ \log \frac{\mu(Yu|Xu)}{p(Yu|Xu)} \right] \\
&= \frac{p(Yu|Xu)}{\mu(Yu|Xu)} \frac{1}{p(Yu|Xu)} \frac{\partial}{\partial \mu_v} [\mu(Yu|Xu)] \\
&= \frac{1}{\mu(Yu|Xu)} \frac{\mathbb{1}[Xv = Xu]}{\mu(Xu)} \left( \mathbb{1}[Yv = Yu] - \mu(Yu|Xu) \right)
\end{aligned}$$

$$= \frac{\mathbb{1}[Xu=Xv \wedge Yu=Yv]}{\mu(Yu, Xu)} - \frac{\mathbb{1}[Xv = Xu]}{\mu(Xu)}. \quad \square$$

**Lemma 8.12.** Let  $p(Y|X)$  be a cpd, and suppose that  $\mu_0, \mu_1 \in \Delta\mathcal{V}(X, Y)$  are joint distributions that have different conditional marginals on  $Y$  given  $X$ ; that is, that there exist  $(x, y) \in \mathcal{V}(X, Y)$  such that  $\mu_0(x, y)\mu_1(x) \neq \mu_1(x, y)\mu_0(x)$ . Then the conditional relative entropy  $D(\mu(X, Y) \parallel \mu(X)p(Y|X))$  is strictly convex in  $\mu$  along the line segment from  $\mu_0$  to  $\mu_1$ . More precisely, for  $t \in [0, 1]$ , if we define  $\mu_t := (1 - t)\mu_0 + t\mu_1$ , then the function

$$t \mapsto D(\mu_t(X, Y) \parallel \mu_t(X)p(Y|X)) \quad \text{is strictly convex.}$$

*Proof.* The function of interest can fail to be strictly convex only if the direction  $\delta$  along  $\mu_1 - \mu_0$  is in the null-space of the Hessian matrix  $\mathbf{H}(\mu)$  of the (conditional) relative entropy. By Lemma 8.11,

$$\mathbf{H}_{(xy), (x'y')} = \frac{\mathbb{1}[x=x' \wedge y=y']}{\mu(x, y)} - \frac{\mathbb{1}[x=x']}{\mu(x)}.$$

Consider a function  $\delta : \mathcal{V}(X, Y) \rightarrow \mathbb{R}$  that is not identically zero, which can be viewed as a vector  $\boldsymbol{\delta} = [\delta(x, y)]_{(x,y) \in \mathcal{V}(X, Y)} \in \mathbb{R}^{\mathcal{V}(X, Y)}$ . We can also view  $\delta$  as a (signed) measure on  $\mathcal{V}(X, Y)$ , that has marginals in the usual sense. In particular, we use the analogous notation

$$\delta(x) := \sum_{y \in \mathcal{V}Y} \delta(x, y).$$

We then compute

$$\begin{aligned} (\mathbf{H}(\mu) \boldsymbol{\delta})_{x,y} &= \sum_{x',y'} \delta(x', y') \left( \frac{\mathbb{1}[x=x' \wedge y=y']}{\mu(x, y)} - \frac{\mathbb{1}[x=x']}{\mu(x)} \right) \\ &= \frac{\delta(x, y)}{\mu(x, y)} - \frac{\delta(x)}{\mu(x)}. \end{aligned}$$

and also

$$\begin{aligned}
\boldsymbol{\delta}^\top \mathbf{H}(\mu) \boldsymbol{\delta} &= \sum_{x,y} \delta(x,y) (\mathbf{H}(\mu) \boldsymbol{\delta})_{x,y} \\
&= \sum_{x,y} \delta(x,y) \left( \frac{\delta(x,y)}{\mu(x,y)} - \frac{\delta(x)}{\mu(x)} \right) \\
&= \sum_{x,y} \frac{\delta(x,y)^2}{\mu(x,y)} - \sum_x \frac{\delta(x)}{\mu(x)} \sum_y \delta(x,y) \\
&= \sum_{x,y} \frac{\delta(x,y)^2}{\mu(x,y)} - \sum_x \frac{\delta(x)^2}{\mu(x)} \\
&= \sum_x \frac{\delta(x)^2}{\mu(x)} \left( \sum_y \frac{\delta(x,y)^2}{\delta(x)^2 \mu(y|x)} - 1 \right). \tag{8.15}
\end{aligned}$$

Now consider another discrete measure  $|\delta|$ , whose value at each component is the absolute value of the value of  $\delta$  at that component, i.e.,  $|\delta|(x,y) := |\delta(x,y)|$ . By construction,  $|\delta|$  is now an unnormalized probability measure:  $|\delta| = kq(X,Y)$ , where  $k = \sum_{x,y} |\delta(x,y)| > 0$  and  $q \in \Delta\mathcal{V}(X,Y)$ .

Note also that  $|\delta|(x)^2 = (\sum_y |\delta(x,y)|)^2 \geq (\sum_y \delta(x,y))^2$ , and strictly so if there are  $y, y'$  such that  $\delta(x,y) < 0 < \delta(x,y')$ . In other words, the vector  $\boldsymbol{\delta}_x = [\delta(x,y)]_{y \in \mathcal{V}Y}$  is either non-negative or non-positive:  $\boldsymbol{\delta}_x \geq 0$  or  $\boldsymbol{\delta}_x \leq 0$  for each  $x$ . Meanwhile,  $|\delta|(x,y)^2 = \delta(x,y)^2$  is unchanged. Thus, for every  $x \in \mathcal{V}X$ , we have:

$$\begin{aligned}
\sum_y \frac{\delta(x,y)^2}{\delta(x)^2 \mu(y|x)} - 1 &\geq \sum_y \frac{|\delta|(x,y)^2}{|\delta|(x)^2 \mu(y|x)} - 1 \\
&= \sum_y \frac{k^2 q(x,y)^2}{k^2 q(x)^2 \mu(y|x)} - 1 \\
&= \sum_y \frac{q(y|x)^2}{\mu(y|x)} - 1 \\
&= \chi^2(q(Y|x) \parallel \mu(Y|x)) \geq 0.
\end{aligned}$$

The final line depicts the  $\chi^2$  divergence between the distributions  $q(Y|x)$  and

$\mu(Y|x)$ , both distributions over  $Y$ . Since it is a divergence, this quantity is non-negative and equals zero if and only if  $q(Y|x) = \mu(Y|x)$ .

Picking up where we left off, we have:

$$\begin{aligned}\boldsymbol{\delta}^\top \mathbf{H}(\mu) \boldsymbol{\delta} &= \sum_x \frac{\delta(x)^2}{\mu(x)} \left( \sum_y \frac{\delta(x,y)^2}{\delta(x)^2 \mu(y|x)} - 1 \right) \\ &\geq \sum_x \frac{\delta(x)^2}{\mu(x)} \left( \sum_y \frac{|\delta|(x,y)^2}{|\delta|(x)^2 \mu(y|x)} - 1 \right) \\ &= \sum_x \frac{\delta(x)^2}{\mu(x)} \chi^2(q(Y|x) \parallel \mu(Y|x)) \geq 0.\end{aligned}$$

As a non-negatively weighted sum of non-negative numbers, this final quantity is non-negative, and equals zero if and only if, for each  $x \in \mathcal{V}X$ , we have either  $q(Y|x) = \mu(Y|x)$ , or  $\delta(x) = 0$ . Furthermore, if  $\boldsymbol{\delta}^\top \mathbf{H}(\mu) \boldsymbol{\delta} = 0$ , then *both* inequalities hold with equality. Therefore, we know that if  $\delta(x) \neq 0$ , then  $\delta_x \geq 0$  or  $\delta_x \leq 0$ . These two conditions are also sufficient to show that  $\boldsymbol{\delta}^\top \mathbf{H}(\mu) \boldsymbol{\delta} = 0$ . To summarize what we know so far:

$$\begin{aligned}\boldsymbol{\delta}^\top \mathbf{H}(\mu) \boldsymbol{\delta} = 0 \iff \forall x \in \mathcal{V}X. \text{ either } (\delta_x \geq 0 \text{ or } \delta_x \leq 0) \text{ and } |\delta|(Y|x) = \mu(Y|x), \\ \text{or } \delta(x) = 0.\end{aligned}$$

The second possibility, however, is a mirage: it cannot occur. Let's now return to the expression we had in (8.15) before considering  $|\delta|$ . We've already shown that the contribution to the sum at each value of  $x$  is non-negative, so if  $\boldsymbol{\delta}^\top \mathbf{H}(\mu) \boldsymbol{\delta}$  is equal to zero, each summand which depends on  $x$  must be zero as well. So if  $x$  is a value of  $X$  for which  $\delta(x) = 0$ , then

$$0 = \frac{1}{\mu(x)} \left( \sum_y \frac{\delta(x,y)^2}{\mu(y|x)} - \delta(x)^2 \right) = \frac{1}{\mu(x)} \sum_y \frac{\delta(x,y)^2}{\mu(y|x)} = \sum_y \frac{\delta(x,y)^2}{\mu(x,y)},$$

which is only possible if  $\delta(x,y) = 0$  for all  $y$ . This allows us to compute, more

simply, that

$$\boldsymbol{\delta}^\top \mathbf{H}(\mu) \boldsymbol{\delta} = 0 \iff \begin{aligned} & (\forall x. \boldsymbol{\delta}_x \geq \mathbf{0} \text{ or } \boldsymbol{\delta}_x \leq \mathbf{0}) \\ & \text{and } \forall (x, y) \in \mathcal{V}(X, Y). \delta(x, y)\mu(x) = \delta(x)\mu(x, y) \end{aligned}$$

Finally, we are in a position to prove the lemma. Suppose that  $\mu_0, \mu_1 \in \Delta\mathcal{V}(X, Y)$  and  $(x^*, y^*) \in \mathcal{V}(X, Y)$  are such that  $\mu_0(x^*, y^*)\mu_1(x^*) \neq \mu_1(x^*, y^*)\mu_0(x^*)$ . So, the quantity

$$gap := \mu_1(x^*, y^*)\mu_0(x^*) - \mu_0(x^*, y^*)\mu_1(x^*) \quad \text{is nonzero.}$$

Then for all  $t \in (0, 1)$  the intermediate point  $\mu_t = (1-t)\mu_0 + t\mu_1$  must have different conditional marginals from both  $\mu_0$  and  $\mu_1$ , as

$$\begin{aligned} & \mu_t(x^*, y^*)\mu_0(x^*) - \mu_0(x^*, y^*)\mu_t(x^*) \\ &= \cancel{(1-t)\mu_0(x^*, y^*)\mu_0(x^*)} + t\mu_1(x^*, y^*)\mu_0(x^*) - \cancel{(1-t)\mu_0(x^*, y^*)\mu_0(x^*)} - t\mu_0(x^*, y^*)\mu_1(x^*) \\ &= t(\mu_1(x^*, y^*)\mu_0(x^*) - \mu_0(x^*, y^*)\mu_1(x^*)) \\ &= t \cdot gap \neq 0, \end{aligned}$$

and analogously for  $\mu_1$ ,

$$\begin{aligned} & \mu_t(x^*, y^*)\mu_1(x^*) - \mu_1(x^*, y^*)\mu_t(x^*) \\ &= \cancel{(1-t)\mu_0(x^*, y^*)\mu_1(x^*)} + \cancel{t\mu_1(x^*, y^*)\mu_1(x^*)} - (1-t)\mu_1(x^*, y^*)\mu_0(x^*) - \cancel{t\mu_1(x^*, y^*)\mu_1(x^*)} \\ &= (1-t)(\mu_0(x^*, y^*)\mu_1(x^*) - \mu_1(x^*, y^*)\mu_0(x^*)) \\ &= -(1-t) \cdot gap \neq 0. \end{aligned}$$

Then for any direction  $\delta := k(\mu_0 - \mu_1)$  parallel to the segment between  $\mu_0$  and  $\mu_1$  (intuitively a tangent vector at  $\mu_t$ , although this fact doesn't affect the

computation), of nonzero length ( $k \neq 0$ ), we have:

$$\begin{aligned}
& \mu_t(x^*, y^*)\delta(x^*) - \delta(x^*, y^*)\mu_t(x^*) \\
&= k \mu_t(x^*, y^*) (\mu_0(x^*) - \mu_1(x^*)) - k (\mu_0(x^*, y^*) - \mu_1(x^*, y^*))\mu_t(x^*) \\
&= k \left( \mu_t(x^*, y^*)\mu_0(x^*) - \mu_t(x^*, y^*)\mu_1(x^*) - \mu_0(x^*, y^*)\mu_t(x^*) + \mu_1(x^*, y^*)\mu_t(x^*) \right) \\
&= k \left( (\mu_t(x^*, y^*)\mu_0(x^*) - \mu_0(x^*, y^*)\mu_t(x^*)) + (\mu_1(x^*, y^*)\mu_t(x^*) - \mu_t(x^*, y^*)\mu_1(x^*)) \right) \\
&= k (+t \text{gap} + (1-t) \text{gap}) \\
&= k \text{gap} \quad \neq 0.
\end{aligned}$$

So at every  $t$ , directions parallel to the segment are not in the null space of  $\mathbf{H}(\mu_t)$ , meaning that  $\boldsymbol{\delta}^\top \mathbf{H}(\mu_t) \boldsymbol{\delta} > 0$  and so our function is strictly convex along this segment.  $\square$

**Proposition 8.5.** *If  $\nu \in [\![\mathcal{M}]\!]_0^*$  and  $(\mu, \mathbf{u})$  solves the problem*

$$\begin{aligned}
& \underset{\mu, \mathbf{u}}{\text{minimize}} \quad \mathbf{1}^\top \mathbf{u} \tag{8.9} \\
& \text{subject to} \quad (-\mathbf{u}, \mu, \mathbf{k}) \in K_{\text{exp}}^{\mathcal{V}\mathcal{X}}, \quad \mu \in \Delta \mathcal{V}\mathcal{X}, \\
& \quad \forall S \xrightarrow{a} T \in \mathcal{A}. \quad \mu(S, T) \nu(S) = \mu(S) \nu(S, T),
\end{aligned}$$

then  $[\![\mathcal{M}]\!]_{0+}^* = \{\mu\}$  and  $\mathbf{1}^\top \mathbf{u} = S\text{Def}_m(\mu)$ .

*Proof.* Suppose that  $(-\mathbf{u}, \mu, \mathbf{k})$  is a solution to problem (8.9). The second constraint, by Proposition 8.3, ensures that  $\mu \in [\![\mathcal{M}]\!]_0^*$ . Then

$$\begin{aligned}
(-\mathbf{u}, \mu, \mathbf{k}) \in K^{\mathcal{V}\mathcal{X}} \implies \forall w \in \mathcal{V}\mathcal{X}. \quad u_w \geq \mu(w) \log \frac{\mu(w)}{k_w} \\
= \mu(w) \log \left( \mu(w) / \prod_{a \in \mathcal{A}} \mu(T_a(w) | S_a(w))^{\alpha_a} \right).
\end{aligned}$$

The same logic as in the proofs of Propositions 8.1 and 8.2 shows that this inequality must be tight, or else  $(-\mathbf{u}, \mu, \mathbf{k})$  would not be optimal for (8.9). So,  $\mathbf{u}$  is a function of  $\mu$ . Also, by Proposition 8.4, the problem objective satisfies

$$\mathbf{1}^\top \mathbf{u} = \sum_{w \in V\mathcal{X}} u_w = SDef_m(\mu).$$

Finally, because  $\mu$  is optimal, it must be the unique distribution  $[\![\mathbf{m}]\!]^*$ , which among those distributions that minimize  $OInc_m$ , also minimizes  $SDef_m$ , meaning  $\mu = [\![\mathbf{m}]\!]^*$ . □

**Proposition 8.7.** *If  $(\mu, \mathbf{u})$  is a solution to (8.10), then*

- (a)  $\mu$  is a calibrated, with  $\Pr_\mu \in [\![\mathbf{m}]\!]_0^*$ , and
- (b) the objective of (8.10) evaluated at  $\mathbf{u}$  equals  $\langle\!\langle \mathbf{m} \rangle\!\rangle_0$ .

*Proof.* The final constraints alone are enough to ensure that  $\mu$  is calibrated. Much like before, the exponential conic constraints tell us that

$$\forall (a, s, t) \in \mathcal{VA}. \quad u_{a,s,t} \geq \mu_{C_a}(s, t) \log \frac{\mu_{C_a}(s, t)}{\mu_{C_a}(s) \mathbb{P}_a(t|s)}$$

and they hold with equality (at least at those indices where  $\beta_a > 0$ ) because  $\mathbf{u}$  is optimal. So

$$\begin{aligned} \sum_{(a,s,t) \in \mathcal{VA}} \beta_a u_{a,s,t} &= \sum_{(a,s,t) \in \mathcal{VA}} \beta_a \mu_{C_a}(s, t) \log \frac{\mu_{C_a}(s, t)}{\mu_{C_a}(s) \mathbb{P}_a(t|s)} \\ &= \sum_a \beta_a \sum_{(s,t) \in \mathcal{V}a} \mu_{C_a}(s, t) \log \frac{\mu_{C_a}(s, t)}{\mu_{C_a}(s) \mathbb{P}_a(t|s)} \\ &= OInc_m(\Pr_\mu). \end{aligned}$$

Because  $\mu$  is optimal, it is the choice of calibrated tree marginal that minimizes this quantity. By Corollary 8.6.1, the distribution  $[\![\mathbf{m}]\!]^*$  can be represented by such

a tree marginal, and by [Proposition 3.4](#) this distribution minimizes  $OInc_m$ . All this is to say that there exist tree marginals of this form whose corresponding distributions attain the minimum value  $OInc_m(\Pr_\mu) = \langle\!\langle m \rangle\!\rangle_0$ . So  $\mu$  must be one of them, as it minimizes  $OInc(\Pr_\mu)$  among such tree marginals by assumption. Thus  $\Pr_\mu \in [\![m]\!]_0^*$  and the objective value of [\(8.10\)](#) equals  $\langle\!\langle m \rangle\!\rangle_0$ .  $\square$

**Proposition 8.8.** *If  $(\mu, \mathbf{u}, \mathbf{v})$  is a solution to [\(8.13\)](#) and  $\beta \geq \gamma\alpha$ , then  $\Pr_\mu$  is the unique element of  $[\![m]\!]_\gamma^*$ , and the objective of [\(8.13\)](#) at  $(\mu, \mathbf{u}, \mathbf{v})$  equals  $\langle\!\langle m \rangle\!\rangle_\gamma$ .*

*Proof.* Suppose that  $(\mu, \mathbf{u}, \mathbf{v})$  is a solution to [\(8.13\)](#). The first and fourth lines of constraints ensures that  $\mu$  is indeed a calibrated tree marginal. The second line of constraints, plays exactly the same role that it did in the previous problems, most directly in the variant [\(8.10\)](#) for  $\gamma = 0$ . In particular, it tells says

$$\forall (a, s, t) \in \mathcal{VA}. \quad u_{a,s,t} \geq \mu_{C_a}(s, t) \log \frac{\mu_{C_a}(s, t)}{\mu_{C_a}(s) \mathbb{P}_a(t|s)}$$

as before, this holds with equality (at least at those indices where  $\beta_a > \alpha_a \gamma$ ) because  $\mathbf{u}$  is optimal. Because  $\beta \geq \gamma\alpha$  by assumption, either  $\beta_a > \gamma\alpha_a$  or the two are equal, for every  $a \in \mathcal{A}$ . Either way, the argument used at this point in [the proof of Proposition 8.7](#) goes through, giving us:

$$\begin{aligned} \sum_{(a, s, t) \in \mathcal{VA}} (\beta_a - \alpha_a \gamma) u_{a,s,t} &= \sum_{(a, s, t) \in \mathcal{VA}} ((\beta_a - \alpha_a \gamma) \mu_{C_a}(s, t) \log \frac{\mu_{C_a}(s, t)}{\mu_{C_a}(s) \mathbb{P}_a(t|s)}) \\ &= \sum_a (\beta_a - \alpha_a \gamma) \sum_{(s, t) \in \mathcal{V}_a} \mu_{C_a}(s, t) \log \frac{\mu_{C_a}(s, t)}{\mu_{C_a}(s) \mathbb{P}_a(t|s)} \\ &= \sum_a (\beta_a - \alpha_a \gamma) D\left(\mu_{C_a}(S_a, T_a) \parallel \mu_{C_a}(S_a) \mathbb{P}_a(T_a|S_a)\right) \end{aligned}$$

This time, though, that's not the problem objective. In this regard, our problem [\(8.13\)](#) is more closely related to [\(8.13\)](#).

Before we get to that, we have to first bring in the final collection of exponential constraints, which show that

$$\forall C \in \mathcal{C}. \forall c \in \mathcal{V}(C). \quad v_{C,c} \geq \mu_C(c) \log \frac{\mu_C(c)}{VCP_C(c)},$$

and yet again these constraints hold with equality, for otherwise  $\mathbf{v}$  would not be optimal (since we assumed  $\gamma > 0$ ). Therefore,

$$\sum_{(C,c) \in \mathcal{VC}} v_{C,c} = \sum_{(C,c) \in \mathcal{VC}} \mu_C(c) \log \frac{\mu_C(c)}{VCP_C(c)} = -H(\Pr_\mu) \quad \text{by Equation (8.12).}$$

The objective of our problem (8.13) is essentially the same as that of (8.6), so the analysis in the proof of Proposition 8.2 applies with only a handful of superficial modifications. Using that proof to take a shortcut, the objective of (8.13) must equal

$$\begin{aligned} & \sum_{(a,s,t) \in \mathcal{VA}} (\beta_a - \alpha_a \gamma) u_{a,s,t} + \gamma \sum_{(C,c) \in \mathcal{VC}} v_{C,c} - \sum_{(a,s,t) \in \mathcal{VA}^+} \alpha_a \gamma \mu_{C_a}(s, t) \log \mathbb{P}_a(t|s) \\ &= \sum_{(a,s,t) \in \mathcal{VA}} (\beta_a - \alpha_a \gamma) \mu_{C_a}(s, t) \log \frac{\mu_{C_a}(s, t)}{\mu_{C_a}(s) \mathbb{P}_a(t|s)} - \gamma H(\Pr_\mu) - \sum_{(a,s,t) \in \mathcal{VA}^+} \alpha_a \gamma \mu_{C_a}(s, t) \log \mathbb{P}_a(t|s) \\ &= \sum_{a \in \mathcal{A}} \beta_a \mathbb{E}_{\mu_{C_a}} [\log \mathbb{P}_a(T_a | S_a)] + \sum_{a \in \mathcal{A}} (\alpha_a \gamma - \beta_a) H_{\Pr_\mu}(T_a | S_a) - \gamma H(\Pr_\mu) \\ &= \llbracket \mathbf{m} \rrbracket_\gamma(\Pr_\mu), \quad . \end{aligned}$$

Finally, since  $\mu$  is such that this quantity is minimized, and because its unique minimizer can be represented as a cluster tree (by Corollary 8.6.1), we conclude that  $\mu$  must be the cluster tree representation of it. Therefore,  $\Pr_\mu$  is the unique element of  $\llbracket \mathbf{m} \rrbracket_\gamma^*$ , and the objective at  $(\mu, \mathbf{u}, \mathbf{v})$  equals  $\langle\!\langle \mathbf{m} \rangle\!\rangle_\gamma$ , as desired.  $\square$

**Proposition 8.9.** *If  $(\mu, \mathbf{u})$  is a solution to (8.14), then  $\mu$  is a calibrated tree marginal and  $\llbracket \mathbf{m} \rrbracket_{0^+}^* = \{\Pr_\mu\}$ .*

*Proof.* Suppose that  $(\mu, \mathbf{u})$  is a solution to (8.14). The exponential cone constraints state that

$$\begin{aligned} \forall C \in \mathcal{C}. \forall c \in \mathcal{V}(C). \quad u_{C,c} &\geq \mu_C(c) \log \frac{\mu_C(c)}{k_{C,c} VCP_C(c)} \\ &= \mu_C(c) \log \frac{\mu_C(c)}{VCP_C(c)} - \mu_C(c) \log \prod_{a \in \mathcal{A}_C} \nu_C(T_a(c)|S_a(c))^{\alpha_a} \\ &= \mu_C(c) \log \frac{\mu_C(c)}{VCP_C(c)} - \mu_C(c) \sum_{a \in \mathcal{A}_C} \alpha_a \log \nu_C(T_a(c)|S_a(c)), \end{aligned}$$

and once again this holds with equality, as each  $u_{C,c}$  is minimal with this property.

The third line of constraints

$$\forall a \in \mathcal{A}. \mu_{C_a}(S_a, T_a) \nu_{C_a}(S_a) = \mu_{C_a}(S_a) \nu_{C_a}(S_a, T_a)$$

and the assumption that  $\Pr_{\boldsymbol{\nu}} \in \llbracket \mathcal{M} \rrbracket_0^*$ , suffice to ensure that  $\Pr_{\boldsymbol{\mu}} \in \llbracket \mathcal{M} \rrbracket_0^*$  by Proposition 8.3. They also allow us to replace each  $\nu_{C_a}(T_a(c)|S_a(c))$  with  $\nu_{C_a}(T_a(c)|S_a(c))$ , in cases where  $S_a(c) \neq 0$ . Therefore, we calculate the objective to be:

$$\begin{aligned} \mathbf{1}^\top \mathbf{u} &= \sum_{C \in \mathcal{C}} \sum_{c \in \mathcal{V}(C)} \left( \mu_C(c) \log \frac{\mu_C(c)}{VCP_C(c)} - \mu_C(c) \sum_{a \in \mathcal{A}_C} \alpha_a \log \nu_C(T_a(c)|S_a(c)) \right) \\ &= \sum_{C \in \mathcal{C}} \sum_{c \in \mathcal{V}(C)} \mu_C(c) \log \frac{\mu_C(c)}{VCP_C(c)} - \sum_{C \in \mathcal{C}} \sum_{c \in \mathcal{V}(C)} \mu_C(c) \sum_{a \in \mathcal{A}} \mathbb{1}[C = C_a] \alpha_a \log \nu_C(T_a(c)|S_a(c)) \\ &= -H(\Pr_{\boldsymbol{\mu}}) - \sum_{a \in \mathcal{A}} \alpha_a \sum_{C \in \mathcal{C}} \mathbb{1}[C = C_a] \sum_{c \in \mathcal{V}(C)} \mu_C(c) \log \nu_C(T_a(c)|S_a(c)) \quad [\text{by (8.12)}] \\ &= -H(\Pr_{\boldsymbol{\mu}}) - \sum_{a \in \mathcal{A}} \alpha_a \sum_{c \in \mathcal{V}(C)_a} \mu_{C_a}(c) \log \nu_{C_a}(T_a(c)|S_a(c)) \\ &= -H(\Pr_{\boldsymbol{\mu}}) - \sum_{a \in \mathcal{A}} \alpha_a \sum_{c \in \mathcal{V}(C)_a} \mu_{C_a}(c) \log \mu_{C_a}(T_a(c)|S_a(c)) \quad \begin{bmatrix} \text{since } \mu_{C_a}(S_a(c)) > 0 \\ \text{whenever } \mu_{C_a}(c) > 0 \end{bmatrix} \\ &= -H(\Pr_{\boldsymbol{\mu}}) + \sum_{a \in \mathcal{A}} \alpha_a H_{\Pr_{\boldsymbol{\mu}}}(T_a|S_a) \\ &= SDef_m(\Pr_{\boldsymbol{\mu}}). \end{aligned}$$

To summarize:  $\Pr_{\boldsymbol{\mu}}$  minimizes  $SDef_m(\Pr_{\boldsymbol{\mu}})$  among calibrated tree marginals with conditional marginals matching those of  $\boldsymbol{\nu}$ . Since we know that there is a

unique distribution that minimizes  $SDef_m$  among the elements  $\llbracket m \rrbracket_0^*$ , and also that this distribution can be represented by a tree marginal (by Corollary 8.6.1), we conclude that  $\mu$  must represent this distribution. Thus,  $\Pr_\mu = \llbracket m \rrbracket^*$  as desired.  $\square$

The next lemma packages the results of Dahl and Andersen (2022); Nesterov et al. (1999) in a precise form that we will be able to make use of.

**Lemma 8.13.** *Fix integers  $n_o, n_e \in \mathbb{N}$ , and let  $n := 3n_e + n_o$ . Suppose that  $K = \mathbb{R}_{\geq 0}^{n_o} \times K_{\text{exp}}^{n_e} \subset \mathbb{R}^n$  is a product cone, consisting of  $n_o$  copies of the non-negative orthant and  $n_e$  copies of the exponential cone. Let  $\mathbf{c} \in [-1, 1]^n$  and  $\mathbf{b} \in [-1, 1]^m$  be vectors, and  $A \in [-1, 1]^{m \times n}$  be a matrix, defining an exponential conic program*

$$\underset{\mathbf{x} \in K}{\text{minimize}} \quad \mathbf{c}^\top \mathbf{x} \quad \text{subject to} \quad A\mathbf{x} = \mathbf{b}. \quad (8.2)$$

*If this program is strictly feasible (i.e., if there exists  $\mathbf{x} \in \text{int } K$  such that  $A\mathbf{x} = \mathbf{b}$ ), as is its dual problem*

$$\underset{\mathbf{s} \in K^*, \mathbf{y} \in \mathbb{R}^m}{\text{maximize}} \quad \mathbf{b}^\top \mathbf{y} \quad \text{subject to} \quad A^\top \mathbf{y} + \mathbf{s} = \mathbf{c},$$

*(i.e., if there exists  $\mathbf{s} \in \text{int } K_*$  such that  $A^\top \mathbf{y} + \mathbf{s} = \mathbf{c}$ ), then both can be simultaneously solved to precision  $\epsilon$  in  $O(n(m+n)^\omega \log \frac{n+m}{\epsilon})$  time, where  $\omega$  is the smallest exponent such that a linear system of  $k$  variables and equations can be solved in  $O(k^\omega)$  time. Furthermore, MOSEK solves this problem in  $O(n(m+n)^3 \log \frac{n+m}{\epsilon})$  time.*

*Proof.* For this, we begin by appealing to the algorithm and analysis of Baden-broek and Dahl (2021), threading details through for this specific choice of cone  $K$ . To finish the proof, however, we will also need to supplement that analysis with some other well-established results of Nesterov et al. (1999) that the authors were no doubt familiar with, but did not bother referencing.

First, we'll need some background material from convex optimization. A *logarithmically homogeneous self-concordant barrier* with parameter  $\nu$  ( $\nu$ -LHSCB) for a cone  $K$  is a thrice differentiable strictly convex function  $F : \text{int } K \rightarrow \mathbb{R}$  satisfying  $F(tx) = F(x) - \nu \log t$  for all  $t > 0$  and  $x \in \text{int } K$ . In some sense, the point of such a barrier function is to augment the optimization objective so that we remain within the cone during the optimization process.

For the positive orthant cone  $\mathbb{R}_{\geq 0}$ , the function  $x \mapsto -\log x$  is a 1-LHSCB. We now fill in some background facts about exponential cones. The *dual* of the exponential cone is

$$\begin{aligned} K_{\exp}^* &:= \{(s_1, s_2, s_3) \in \mathbb{R}^3 : \forall (x_1, x_2, x_3) \in K_{\exp}. x_1 s_1 + x_2 s_2 + x_3 s_3 \geq 0\} \\ &= \{(s_1, s_2, s_3) : -s_1 \log(-s_1/s_3) + s_1 - s_2 \leq 0, s_1 \leq 0, s_3 \geq 0\}. \end{aligned}$$

Consider points  $x = (x_1, x_2, x_3) \in K_{\exp}$ . The function

$$F_{\exp}(x) := -\log \left( x_2 \log \frac{x_1}{x_2} - x_3 \right) - \log x_1 x_2 \quad (8.16)$$

is a 3-LHSCB for  $K_{\exp}$ , since

$$\begin{aligned} F_{\exp}(tx) &= -\log \left( tx_2 \log \frac{tx_1}{tx_2} - tx_3 \right) - \log(t^2 x_1 x_2) \\ &= -\log \left( t \left( \log \frac{x_1}{x_2} - x_3 \right) \right) - \log(x_1 x_2) - 2 \log t \\ &= F_{\exp}(x) - 3 \log t \end{aligned}$$

Such barrier functions can be combined to act on product cones by summation. Concretely, suppose that for each  $i \in \{1, \dots, k\}$ , we have a  $\nu_i$ -LHSCB  $F_i : \text{int } K_i \rightarrow \bar{\mathbb{R}}$ . For  $x = (x_i)_{i=1}^k \in \prod_i K_i$ , the function  $F(x) := \sum_{i=1}^k F_i(x_i)$  is a  $(\sum_i \nu_i)$ -LHSCB for  $\prod_i K_i$ , since

$$F(tx) = \sum_{i=1}^k F_i(tx_i) = \sum_{i=1}^k (F_i(x_i) - \nu_i \log t) = F(x) - \sum_{i=1}^k \nu_i.$$

In this way, our product cone  $K = \mathbb{R}_{\geq 0}^{n_o} \times K_{\text{exp}}^{n_e}$  admits a LHSCB  $F$  with parameter  $\nu = n_o + 3n_e = n$ . Furthermore it can be evaluated in  $O(n)$  time, as can each component of its gradient  $F'(x)$  and Hessian  $F''(x) \in \mathbb{R}^{n \times n}$  at  $x$ , all of which can be expressed analytically. In addition, the convex conjugate of  $F$  also has a known analytic form.

Generally speaking, the idea behind primal-dual interior point methods ([Nesterov and Nemirovskii 1994](#)) such as the one behind MOSEK, is to maintain both a point  $x \in K$  and a dual point  $s \in K_*$  (as well as  $y \in \mathbb{R}^m$ ) and iteratively update them, as we slowly relax the barrier and approach a point on the boundary of the cone. The quantity  $\mu(z) := \langle s, x \rangle / \nu \geq 0$ , called the complementarity gap, is a measure of how close the process is to converging.

Because the initial points may not satisfy the constraints, instead the standard algorithms work with “extended points”  $\bar{x} = (x, \tau)$  and  $\bar{s} = (s, \kappa)$ , for which the analogous complementarity gap is  $\mu^e(\bar{x}, \bar{s}) := (\langle x, s \rangle + \kappa\tau) / (\nu + 1)$ . Altogether, the data at each iteration may be summarized as a point  $z = (y, x, \tau, s, \kappa) \in \mathbb{R} \times (K \times \mathbb{R}_{\geq 0}) \times (K_* \times \mathbb{R}_{\geq 0})$ . The primary object of interest is then something called the *homogenous self-dual* model. Originally due to [Nesterov et al. \(1999\)](#) and also used by others ([Skajaa and Ye 2015](#)), it can be defined as a linear operator:

$$G : \bar{\mathbb{R}}^{m+2n+2} \rightarrow \bar{\mathbb{R}}^{n+m+1}$$

$$G(y, x, \tau, s, \kappa) := \begin{bmatrix} 0 & A & -b \\ -A^\top & 0 & c \\ b^\top & -c^\top & 0 \end{bmatrix} \begin{bmatrix} y \\ x \\ \tau \end{bmatrix} - \begin{bmatrix} 0 \\ s \\ k \end{bmatrix}.$$

The reason for our interest is that if  $z$  is such that  $G(z) = 0$  and  $\tau > 0$ , then  $(x/\tau)$  is a solution to the primal problem, and  $(y, s)/\tau$  is a solution to the dual problem ([Skajaa and Ye 2015](#), Lemma 1), while if  $G(z) = 0$  and  $\kappa > 0$ , then at least one of the two problems is infeasible.

We now are in a better position to describe the algorithm. According to the MOSEK documentation ([Dahl and Andersen 2022](#)), for the exponential cone, begins with an initial point

$$\mathbf{v} := (1.291, 0.805, -0.828) \in (K_{\exp} \cap K_{\exp}^*)$$

for this particular cone  $K$ , the algorithm begins at the initial point

$$z_0 := (y_0, x_0, \tau_0, s_0, \kappa_0)$$

where  $x_0 = s_0 = (\overbrace{1, \dots, 1}^{n_o \text{ copies}}, \overbrace{\mathbf{v}, \dots, \mathbf{v}}^{n_e \text{ copies}}) \in (\mathbb{R}_{\geq 0})^{n_o} \times (K_{\exp} \cap K_{\exp}^*)^{n_e}$ ,

$$y_0 = \mathbf{0} \in \mathbb{R}^m, \quad \tau_0 = \kappa_0 = 1.$$

At each iteration, the first step is to predict a direction for which [Badenbroek and Dahl \(2021\)](#) compute a scaling matrix  $W$ . To describe it, we first need to define *shadow iterates*

$$\tilde{x} := -F'_*(s) \quad \text{and} \quad \tilde{s} := -F'(x).$$

which are in a sense reflections of  $s$  and  $x$  across their barrier functions, and can be computed in in  $O(n)$  time. The analogous notion of complementarity can then be defined as  $\tilde{\mu}(z) := \langle \tilde{x}, \tilde{s} \rangle / \nu$ . The scaling matrix, which we do not interpret here, can then be calculated as:

$$W := \mu F''(x) + \frac{ss^\top}{\nu\mu} - \frac{\mu\tilde{s}\tilde{s}^\top}{\nu} + \frac{(s - \mu\tilde{s})(s - \mu\tilde{s})^\top}{(s - \mu\tilde{s})^\top(x - \mu\tilde{x})} - \frac{\mu[F''(x)\tilde{x} - \tilde{s}][F''(x)\tilde{x} - \tilde{s}]^\top}{\tilde{x}^\top F''(x)\tilde{x} - \nu\tilde{\mu}^2} \tag{8.17}$$

Doing so requires  $O(n^2)$  steps (although it may be parallelized). The first four terms clearly require  $O(n^2)$  steps, since each one is an outer product resulting in a  $n \times n$  matrix. The last term computes a matrix-vector product (which requires  $O(n^2)$  steps), and computes an outer product with the resulting vector, which takes  $O(n^2)$  steps as well.

The next step involves finding a solution  $\Delta z^{\text{aff}} = (\dots)$  to the system of equations

$$G(\Delta z^{\text{aff}}) = -G(z) \quad (8.18\text{a})$$

$$\tau \Delta \kappa^{\text{aff}} + \Delta \tau^{\text{aff}} = -\tau \kappa \quad (8.18\text{b})$$

$$W \Delta x^{\text{aff}} + \Delta s^{\text{aff}} = -s. \quad (8.18\text{c})$$

(8.18a-c) describe a system of  $(n+m+1) + 1 + (n) = 2n+m+2$  equations and equally many unknowns, and solved in  $O((n+m)^\omega)$  steps. It may be possible to exploit the sparsity of  $G$  to do better.

The next step is to center that search direction so that it lies on the central path. This is done by finding a solution  $\Delta z^{\text{cen}}$  to

$$G(\Delta z^{\text{cen}}) = G(z) \quad (8.19\text{a})$$

$$\tau \Delta \kappa^{\text{cen}} + \kappa \Delta \tau^{\text{cen}} = \mu^e \quad (8.19\text{b})$$

$$W \Delta x^{\text{cen}} + \Delta s^{\text{cen}} = \mu^e \tilde{s}, \quad (8.19\text{c})$$

which again can be done in  $O((n+m)^3)$  steps with Gaussian elimination, or with a fancier solver in  $O((n+m)^2 \cdot 332)$  steps. The two updates are then applied to the current point  $z$  to obtain

$$z_+ = (y_+, x_+, \tau_+, s_+, \kappa_+) := z + \alpha(\Delta z^{\text{aff}} + \gamma \Delta z^{\text{cen}}).$$

Finally, a “correction step”, which is the primary innovation of [Badenbroek and Dahl \(2021\)](#) and used in MOSEK’s algorithm, is a third direction  $\Delta z_+^{\text{cor}}$ , which is found by solving the system of equations

$$G(\Delta z^{\text{cor}}) = 0 \quad (8.20\text{a})$$

$$\tau_+ \Delta \kappa^{\text{cor}} + \kappa_+ \Delta \tau^{\text{cor}} = 0 \quad (8.20\text{b})$$

$$W_+ \Delta x_+^{\text{cor}} + \Delta s^{\text{cen}} = \mu^e \tilde{s}, \quad (8.20\text{c})$$

where  $W_+$  is defined the same way that  $W$  is, except that it uses the components of  $z_+$  instead of  $z$ . After adding the correction step  $\Delta z_+^{\text{cor}}$  to  $z$ , we repeat the entire process. The full algorithm, then, is given by [Algorithm 2](#) below.

---

**Algorithm 2 [Badenbroek and Dahl]**


---

```

 $z \leftarrow (y_0, x_0, \tau_0, s_0, \kappa_0);$ 
while do
    Compute scaling matrix  $W$  as in (8.17);
    Find the solution  $\Delta z^{\text{aff}}$  to (8.18a-c), and the solution  $\Delta z^{\text{cen}}$  to (8.19a-c);
     $z_+ \leftarrow z + \alpha(\Delta z^{\text{aff}} + \gamma\Delta z^{\text{cen}});$ 
    Compute the scaling matrix  $W_+$ ;
    Find the solution  $\Delta z_+^{\text{cor}}$  to (8.20a-c);
     $z \leftarrow z_+ + \Delta z_+^{\text{cor}};$ 

```

---

We have verified that each iteration of this process can be done in  $O((n+m)^\omega)$  time. Their main result ([Badenbroek and Dahl 2021](#), Theorem 3), states that for every  $\epsilon \in (0, 1)$ , the algorithm results in a solution  $z$  satisfying

$$\mu^e(z) \leq \epsilon \quad \text{and} \quad \|G(z)\| \leq \epsilon \|G(z_0)\|$$

in  $O(n \log(1/\epsilon))$  iterations, for a total cost of  $O(n(m+n)^3 \log(1/\epsilon))$  time with Gaussian elimination, or  $O(n(m+n)^{2.332} \log(1/\epsilon))$  time using the linear solver with best known asymptotic complexity as of 2022 [Duan et al. \(2022\)](#).

**Verifying that the solution is approximately optimal.** What we have at this point is not quite enough: simply because the residual quantity  $G(z)$  is approximately zero (so that we have approximately solved the homogenous model), does not mean that we've approximately solved the original problem. Specifically, it's entirely possible a priori that the parameter  $\tau$  goes to zero at the same rate as everything else, and the quantity  $(x/\tau)$  does not converge to a solution to the primal problem. To address this issue, we must also trace the analysis of

the seminal work of [Nesterov et al. \(1999\)](#), who use slightly different quantities, conflicting with the notation we have been using thus far.

Following [Nesterov et al. \(1999\)](#), pg. 231), fix an initial point  $z_0$ , and let *shifted feasible set*  $\mathcal{F} := \{z \in \mathbb{R} \times K \times \mathbb{R}_{\geq 0} \times K^* \times \mathbb{R}_{\geq 0} : G(z) = G(z_0)\}$  be the collection of all points that have the same residual as  $z_0$ . [Nesterov, Todd, and Ye](#) also refer to a complementary gap by  $\mu(z)$  and define it identically, but the meaning of this parameter is different, because the set  $\mathcal{F}$  on which it's defined is quite distinct from (if closely related to) the iterates of [Badenbroek and Dahl](#)'s algorithm. In the service of clarity, will call this quantity  $\mu^N(z^N)$ , for  $z^N = (y^N, x^N, \tau^N, s^N, \kappa^N) \in \mathcal{F}$ .

Although we made a point of emphasizing that the two are distinct, the actual relationship between them is straightforward. Let  $z = (y, x, \tau, s, \kappa)$  be the final output of [Badenbroek and Dahl \(2021\)](#). In proving their main theorem, they also prove that  $G(z) = \epsilon G(z_0)$ , and  $\mu^e = \epsilon$ ; because  $G$  is linear, we know that  $G(z/\epsilon) = G(z_0)$ . This means that  $z^N := z/\epsilon \in \mathcal{F}$ . Therefore,

$$\mu^N(z^N) = \frac{1}{\nu + 1} \left( \left\langle \frac{s}{\epsilon}, \frac{x}{\epsilon} \right\rangle + \frac{\tau \kappa}{\epsilon \epsilon} \right) = \frac{1}{\epsilon^2} \mu^e(z) = \frac{1}{\epsilon}.$$

So, roughly speaking,  $\mu^N$  and  $\mu^e$  are reciprocals. [Badenbroek and Dahl](#) also prove that, every iterate  $z$  satisfies their assumption (A2): for a fixed constant  $\beta$  (equal to 0.9 in their analysis),  $\beta \mu^e(z) \leq \tau \kappa$ . Consequently, it happens that the same inequality holds with Nesterov's notation:

$$\tau^N \kappa^N = \frac{\tau \kappa}{\epsilon \epsilon} = \frac{\tau \kappa}{\epsilon^2} \geq \frac{\beta \epsilon}{\epsilon^2} = \frac{\beta}{\epsilon} = \beta \mu^N(z^N).$$

This witnesses that  $z^N = \frac{z}{\epsilon}$  satisfies equation (81) of [Nesterov et al.](#), which allows us to apply one of their main theorems, which addresses these issues. Supposing that the original problem is solvable, let  $(x^*, s^*)$  be any solution to the primal and dual problems, and define the value  $\psi := 1 + \langle s_0, x^* \rangle + \langle s^*, x_0 \rangle \geq 1$ , which

depends only on the problem and the choice of initialization. Then Theorem 1, part 1 of [Nesterov, Todd, and Ye](#), allows us to conclude that

$$\frac{\kappa}{\epsilon} \leq \psi \quad \text{and} \quad \frac{\tau}{\epsilon} \geq \frac{\beta}{\epsilon\psi} \quad \iff \quad \kappa \leq \epsilon\psi \quad \text{and} \quad \tau \geq \frac{\beta}{\psi}.$$

Finally, the original theorem guarantees that  $\|G(x)\| \leq \epsilon\|G(z_0)\|$ , meaning that

$$\left\| A\left(\frac{x}{\tau}\right) - b \right\|_\tau + \left\| A^\top\left(\frac{y}{\tau}\right) - \frac{s}{\tau} - c \right\|_\tau + \left\| b^\top\left(\frac{y}{\tau}\right) - c^\top\left(\frac{x}{\tau}\right) - \frac{\kappa}{\tau} \right\|_\tau \leq \epsilon\|G(z_0)\|.$$

Since the euclidean norm is an upper bound on the deviation in any component ( $\|v\| := \sqrt{\sum_i v_i^2} \geq \sqrt{\max_i v_i^2} = \max_i v_i =: \|v\|_\infty$ ), this means that in light of our bound on  $\tau$  above, we have

$$\left\| A\left(\frac{x}{\tau}\right) - b \right\|_\infty + \left\| A^\top\left(\frac{y}{\tau}\right) + \frac{s}{\tau} - c \right\|_\infty + \left\| b^\top\left(\frac{y}{\tau}\right) - c^\top\left(\frac{x}{\tau}\right) - \frac{\kappa}{\tau} \right\|_\infty \leq \epsilon \frac{\beta\|G(z_0)\|}{\psi}.$$

The first two components show that the total constraint violation (in the primal and dual problems, respectively) is at most  $\epsilon\beta/\psi\|G(z_0)\|$ . Meanwhile, the final component shows that the duality gap  $gap = b^\top\left(\frac{y}{\tau}\right) - c^\top\left(\frac{x}{\tau}\right)$ , which is positive and an upper bound on the difference between the objective at  $x/\tau$  and the optimal objective value, satisfies

$$gap \leq gap + \frac{\kappa}{\tau} \leq \frac{\epsilon\beta\|G(z_0)\|}{\psi}.$$

Thus  $x/\tau$  is an  $(\epsilon\|G(z_0)\|)$ -approximate solution to the original exponential conic problem. Since also  $\psi \geq 1$ , we may freely drop it to get a looser bound. All that remains is to investigate  $\|G(z_0)\|$ , the residual norm of the initial point chosen by the MOSEK solver, which equals:

$$\|G(z_0)\| = \|Ax_0 - b\| + \|A^\top y_0 + s_0 - c\| + |c^\top x - b^\top y + 1|.$$

Making use of our assumption that every component of  $A$ ,  $b$ , and  $c$  is at most

one, we find that

$$\begin{aligned} \|Ax_0 - b\|^2 &= \sum_j (\sum_i A_{j,i}(1.3) - b_j)^2 \leq m(1.3n + 1)^2 \in O(mn^2) \subset O((m+n)^3) \\ \|A^\top y_0 + s_0 - c\|^2 &= \sum_i (\sum_j (A_{j,i}) \leq n(m+2)^2 \in O(nm^2) \subset O((m+n)^3) \\ |c^\top x - b^\top y + 1|^2 &\leq (1.3n + m + 1)^2 \in O((n+m)^2) \subset O((n+m)^3). \end{aligned}$$

Therefore, the residual of the initial point is  $G(z_0) \in O((n+m)^{3/2})$ .

To obtain a solution at most  $\epsilon_0$  away from the true solution in any coordinate, we need to select  $\epsilon$  small enough that the final output of the algorithm  $z$  satisfies

$$\epsilon \|G(z_0)\| \leq \epsilon_0 \iff \frac{1}{\epsilon} \geq \frac{1}{\epsilon_0} \|G(z_0)\|$$

It therefore suffices to choose  $\frac{1}{\epsilon} \in O(\frac{1}{\epsilon_0}(n+m)^{3/2})$ , leading to  $\log \frac{1}{\epsilon} = O(\log \frac{n+m}{\epsilon_0})$  iterations. Thus, we arrive at our total advertised asymptotic complexity of time

$$O\left(n(n+m)^\omega \log \frac{n+m}{\epsilon_0}\right).$$

In particular, to attain machine precision, we can fix  $\epsilon_0$  to be the smallest gap between numbers representable (say with 64-bit floats, leading to  $\epsilon_0 = 10^{-78}$  in the worst case), and omit the dependance on  $\epsilon_0$  for the price of relatively small constant (78, for 64-bit floats).  $\square$

Having combed through all of the details of the analysis of [Badenbroek and Dahl \(2021\)](#) and [Nesterov et al. \(1999\)](#) for exponential conic programs as we have defined them, we are ready to show that this algorithm solves the problems presented in [Section 8.4](#) within polynomial time.

In the results that follow, we use the symbol  $O_{\text{BP}}(\cdot)$  to describe the complexity under the *bounded precision* assumption: the numerical values of  $(\alpha, \beta, \mathbb{P})$  that

describe the PDG, as well as  $\gamma$ , lie within a fixed range, e.g., are 64-bit floating point numbers. Correspondingly, we use  $\tilde{O}_{\text{BP}}(\cdot)$  to describe the complexity under the same assumption, but hiding logarithmic factors for parameters on which the complexity also depends polynomially.

**Lemma 8.14.** *Problem (8.10) can be solved to  $\epsilon$  precision in time*

$$O\left((\mathcal{V}\mathcal{A} + \mathcal{V}\mathcal{C})^{1+\omega} \left( \log \frac{|\mathcal{V}\mathcal{A}| + |\mathcal{V}\mathcal{C}|}{\epsilon} + \log \frac{\beta^{\max}}{\beta^{\min}} \right) \right) \subset \tilde{O}_{\text{BP}}\left(|\mathcal{V}\mathcal{A} + \mathcal{V}\mathcal{C}|^4 \log \frac{1}{\epsilon}\right),$$

where  $\beta^{\max} := \max_{a \in \mathcal{A}} \beta_a$  is the largest value of  $\beta$ , and  $\beta^{\min} := \min_{a \in \mathcal{A}} \{\beta_a : \beta_a > 0\}$  is the smallest positive one.

*Proof.* Problem (8.10) can be translated via the DCP framework to the following exponential conic program, which has:

- variables  $x = (\mathbf{u}, \mathbf{v}, \mathbf{w}, \boldsymbol{\mu}) \in K_{\text{exp}}^{\mathcal{V}\mathcal{A}} \times \mathbb{R}_{\geq 0}^{\mathcal{V}\mathcal{C}}$ , where
  - $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \bar{\mathbb{R}}^{\mathcal{V}\mathcal{A}}$  are all vectors over  $\mathcal{V}\mathcal{A}$ , that at index  $\iota = (a, s, t) \in \mathcal{V}\mathcal{A}$ , have components  $u_\iota, v_\iota$ , and  $w_\iota$ , respectively;
  - $\boldsymbol{\mu} = [\mu_C(C=c)]_{C \in \mathcal{C}, c \in \mathcal{V}(C)} \in \bar{\mathbb{R}}^{\mathcal{V}\mathcal{C}}$  is a vector representation of a tree marginal over clusters  $\mathcal{C}$ ;

- constraints as follows:

- two linear constraints for every  $(a, s, t) \in \mathcal{V}\mathcal{A}$  to ensure that

$$v_{a,s,t} = \mu_{C_a}(s, t) \quad \left( = \sum_{\bar{c} \in \mathcal{V}(C_a \setminus \{S_a, T_a\})} \mu_{C_a}(\bar{c}, s, t) \right)$$

$$\text{and} \quad w_{a,s,t} = \mu_{C_a}(S_a=s) \mathbb{P}_a(T_a=t \mid S_a=s)$$

$$\left( \text{or, more precisely, } w_{a,s,t} = \mathbb{P}_a(T_a=t \mid S_a=s) \sum_{\bar{c} \in \mathcal{V}(C_a \setminus \{S_a\})} \mu_{C_a}(\bar{c}, s) \right);$$

- for every edge  $(C-D) \in \mathcal{T}$ , and every value  $\omega \in \mathcal{V}(C \cap D)$  of the variables that clusters  $C$  and  $D$  have in common, a linear constraint

$$\sum_{\bar{c} \in \mathcal{V}(C \setminus D)} \mu_C(\bar{c}, \omega) = \sum_{\bar{d} \in \mathcal{V}(D \setminus C)} \mu_D(\bar{d}, \omega);$$

- and one constraint for each cluster  $C \in \mathcal{C}$  to ensure that  $\mu_C$  lies on the probability simplex, i.e.,

$$\sum_{c \in \mathcal{V}(C)} \mu_C(c) = 1.$$

Altogether this means that we have an exponential conic program in the form of [Lemma 8.13](#), with  $n = 3|\mathcal{VA}| + |\mathcal{VC}|$  variables, and  $m = 2|\mathcal{VA}| + |\mathcal{VT}| + |\mathcal{C}|$  constraints, where  $\mathcal{VT} = \{(C-D, \omega) : C-D \in \mathcal{T}, \omega \in \mathcal{V}(C \cap D)\}$ . Since we can simply disregard variables whose value sets are singletons, we can assume  $\mathcal{V}(C) > 1$ ; summing over all clusters yields  $|\mathcal{VC}| > |\mathcal{C}|$ . At the same time, since  $|\mathcal{VT}| \leq |\mathcal{VC}|$ , we have

$$m, n, (m+n) \in O(|\mathcal{VA}|, |\mathcal{VC}|).$$

We now give the explicit construction of the data  $(A, b, c)$  of the exponential conic program that [\(8.10\)](#) compiles to. The variables are indexed by tuples of the form  $i = (\ell, a, s, t)$  for  $(a, s, t) \in \mathcal{VA}$  and  $\ell \in \{u, v, w\}$ , or by tuples of the form  $(C, c)$ , for  $c \in \mathcal{V}(C)$  and  $C \in \mathcal{C}$ , while the constraints are indexed by tuples of the form  $j = (\ell, a, s, t)$  for  $(a, s, t) \in \mathcal{VA}$  and  $\ell \in \{v, w\}$ , of the form  $(C-D, \omega)$ , for an edge  $(C-D) \in \mathcal{T}$  and  $\omega \in \mathcal{V}(C \cap D)$ , or simply by  $(C)$ , the name of a cluster  $C \in \mathcal{C}$ . The problem data  $A = [A_{j,i}]$ ,  $b = [b_j]$ ,  $c = [c_i]$  of this program are zero,

except (possibly) for the components:

$$c_{(u,a,s,t)} = \beta_a$$

$$A_{(v,a,s,t),(C,c)} = \mathbb{1}[C=C_a \wedge S_a(c)=s \wedge T_a(c)=t]$$

$$A_{(w,a,s,t),(C,c)} = \mathbb{P}_a(T_a=t \mid S_a=s) \mathbb{1}[C=C_a \wedge S_a(c)=s]$$

$$A_{(w,a,s,t),(w,a,s,t)} = -1$$

$$A_{(v,a,s,t),(v,a,s,t)} = -1$$

$$A_{(C-D,\omega),(C',c)} = \mathbb{1}[C=C'] - \mathbb{1}[C'=D]$$

$$A_{(C),(C,c)} = 1$$

$$b_{(C)} = 1,$$

where  $\mathbb{1}[\varphi]$  is equal to 1 if  $\varphi$  is true, and zero if  $\varphi$  is false. We note that we can equivalently divide each  $\beta_a$  by  $\max_a \beta_a$  without affecting the problem, although this could affect the approximation accuracy by the same factor. Thus, we get another factor of

$$\log(\max\{1\} \cup \{\beta_a : a \in \mathcal{A}\}) \subseteq O(\log(1 + \max_a \beta_a)).$$

Finally, to find a point that is  $\epsilon$ -close (say, in 2-norm) to the limiting point  $\mu^*$  on the central path, as opposed to simply one that for which the suboptimality gap is at most  $\epsilon$ , we can appeal to strong concavity of the objective function. (Conditional) relative entropy is 1-strongly convex, and each relative entropy term is scaled by  $\beta_a$ . Furthermore, we're only considering marginal conditional entropy, so this convexity may not hold in all directions. Still, if the next step direction  $\delta$  is not far from the gradient (as is the case if the interior point method has nearly converged), then, in that direction, the objective will be at least  $(\min_a \{\beta_a : \beta_a > 0\})$ -strongly convex. Therefore, by multiplying the requested precision by an additional factor of  $\min_a \{\beta_a : \beta_a > 0\}$ , we can guarantee that our

point is  $\epsilon$ -close to  $\mu^*$ , and not just in complementarity gap.

To summarize, applying Lemma 8.13, we find that we can solve problem (8.10) in time

$$O\left(\left(|\mathcal{VA}| + |\mathcal{VC}|\right)^{1+\omega}\left(\log \frac{|\mathcal{VA}| + |\mathcal{VC}|}{\epsilon} + \log \frac{\beta_{\max}}{\beta_{\min}}\right)\right) \subset \tilde{O}_{BP}\left(\left(|\mathcal{VA}| + |\mathcal{VC}|\right)^4 \log \frac{1}{\epsilon}\right).$$

The factor of  $\log \frac{\beta_{\max}}{\beta_{\min}}$  can be treated as a constant under the bounded precision assumption.  $\square$

We now quickly step through the analogous construction for problems (8.13) and (8.14), which solve the  $\hat{\gamma}$ -inference problem, and  $0^+$ -inference, respectively.

**Lemma 8.15.** *Problem (8.13) is solved to precision  $\epsilon$  in time*

$$\begin{aligned} O\left(|\mathcal{VA} + \mathcal{VC}|^{1+\omega}\left(\log \frac{|\mathcal{VA}| + |\mathcal{VC}|}{\epsilon} + \log(1 + \|\boldsymbol{\beta}\|_\infty) + \log \log \frac{1}{p^{\min}}\right)\right) \\ \subset \tilde{O}_{BP}\left(|\mathcal{VA} + \mathcal{VC}|^4 \log \frac{1}{\epsilon}\right) \end{aligned}$$

where  $p^{\min}$  is the smallest nonzero probability in the PDG.

*Proof.* Problem (8.13) has

► variables  $x = (\mathbf{u}, \mathbf{y}, \mathbf{w}, \mathbf{v}, \boldsymbol{\mu}, \mathbf{z}) \in K_{\exp}^{\mathcal{VA}} \times K_{\exp}^{\mathcal{VC}}$  where

- $\mathbf{u}, \mathbf{y}, \mathbf{w} \in \bar{\mathbb{R}}^{\mathcal{VA}}$  are all vectors over  $\mathcal{VA}$  that at index  $\iota = (a, s, t) \in \mathcal{VA}$ , have components  $u_\iota, v_\iota$ , and  $w_\iota$ , respectively;
- Meanwhile,  $\mathbf{v}, \boldsymbol{\mu}, \mathbf{z} \in \bar{\mathbb{R}}^{\mathcal{VC}}$  are all vectors over  $\mathcal{VC}$  which at index  $(C, c) \in \mathcal{VC}$ , have components  $v_{C,c}, \mu_C(c)$ , and  $z_{C,c}$ , respectively. Once again,  $\boldsymbol{\mu} = [\mu_C(C=c)]_{C \in \mathcal{C}, c \in \mathcal{V}(C)} \in \bar{\mathbb{R}}^{\mathcal{VC}}$  is intended to be a vector representation of a tree marginal.

► constraints as follows:

- two linear constraints for each  $(a, s, t) \in \mathcal{VA}$ , to ensure that

$$y_{a,s,t} = \mu_{C_a}(s, t) \quad \text{and} \quad w_{a,s,t} = \mu_{C_a}(S_a=s) \mathbb{P}_a(T_a=t \mid S_a=s),$$

- for every edge  $(C-D) \in \mathcal{T}$ , and every value  $\omega \in \mathcal{V}(C \cap D)$  of the variables that clusters  $C$  and  $D$  have in common, a linear constraint

$$\sum_{\bar{c} \in \mathcal{V}(C \setminus D)} \mu_C(\bar{c}, \omega) = \sum_{\bar{d} \in \mathcal{V}(D \setminus C)} \mu_D(\bar{d}, \omega)$$

- for every  $(a, s, t) \in \mathcal{VA}^0$ , a linear constraint that ensures

$$0 = \mu_{C_a}(S_a=s, T_a=t) \quad \left( = \sum_{\bar{c} \in \mathcal{V}(C \setminus \{S_a, T_a\})} \mu_{C_a}(\bar{c}, s, t) \right)$$

- a linear constraint for every value  $c \in \mathcal{V}(C)$  of every cluster  $C \in \mathcal{C}$ , to ensure that

$$z_{C,c} = \mu_C(VCP_C(c)) \quad \left( = \sum_{\bar{c} \in \mathcal{V}(C \setminus VCP_C)} \mu_C(\bar{c}, VCP_C(c)) \right)$$

- and one constraint for each cluster  $C \in \mathcal{C}$  to ensure that  $\mu_C$  lies on the probability simplex, i.e.,  $\sum_{c \in \mathcal{V}(C)} \mu_C(c) = 1$ .

So in total, there are  $n = |\mathcal{VA}| + |\mathcal{VC}|$  variables, and  $m = 2|\mathcal{VA}| + |\mathcal{VT}| + |\mathcal{VA}^0| + |\mathcal{VC}| + |\mathcal{C}|$  constraints. The same arguments made in [Lemma 8.14](#) show that both  $n, m \in O(|\mathcal{VA}| + |\mathcal{VC}|)$ .

Also like before, it is easy to see that the components of  $A$  and  $b$  are all at most 1. However, we will need to rescale the objective  $c$  in order for each of its components to be most 1. We can do this by dividing it by  $\max\{-\beta_a \log p_a(t|s)\}_{(a,s,t) \in \mathcal{VA}} \cup \{1\}$ .

Finally, to ensure that we have a solution that is  $\epsilon$ -close to the end of the central path, as opposed to one that is merely  $\epsilon$ -close in complementarity gap, we must appeal to convexity. As in the proof of [Lemma 8.14](#), this amounts to reducing the target accuracy by a factor of the smallest possible coefficient of strong convexity, along the next step direction. In this case, the bound is simpler: because negative entropy is (unconditionally) 1-strongly convex, and since  $\beta \geq \alpha\gamma$ , the remaining terms are convex, this could be, at worst,  $\frac{1}{\gamma}$ .

This gives rise to our result: problem [\(8.13\)](#) can be solved in

$$\begin{aligned} O\left(|\mathcal{VA} + \mathcal{VC}|^{1+\omega} \left( \log \frac{|\mathcal{VA} + \mathcal{VC}|}{\epsilon} + \log \frac{1}{\gamma} \left( 1 + \max_{(a,s,t) \in \mathcal{VA}} \beta_a \log \frac{1}{\mathbb{P}_a(t|s)} \right) \right) \right) \\ \subset O\left(|\mathcal{VA} + \mathcal{VC}|^{1+\omega} \left\{ \log \frac{|\mathcal{VA} + \mathcal{VC}|}{\epsilon} + \log \frac{\beta^{\max}}{\gamma} + \log \log \frac{1}{p^{\min}} \right\} \right) \\ \subset \tilde{O}_{\text{BP}}\left(|\mathcal{VA} + \mathcal{VC}|^{1+\omega} \left( \log \frac{1}{\epsilon} \right)\right) \end{aligned}$$

operations, where  $p$  is the smallest nonzero probability in the PDG, and  $\beta^{\max}$  is the largest confidence in the PDG larger than 1.  $\square$

**Lemma 8.16.** *Problem [\(8.14\)](#) is solved to precision  $\epsilon$  in*

$$O\left(|\mathcal{VC}||\mathcal{VA} + \mathcal{VC}|^\omega \log \frac{|\mathcal{VA} + \mathcal{VC}|}{\epsilon}\right) \subset \tilde{O}_{\text{BP}}\left(|\mathcal{VC} + \mathcal{VA}|^4 \log \frac{1}{\epsilon}\right) \text{ time.}$$

*Proof.* Problem [\(8.14\)](#) is slightly more straightforward; having done [Lemmas 8.14](#) and [8.15](#) in depth, we do this one more quickly. In the standard form, problem [\(8.14\)](#), has variables  $x = (\mathbf{u}, \boldsymbol{\mu}, \mathbf{w}) \in K_{\text{exp}}^{\mathcal{VC}}$ . The constraints are:

- one linear constraint for each  $(C, c) \in \mathcal{VC}$ , to ensure that

$$w_{C,c} = k_{(C,c)} \mu_C(VCP_C(c)) \quad \left( = \sum_{\bar{c} \in \mathcal{V}(C \setminus VCP_C)} \mu_C(\bar{c}, VCP_C(c)) \right)$$

- for every edge  $(C-D) \in \mathcal{T}$ , and every value  $\omega \in \mathcal{V}(C \cap D)$  of the variables that clusters  $C$  and  $D$  have in common, a linear constraint

$$\sum_{\bar{c} \in \mathcal{V}(C \setminus D)} \mu_C(\bar{c}, \omega) = \sum_{\bar{d} \in \mathcal{V}(D \setminus C)} \mu_D(\bar{d}, \omega)$$

- for every  $(a, s, t) \in \mathcal{VA}$ , a linear constraint that ensures

$$\mu_{C_a}(S_a=s, T_a=t) \nu_{C_a}(S_a=s) = \nu_{C_a}(S_a=s, T_a=t) \mu_{C_a}(S_a=s).$$

This is linear, because recall that  $\nu$  is a constant in this optimization problem, found by having previously solved (8.10).

- and one constraint for each cluster  $C \in \mathcal{C}$  to ensure that  $\mu_C$  lies on the probability simplex.

So in total, there are  $n = 3|\mathcal{VC}|$  variables, and  $m = |\mathcal{VC}| + |\mathcal{VT}| + |\mathcal{VA}| + |\mathcal{C}|$  constraints. Once again the components of  $A$  and  $b$  are all at most one, and now the components of the cost function  $c = \mathbf{1}$  are identically one. Furthermore, our objective is 1-strongly convex, so no additional multiplicative terms are required to convert an  $\epsilon$ -close solution in the sense of suboptimality, to an  $\epsilon$ -close solution in the sense of proximity to the true solution.

Therefore (8.14) can be solved in

$$O\left(|\mathcal{VC}||\mathcal{VA} + \mathcal{VC}|^\omega \log \frac{|\mathcal{VA} + \mathcal{VC}|}{\epsilon}\right) \subset \tilde{O}_{\text{BP}}(|\mathcal{VC} + \mathcal{VA}|^4 \log \frac{1}{\epsilon})$$

operations. □

**Theorem 8.10.** *Let  $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathbb{P}, \boldsymbol{\alpha}, \boldsymbol{\beta})$  be a proper discrete PDG with  $N = |\mathcal{X}|$  variables each taking at most  $V$  values and  $A = |\mathcal{A}|$  arcs, in which each component of  $\boldsymbol{\beta} \in \mathbb{R}^{\mathcal{A}}$  and  $\mathbb{P} \in \mathbb{R}^{|\mathcal{VA}|}$  is specified in binary with at most  $k$  bits. Suppose that*

$\gamma \in \{0^+\} \cup (0, \min_{a \in \mathcal{A}} \frac{\beta_a}{\alpha_a}]$ . If  $(\mathcal{C}, \mathcal{T})$  is a tree decomposition of  $(\mathcal{X}, \mathcal{A})$  of width  $T$  and  $\mu^* \in \mathbb{R}^{V\mathcal{C}}$  is the unique calibrated tree marginal over  $(\mathcal{C}, \mathcal{T})$  that represents the  $\hat{\gamma}$ -semantics of  $\mathcal{M}$ , then

(a) Given  $\mathcal{M}$ ,  $\gamma$ , and  $\epsilon > 0$ , we can find a calibrated tree marginal  $\epsilon$  close in  $\ell_2$  norm to  $\mu^*$  in time

$$\begin{aligned} O\left(|\mathcal{VA} + \mathcal{VC}|^4 \left(\log |\mathcal{VA} + \mathcal{VC}| + \log \frac{1}{\epsilon}\right) k^2 \log k\right) \\ \subseteq \tilde{O}\left(k^2 |\mathcal{VA} + \mathcal{VC}|^4 \log^{1/\epsilon}\right) \\ \subseteq \tilde{O}\left(k^2 (N+A)^4 V^{4(T+1)} \log^{1/\epsilon}\right). \end{aligned}$$

(b) The unique tree marginal closest to  $\mu^*$  in which every component is represented with a  $k$ -bit binary number, can be calculated in time<sup>1</sup>

$$\tilde{O}\left(k^2 |\mathcal{VA} + \mathcal{VC}|^4\right) \subseteq \tilde{O}\left(k^2 (N+A)^4 V^{4(T+1)}\right).$$

*Proof.* Suppose that the PDG has  $N$  variables (each of which can take at most  $V$  distinct values), and  $A$  hyperarcs, which together form a structure has tree-width  $T$ . Then each cluster (of which there are at most  $N$ ) can have at most  $T+1$  variables, and so can take at most  $V^T$  values. Therefore,  $|\mathcal{VC}| \leq NV^{T+1}$ . Since each arc must be entirely contained within some cluster,  $|\mathcal{VA}| \leq AV^T$ . So,  $|\mathcal{VA} + \mathcal{VC}| \leq (N+A)V^{T+1}$ .

By Lemma 8.15, we know that, for  $\gamma \in (0, \min_a \frac{\beta_a}{\alpha_a}]$ , a tree marginal  $\epsilon$ -close (in  $\ell_2$  norm) to the one that represents the unique distribution in the  $\hat{\gamma}$ -semantics can be found in time in time

$$O\left((N+A)^4 V^{4T+4} \log \left(V^{T+1}(N+A) \frac{1}{\epsilon} \frac{\beta^{\max}}{\gamma} + \log \frac{1}{p^{\min}}\right)\right).$$

Similarly, by Lemmas 8.14 and 8.16 a tree marginal  $\epsilon$ -close to the one representing the  $0^+$  semantics can be found in time

$$\begin{aligned} O\left(|\mathcal{VC} + \mathcal{VA}|^4 \log \frac{|\mathcal{VC} + \mathcal{VA}|}{\epsilon}\right) + O\left(|\mathcal{VC} + \mathcal{VA}|^4 \log \frac{|\mathcal{VC} + \mathcal{VA}| \beta^{\max}}{\epsilon \beta^{\min}}\right) \\ \subseteq O\left((N + A)^4 V^{4(T+1)} \log \left(V^{T+1}(N + A) \frac{1}{\epsilon} \frac{\beta^{\max}}{\beta^{\min}}\right)\right). \end{aligned}$$

Either way, a tree marginal  $\epsilon$ -close to the one that represents the  $\hat{\gamma}$ -semantics, for  $\gamma \in \{0^+\} \cup (0, \min_a \frac{\beta_a}{\alpha_a})$ , can be found in

$$O\left(|\mathcal{VC} + \mathcal{VA}|^4 \log \left(\frac{|\mathcal{VC} + \mathcal{VA}| \beta^{\max}}{\epsilon \beta^{\min}} + \log \frac{1}{p^{\min}}\right)\right) \subseteq \tilde{O}_{\text{BP}}\left(|\mathcal{VC} + \mathcal{VA}|^4 \log \frac{1}{\epsilon}\right)$$

arithmetic operations, each of which can be done in  $O(k \log k)$  time.

If  $\beta, \mathbb{P}$ , and  $\gamma$  are all binary numbers specified in  $k$  bits, then  $\log_2 \frac{\beta^{\max}}{\beta^{\min}} \leq 2k$  and  $\log \log \frac{1}{p^{\min}} \leq \log k + \log(2)$ . Thus, under these assumptions, such a tree marginal can be found in

$$O\left(|\mathcal{VC} + \mathcal{VA}|^4 \left(\log \frac{|\mathcal{VC} + \mathcal{VA}|}{\epsilon} + k + \log k\right) k \log k\right) \subseteq \tilde{O}\left(|\mathcal{VC} + \mathcal{VA}|^4 \log \left(\frac{1}{\epsilon}\right) k\right)$$

time. Finally, we prove part (b). The  $\infty$ -norm is smaller than the  $\ell_2$  norm, so if  $\|\mu - \mu^*\|_2 < 2^{-(k+1)}$ , then any change to  $\mu$  of size  $2^{-k}$  or larger will cause it to be further from  $\mu^*$ . Thus, selecting  $\epsilon = 2^{-(k+1)}$  produces the tree marginal of  $k$ -bit numbers that is closest to  $\mu^*$ . Plugging in this value of  $\epsilon$ , we find that finding it takes  $\tilde{O}(|\mathcal{VC} + \mathcal{VA}|^4 k^2)$  time.  $\square$

**Lemma 8.17.** *Let  $k \geq 1$  be a fixed integer, and  $\Phi, K_0, K_1, \dots, K_k$  be parameters. Given a procedure that produces  $\epsilon$ -approximate unconditional probabilities in  $O(\Phi \cdot (K_0 + \sum_{i=1}^k K_i \log^i \frac{1}{\epsilon}))$  time, we can approximate conditional probabilities  $\Pr(B|A)$  to within  $\epsilon$  in  $O(\Phi \cdot (K_0 \log \log \frac{1}{\Pr(A)} + \sum_{i=1}^k K_i \log^i \frac{1}{\epsilon \Pr(A)}))$  time.*

*Proof.* Let  $f$  be our algorithm for approximating unconditional probabilities. If  $A$  is an event and  $\epsilon > 0$ , we write  $f(A; \epsilon)$  for the corresponding approximation to

$\Pr(A)$ , which by definition satisfies

$$\Pr(A) - \epsilon \leq f(A; \delta) \leq \Pr(A) + \epsilon.$$

Now suppose that  $A$  and  $B$  are both events, and we want to find the conditional probability  $\Pr(B|A)$ . To do so, we can run the following algorithm.

---

```

1:  $\delta \leftarrow \epsilon;$ 
2: loop
3:   let  $a \leftarrow f(A; \delta);$ 
4:   if  $a > 2\delta$  then
5:     let  $\delta^* \leftarrow \epsilon(a - \delta)/3;$ 
6:     let  $p \leftarrow f(A; \delta^*)$  and  $q \leftarrow f(A \cap B; \delta^*);$ 
7:     return  $q / (p + \delta^*).$ 
8:   else
9:      $\delta \leftarrow \delta^2;$ 

```

---

**Proof of correctness.** We claim that the final output of the algorithm is within  $\epsilon$  of the true conditional probability  $\Pr(B|A)$ . In the first iteration in which  $a > 2\delta$  (line 4), we know that  $\delta \leq a - \delta \leq \Pr(A)$ .

By assumption,

$$\Pr(A) - \delta^* \leq p \leq \Pr(A) + \delta^* \quad \text{and} \quad \Pr(A \cap B) - \delta^* \leq q \leq \Pr(A \cap B) + \delta^*,$$

from which it follows that

$$\frac{\Pr(A \cap B) - \delta^*}{\Pr(A) + 2\delta^*} \leq \frac{q}{p + \delta^*} \leq \frac{\Pr(A \cap B) + \delta^*}{\Pr(A)}. \quad (8.21)$$

We now extend the bounds on  $q/(p + \delta^*)$  in both directions, starting with the

upper bound. Because  $a - \delta \leq \Pr(A)$ , the RHS of (8.21) is at most

$$\begin{aligned}\frac{\Pr(A \cap B) + \delta^*}{\Pr(A)} &= \Pr(B|A) + \frac{\delta^*}{\Pr(A)} = \Pr(B|A) + \frac{\epsilon(a - \delta)}{3\Pr(A)} \\ &\leq \Pr(B|A) + \frac{\epsilon\Pr(A)}{3\Pr(A)} < \Pr(B|A) + \epsilon.\end{aligned}$$

The analysis of the lower bound (the LHS of (8.21)) is slightly more complicated, but we still find that

$$\begin{aligned}\frac{\Pr(A \cap B) - \delta^*}{\Pr(A) + 2\delta^*} &= \Pr(B|A) - \frac{\mu^*(x, y)}{\Pr(A)} + \frac{\Pr(A \cap B) - \delta^*}{\Pr(A) + 2\delta^*} \\ &= \Pr(B|A) + \frac{-\Pr(A)\Pr(\overline{A \cap B}) - 2\delta^*\Pr(A \cap B) + \Pr(A)\Pr(\overline{A \cap B}) - \delta^*\Pr(A)}{\Pr(A)(\Pr(A) + 2\delta^*)} \\ &= \Pr(B|A) + \frac{-2\delta^*\Pr(B|A) - \delta^*}{\Pr(A) + 2\delta^*} \\ &= \Pr(B|A) - \delta^*\left(\frac{2\Pr(B|A) + 1}{\Pr(A) + 2\delta^*}\right) \\ &\geq \Pr(B|A) - \delta^*\frac{3}{\Pr(A) + \delta^*} \quad \left[ \text{since } \Pr(B|A) \leq 1, \text{ and thus } -2\Pr(B|A) \geq -2 \right] \\ &\geq \Pr(B|A) - \delta^*\frac{3}{\Pr(A)} \quad \left[ \text{as eliminating } \delta^* \text{ makes this more negative} \right] \\ &= \Pr(B|A) - \frac{\epsilon(a - \delta)}{3} \frac{3}{\Pr(A)} \quad \left[ \text{by definition of } \delta^* \right] \\ &\geq \Pr(B|A) - \frac{\epsilon\Pr(A)}{\Pr(A)} \quad \left[ \text{since } -(a - \delta) \geq -\Pr(A) \right] \\ &= \Pr(B|A) - \epsilon.\end{aligned}$$

These two arguments extend the bounds of (8.21) in both directions. Chaining all of these inequalities together, we have shown that our procedure returns a number output satisfying

$$\Pr(B|A) - \epsilon \leq \text{output} \leq \Pr(B|A) + \epsilon,$$

and hence calculates the desired conditional probability to within  $\epsilon$ .

**Analysis of Runtime.** Let  $m$  denote the total number of iterations of the algorithm. We deal with the simple case of  $m = 1$  separately. If  $m = 1$ , then

already in the first iteration  $a > 2\delta = 2\epsilon$ , so by definition  $\delta^* > \frac{1}{3}\epsilon^3$ . Line 6 is just two calls to the procedure, and takes

$$\begin{aligned} O\left(\Phi\left(K_0 + \sum_{i=1}^k K_i \log^i \frac{1}{\delta^*}\right)\right) &= O\left(\Phi\left(K_0 + \sum_{i=1}^k K_i \log^i \frac{3}{\epsilon^3}\right)\right) \\ &\subseteq O\left(\Phi\left(K_0 + \sum_{i=1}^k K_i \log^i \frac{1}{\epsilon}\right)\right) \text{ time.} \end{aligned} \quad (8.22)$$

Now consider the case where  $m > 1$ . Observe that, in the final iteration,  $\delta = \epsilon^{2^{m-1}}$ . The procedure halts when  $a > 2\delta$ , and the smallest possible value of  $a$  that our approximation can return is  $\Pr(A) - \delta$ . Thus, the procedure must halt by the time  $\Pr(A) > 3\delta = 3\epsilon^{2^{m-1}}$ . On the other hand, since  $m - 1$  iterations are not enough to ensure termination, it must be that  $\Pr(A) - \delta' \leq 2\delta'$ , where  $\delta' := \epsilon^{2^{m-2}}$  is the value of  $\delta$  in the penultimate iteration. Together, these two facts give us a relationship between  $m$  and  $\Pr(A)$ :

$$\begin{aligned} 3\epsilon^{2^{m-2}} &\geq \Pr(A) > 3\epsilon^{2^{m-1}} \\ \iff -\log_2 3 - 2^{m-2} \log_2 \epsilon &\leq -\log_2 \Pr(A) < -\log_2 3 - 2^{m-1} \log_2 \epsilon \\ \iff 2^{m-2} &\leq \left(\log_2 \frac{3}{\Pr(A)}\right) / \log_2(1/\epsilon) < 2^{m-1} \end{aligned} \quad (8.23)$$

In particular, the first inequality tells us that the number of required iterations is at most

$$m \leq 2 + \log_2 \log_2 \frac{3}{\Pr(A)} - \log_2 \log_2 \frac{1}{\epsilon} = 2 + \log_2 \log_\epsilon \frac{\Pr(A)}{3}.$$

Across all iterations, the total cost of line 3 is on the order of

$$\begin{aligned} m\Phi K - \Phi \sum_{i=1}^k K_i \sum_{j=1}^m \log^i(\epsilon^{2^{j-1}}) \\ = m\Phi K - \Phi \sum_{i=1}^k K_i \log^i(\epsilon) \sum_{j=0}^{m-1} 2^{kj} \end{aligned}$$

$$\begin{aligned}
&= m\Phi K + \Phi \sum_{i=1}^k K_i \log^i \frac{1}{\epsilon} \frac{2^{im} - 1}{2^i - 1} \\
&< \left( \log \log \frac{3}{\Pr(A)} - \log \log \frac{1}{\epsilon} \right) \Phi K_0 \\
&\quad + \Phi \sum_{i=1}^k K_i \log^i \left( \frac{1}{\epsilon} \right) \cdot \left[ 4^i \left( \log^i \frac{3}{\Pr(A)} \right) / \log^i \left( \frac{1}{\epsilon} \right) \right] / (2^i - 1) \\
&\leq \Phi K_0 \log \log \frac{3}{\Pr(A)} + \Phi \sum_{i=1}^k K_i \frac{4^i}{2^i - 1} \log^i \frac{3}{\Pr(A)} \\
&\subseteq O\left(\Phi \cdot \left( K_0 \log \log \frac{1}{\Pr(A)} + \sum_{i=1}^k K_i \log^i \frac{1}{\Pr(A)} \right)\right). \tag{8.24}
\end{aligned}$$

Line 6 is the last part of the procedure that incurs a nontrivial cost. The procedure executes it one time, in the final iteration. Because  $a > 2\delta$  at this point, we know that

$$\delta^* = \frac{\epsilon}{3}(a - \delta) > \frac{\epsilon}{3}\delta = \frac{\epsilon}{3}\epsilon^{2^{m-1}} = \frac{\epsilon}{3} \frac{9}{9} \epsilon^{2(2^{m-2})} = \frac{\epsilon}{27} \left( 3\epsilon^{2^{m-2}} \right)^2 \geq \frac{\epsilon}{27} \Pr(A)^2.$$

Thus line 6 requires time

$$O\left(\Phi \cdot \left( K_0 + \sum_{i=1}^k K_i \log^i \frac{27}{\Pr(A)^2 \epsilon} \right)\right) \subseteq O\left(\Phi K_0 + \Phi \sum_{i=1}^k K_i \left( \log \frac{1}{\Pr(A)} + \log \frac{1}{\epsilon} \right)^i\right). \tag{8.25}$$

Summarizing, the total running time is (at most) the sum of (8.22), (8.24), and (8.25), or explicitly,

$$O\left(\Phi \cdot \left( K \log \log \frac{1}{\Pr(A)} + \sum_{i=1}^k K_i \log^i \frac{1}{\epsilon \Pr(A)} \right)\right).$$

□

### Theorem 9.1. \*

*Proof.* Theorem 8.10 gives us an approximation to a calibrated tree marginal that represents the distribution of interest, and Lemma 8.17 allows us to approximate

conditional probabilities once we can approximate unconditional ones. The final ingredient is to approximate unconditional probabilities using an approximate tree marginal.

Concretely, suppose that we are looking to find  $\mu^*(X=x)$ , where  $\mu^* \in \llbracket m \rrbracket_\gamma$ . Once we have a calibrated tree marginal  $\mu$  that represents  $\mu^*$ , calculating a marginal  $\mu^*(X=x)$  (exactly) from  $\mu$  can be done with standard methods ([Koller and Friedman 2009](#), §10.3.3). In the worst case, it requires taking a marginal of every cluster, which can be done in  $O(|\mathcal{VC}|) \subseteq O(NV^{T+1})$  arithmetic operations.

The wrinkle is that  $\mu$  only *approximately* represents  $\mu^*$ , in the sense that there is some  $\mu^*$  that does represent  $\mu^*$  such that the L2 norm of  $\mu^* - \mu$  is small. As usual, we write  $\mu_C$  for the components of  $\mu$  that are associated with cluster  $C$ . For each  $C \in \mathcal{C}$ , let  $E_C$  denote the event that  $(X \cap C) = x|_C$ . That is, the variables of  $X$  that lie in cluster  $C$  take the values prescribed by  $x$ . Then

$$\begin{aligned} |\Pr_{\mu}(X=x) - \Pr_{\mu^*}(X=x)| &\leq \sum_{C \in \mathcal{C}} |\Pr_{\mu_C^*}(E_C) - \Pr_{\mu_C}(E_C)| \\ &\leq \sum_{C \in \mathcal{C}} \|\mu_C^* - \mu_C\|_1 = \|\mu^* - \mu\|_1. \end{aligned}$$

Applying the L2-L1 norm inequality to the vector  $\mu - \mu^*$ , we find

$$\|\mu - \mu^*\|_1 \leq \|\mu - \mu^*\|_2 \sqrt{|\mathcal{VC}|} \leq \sqrt{NV^{T+1}} \|\mu - \mu^*\|_2.$$

Thus, to answer unconditional queries about  $X$  within (absolute) precision  $\epsilon$ , it suffices to find a tree marginal within  $\epsilon/\sqrt{NV^{T+1}}$  of  $\mu^*$  by L2 norm.

From the proof of [Theorem 8.10](#), we know that we can find such a  $\mu$  in

$$\begin{aligned} O\left((N+A)^4 V^{4(T+1)} \log\left(\frac{(N+A)^4 V^{4(T+1)} \cdot N^{\frac{1}{2}} V^{\frac{T+1}{2}} \beta^{\max}}{\epsilon} + \log \frac{1}{p^{\min}}\right)\right) \\ \subseteq \tilde{O}\left((N+A)^4 V^{4(T+1)} \left(\log \frac{1}{\epsilon} + \log \frac{\beta^{\max}}{\beta^{\min}}\right)\right) \\ \subseteq \tilde{O}_{\text{BP}}\left(|\mathcal{VA} + \mathcal{VC}|^4 \log \frac{1}{\epsilon}\right) \end{aligned}$$

arithmetic operations, which dominates the number of operations required to then find the marginal probability  $\Pr_{\mu}(X=x)$  given the tree marginal  $\mu$ . Thus, the complexity of finding unconditional probabilities is the same. The arithmetic operations need to be done to precision at most  $k \in O(\log 1/\epsilon)$ , and can be done in time  $O(k \log k)$ . Thus, unconditional inference can be done in

$$\begin{aligned} O\left(\frac{(N+A)^{4.5}}{V^{4.5(T+1)}} \log\left(\frac{(N+A)^4 V^{4(T+1)} \cdot N^{\frac{1}{2}} V^{\frac{T+1}{2}} \beta^{\max}}{\epsilon} + \log \frac{1}{p^{\min}}\right) \log \frac{1}{\epsilon} \log \log \frac{1}{\epsilon}\right) \\ \subseteq \tilde{O}\left((N+A)^4 V^{4(T+1)} \left(\log \frac{1}{\epsilon} + \log \frac{\beta^{\max}}{\beta^{\min}}\right) \log \frac{1}{\epsilon}\right) \text{ time.} \end{aligned}$$

Now that we have characterized the cost of unconditional inference, we can apply [Lemma 8.17](#) with  $\Phi := (N + A)^4 V^{4(T+1)}$ ,  $k = 2$ ,  $K_0 = 0$ ,  $K_1 := \log \Phi + \log \frac{\beta^{\max}}{\beta^{\min}} + \log \log \frac{1}{p^{\min}}$ , and  $K_2 = 1$  to find that conditional probabilities can be found in

$$\tilde{O}\left((N+A)^4 V^{4(T+1)} \log \frac{1}{\epsilon \mu^*(x)} \left(\log \frac{\beta^{\max}}{\beta^{\min}} + \log \frac{1}{\epsilon \mu^*(x)}\right)\right) \text{ time,}$$

where  $\mu^*(x)$  is shorthand for  $\mu^*(X=x)$ . □

## 8.B The Convex-Concave Procedure, and Implementation Details

Optimization problems (8.6) and (8.13) can be extended to apply slightly more broadly. There are some cases where there is a unique optimal distribution but

$\gamma$  is large enough that  $\beta \not\geq \gamma\alpha$ . In these cases, our convex program will fail to satisfy the dcp requirements, and so we cannot compile it to an exponential conic program—but it turns out to still be a useful building block. We now describe how we can still do inference in some of these cases with the convex-concave procedure, or CCCP (Yuille and Rangarajan 2003). This will give us a local minimum of the PDG scoring function  $\llbracket m \rrbracket_\gamma$ , without requiring us to write this scoring function in a way that proves its convexity, (as is necessary in order to specify a disciplined convex program). At this point, if we happen to know that the problem is convex (or even just pseudo-convex) for other reasons, then finding this distribution suffices for inference. We now describe how this can be done in more detail.

Suppose  $\beta_a < \gamma\alpha_a$  some  $a \in \mathcal{A}$ . In this case  $\llbracket m \rrbracket_\gamma$  may not be convex, in general.<sup>6</sup> However, we do know how to decompose  $\llbracket m \rrbracket_\gamma$  into a sum of a convex function  $f(\mu)$  and a concave one  $g(\mu)$ . Concretely: each term on the second line of (8.5) is either convex or concave, depending on the sign of the quantity  $\gamma\alpha_a - \beta_a$ . Once we sort the terms into convex terms  $f(\mu)$  and strictly concave terms  $g(\mu)$ , the CCCP tells us to repeatedly solve  $f$  plus a linear approximation to  $g$ . In more detail, the algorithm proceeds as follows. First, choose an initial guess  $\mu_0$ , and iteratively use the convex solver as in the main part of the chapter to compute

$$\mu_{t+1} := \arg \min_{\mu} f(\mu) + (\mu - \mu_t)^T \nabla g(\mu_t).$$

This can be slow because each iteration of the solver is expensive. Still, it is

---

<sup>6</sup>Consider the PDG  $(\rightarrow X, Y \leftarrow)$  for instance, which has arcs to  $X$  and  $Y$ , both with  $\alpha = 1$  and  $\beta = 0$ . The minimizers of  $\llbracket \rightarrow X, Y \leftarrow \rrbracket_\gamma$  are the distributions that make  $X$  and  $Y$  independent. It is easily seen that this set is not convex:  $X$  and  $Y$  are independent if either variable is deterministic, and every distribution is a convex combination of deterministic distributions.

guaranteed to make progress, since

$$\begin{aligned}
f(\mu_{t+1}) + g(\mu_{t+1}) &< f(\mu_{t+1}) + (\mu_{t+1} - \mu_t)^\top \nabla g(\mu_t) + g(\mu_t) \\
&\leq f(\mu_t) + (\mu_t - \mu_t)^\top \nabla g(\mu_t) + g(\mu_t) \\
&= f(\mu_t) + g(\mu_t).
\end{aligned}$$

Furthermore, because in our case  $g$  is bounded, the process eventually converges a local minimum of  $\llbracket m \rrbracket_\gamma$ . This alone, however, is not sufficient for inference, because we may not be able to use this local minimum to answer queries in a way that is true of *all* minimizing distributions. But, if it happens there is a unique local minimum, then the CCCP will find it, leading to an inference procedure.

Notice that if  $\beta \geq \gamma\alpha$ , then the concave part  $g$  is identically zero, and CCCP converges after making just one call to the convex solver. Therefore, in the cases we could already handle, this extension reduces to the algorithm we described before. For this reason, all of our code that handles problems (8.6) and (8.13) is augmented with the CCCP.

Compared to the black-box optimization baselines (Adam and LBFGS), which also only find one minimum, the CCCP still has some advantages. One can see in [Figure 8.C.4](#), for example, that when  $\gamma = 2 > 1 = \max_a (\beta_a/\alpha_a)$ , CCCP performs better than the baselines. In fact, the CCCP-augmented solver could probably even higher accuracy, if were we not limiting it to a maximum of only five iterations.

## 8.C Details on the Empirical Evaluation

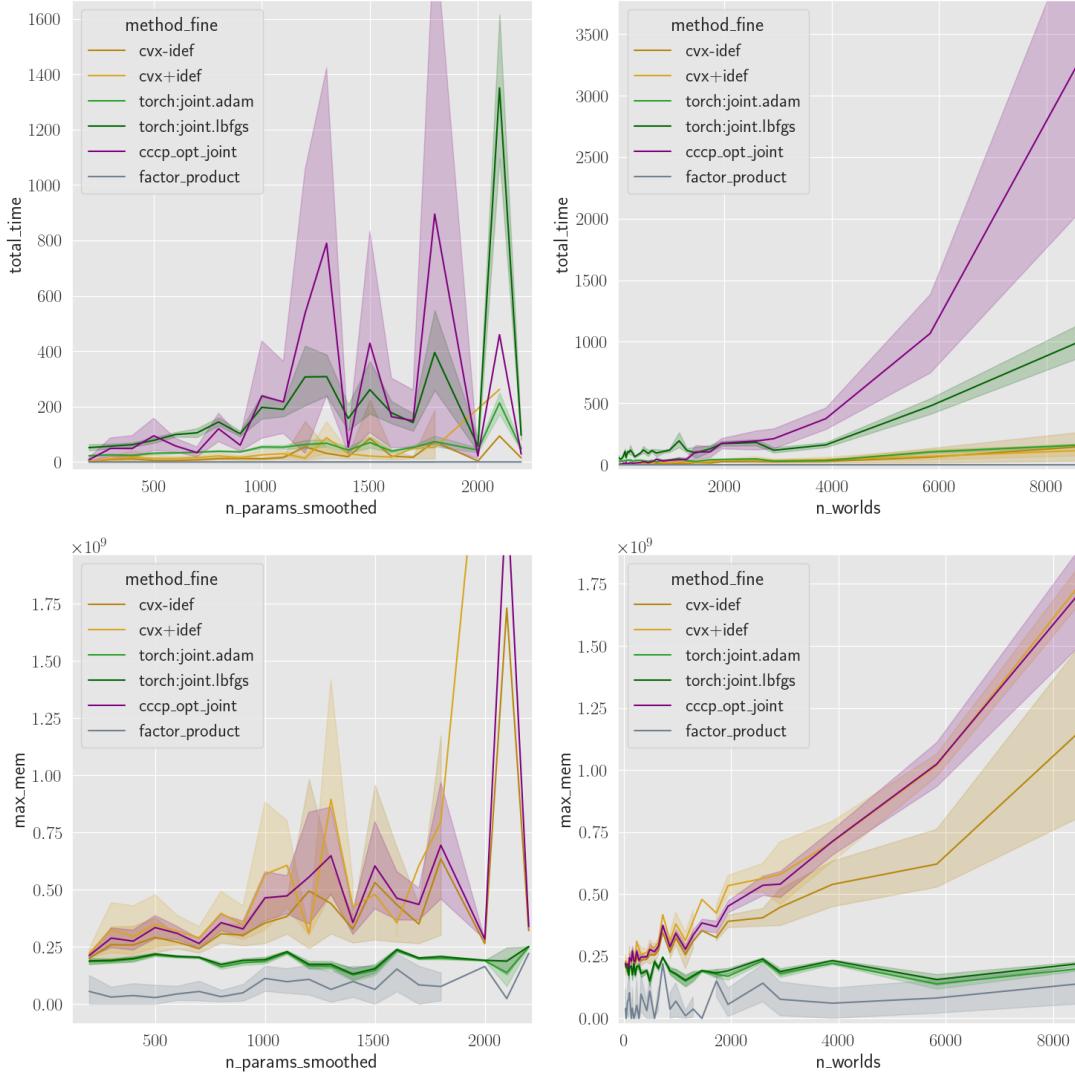
Imagine a very steep  $V$ -shaped canyon, and inside a small slow-moving stream at a gentle incline. The end of the river may be very far away, and the whole landscape may be smooth and strongly convex, but the gradient will still almost always point perpendicular to it, and rather towards the center of the river. This intuition may help explain why, even though  $\|\mathbf{m}\|_\gamma$  is infinitely differentiable in  $\mu$  and  $\gamma$ -strongly convex, it can still be challenging to optimize, especially when the  $\beta$ 's are very different, or when  $\gamma$  is small. For example, a solution to (8.10) finds a minimizer of  $OInc$ , but such minimizers may be very far away from  $\|\mathbf{m}\|_{0+}^*$ , despite sharing an objective value.

We now see how this is true even when working with very small PDGs and joint distributions.

### 8.C.1 Synthetic Experiment: Comparison with Black-Box Optimizers, on Joint Distributions.

Here is a more precise description of our first synthetic experiment, on joint distributions, which contrasts the convex optimization approaches of Section 8.3 with black-box optimizers.

- generate 300 PDGs, each of which has the following quantities, to each of which we choose the following natural numbers uniformly at random:
  - $N \in \{5, \dots, 9\}$  of variables (so that  $\mathcal{X} := \{1, \dots, N\}$ ),
  - $V_X \in \{2, 3\}$  values per variable (so that  $|\mathcal{V}X| = V_X$  for each  $X \in \mathcal{X}$ )



*Figure 8.C.1:* Resource costs for the joint-distribution optimization setting of Section 8.3. We measure computation time (total\_time, top) and maximum memory usage (max\_mem, bottom) for the various optimization methods (by color), as the size of the PDG increases, as measured by the number of parameters in the PDG ( $n_{\text{params}} = \mathcal{V}\mathcal{A}$ , left), and the size of a joint distribution over its variables ( $n_{\text{worlds}} = \mathcal{V}\mathcal{X}$ , right). Note that the convex solvers for the 0 and  $0^+$  semantics are significantly faster than LBFGS, and on par with Adam. However, all three convex-solver based approaches require significantly more memory than the black-box optimizers.

- $A \in \{7, \dots, 14\}$  hyperarcs, each  $a \in \{1, \dots, A\} =: \mathcal{A}$  of which has
  - $N_a^S \in \{0, 1, 2, 3\}$  sources, and
  - $N_a^T \in \{1, 2\}$  targets.
- For each arc  $a \in \mathcal{A}$ ,  $N_a^S$  of the  $N$  variables are chosen without replacement to be sources  $S_a \subseteq N$ , and  $N_a^T$  of remaining variables are chosen to be targets. Finally, to each value of  $S_a$  and  $T_a$ , a number  $p_{a,s,t} \in [0, 1]$  is chosen uniformly at random, and the cpd

$$\mathbb{P}_a(T_a=t \mid S_a=s) = \frac{p_{a,s,t}}{\sum_{t' \in \mathcal{V}(T)} p_{a,s,t'}} \quad \text{is given by normalizing appropriately.}$$

This defines a PDG  $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathbb{P}, \mathbf{1}, \mathbf{1})$ , that has  $\alpha = \beta = 1$ , which will allow us to compare against belief propagation and other graphical models at  $\gamma = 1$ . The complexity of this PDG is summarized by two numbers:

- $n\_params := \mathcal{V}\mathcal{A}$ , the total number of parameters in all cpds of  $\mathcal{M}$ , and
- $n\_worlds := \mathcal{V}\mathcal{X}$ , the dimension of joint distributions over  $\mathcal{M}$ 's variables.
- Run MOSEK on (8.4) to find a distribution that minimizes  $OInc$ ; we refer to this method as `cvx-idef`
- Use the result to run MOSEK on (8.9) to find the special distribution  $[\mathcal{M}]_{0+}^*$ ; we refer to this method as `cvx+idef`. These names are due to the fact that  $SDef$  is called  $IDef$  in previous work (Richardson and Halpern 2021; Richardson 2022); thus, this refers to using the convex solver to compute minimizers of  $OInc$  with and without considering  $IDef$ .
- Run the pytorch baselines. Let  $\theta = [\theta_x]_{x \in \mathcal{V}\mathcal{X}} \in \mathbb{R}^{\mathcal{V}\mathcal{X}}$  be a vector of optimization variables, and choose a representation of the joint distribution,

either:

$$\mu_\theta(\mathbf{x}) = \frac{\max\{\theta_{\mathbf{x}}, 0\}}{\sum_{\mathbf{y} \in \mathcal{VX}} \max\{\theta_{\mathbf{y}}, 0\}} \quad (\text{renormalized simplex})$$

or

$$\mu_\theta(\mathbf{x}) = \frac{\exp(\theta_{\mathbf{x}})}{\sum_{\mathbf{y} \in \mathcal{VX}} \exp(\theta_{\mathbf{y}})} \quad (\text{Gibbs})$$

See [Figure 8.C.2](#) for a visual representation of the (relatively minor) effect of this choice.

- For each value of the trade-off parameter  $\gamma \in \{0, 10^{-8}, 10^{-4}, 10^{-2}, 1\}$ , and each learning rate  $lr \in 1E - 3, 1E - 2, 1E - 1, 1E0$ , and each optimizer  $opt \in \{\text{adam}, \text{L-BFGS}\}$ , run  $opt$  over the parameters  $\theta$  to minimize  $\llbracket m \rrbracket_\gamma(\mu_\theta)$  until convergence (or a maximum of 1500 iterations)
- We collect the following data about the resulting distribution and the process of computing it:
  - the total time taken to arrive at  $\mu$ ;
  - the maximum memory taken by the process computing  $\mu$ ;
  - the objective and its component values:

$$\begin{aligned} \text{inc} &:= SDef_m(\mu), & \text{obj} &:= OInc_m(\mu) + \gamma SDef_m(\mu) = \llbracket m \rrbracket_\gamma(\mu) \\ \text{idef} &:= SDef_m(\mu), \end{aligned}$$

The numbers can then be recreated by running our experimental script as follows:

```
python random_expts.py -N 300 -n 5 9 -e 7 14 -v 2 3
--ozrs lbfsgs adam
--learning-rates 1E0 1E-1 1E-2 1E-3
--gammas 0 1E-8 1E-4 1E-2 1E0
--num-cores 20
--data-dir random-joint-data
```

which creates a folder called `random-joint-data`, and fills it with `.mpt` files corresponding to each distribution and the method / parameters that gave rise to it.

**Analyzing the Results.** Look at Figure 8.C.1. Our theoretical analysis, and in particular the proof of Lemma 8.14, suggest that the magnitudes of  $\mathcal{V}\mathcal{X}$  and  $\mathcal{V}\mathcal{A}$  play similar roles in the asymptotic complexity of PDG inference. Our experiments reveal that, at least for random PDGs, the number of worlds is the far more important of the two; observe how much more variation there is on the left side of the figure than the right—and now note that the left side has been smoothed, while the right side has not. The black-box py-torch based approaches clearly have an edge in that they can handle larger models, as evidenced by the cut-offs on the right-hand side of Figure 8.C.8, when with 5GB memory.

Note that the exponential-cone-based methods for the observational limit (gold) are actually faster than L-BFGS (the black-box optimizer with the lowest gap), and also seem to be growing at a slower rate. However, they use significantly more memory, and cannot handle large models. In addition to being faster, our techniques also seem to be more precise; they achieve objective values that are consistently much better than the black-box methods.

Now look at Figure 8.C.3, which contains a break-down of the information in Figure 8.1. The bottom half of the figure is just the same information, but with

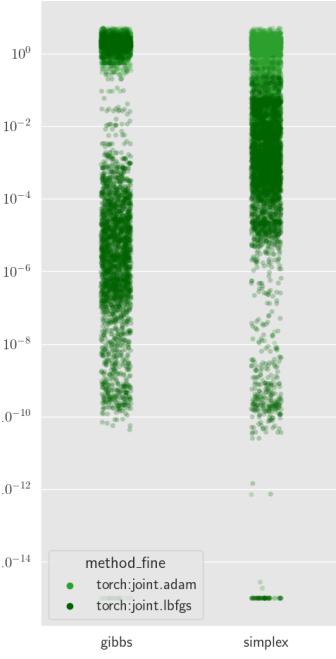
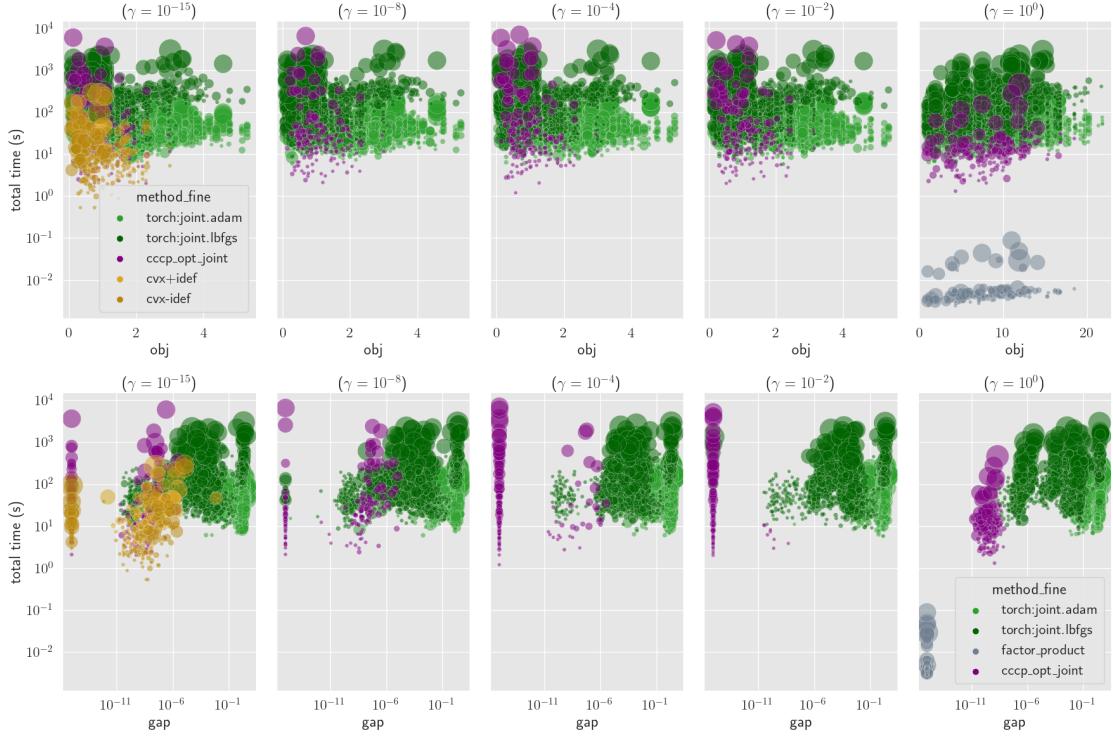


Figure 8.C.2: differences in performance between the Gibbs and simplex parameterizations of probabilities.

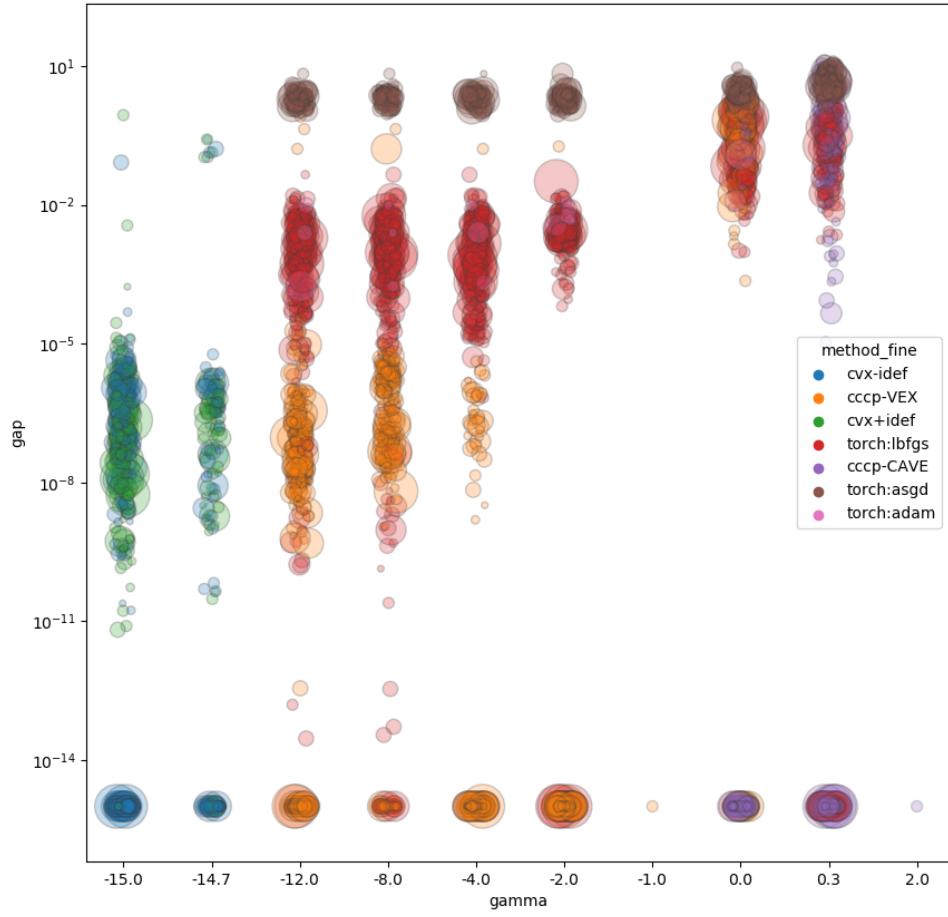


*Figure 8.C.3:* An un-compressed version of the information in [Figure 8.1](#), that groups by the value of  $\gamma$ , and also gives the absolute values of the objectives (top row) in addition to the relative gaps (bottom row).

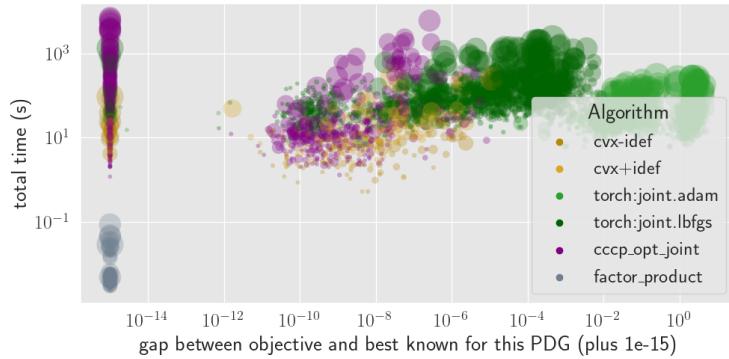
each value of  $\gamma$  separated out, so that the special cases of the factor product and  $0^+$  inference become clear, while the top half shows why it's more important to look at the gap than the actual objective value for these random PDGs. [Figure 8.C.3](#) also makes it clearer how larger problems take longer, and especially so for `cccp` (violet), which solves the most complex version of the problem (8.6).

## 8.C.2 Synthetic Experiment: Comparing with Black-Box Optimizers, on Tree Marginals

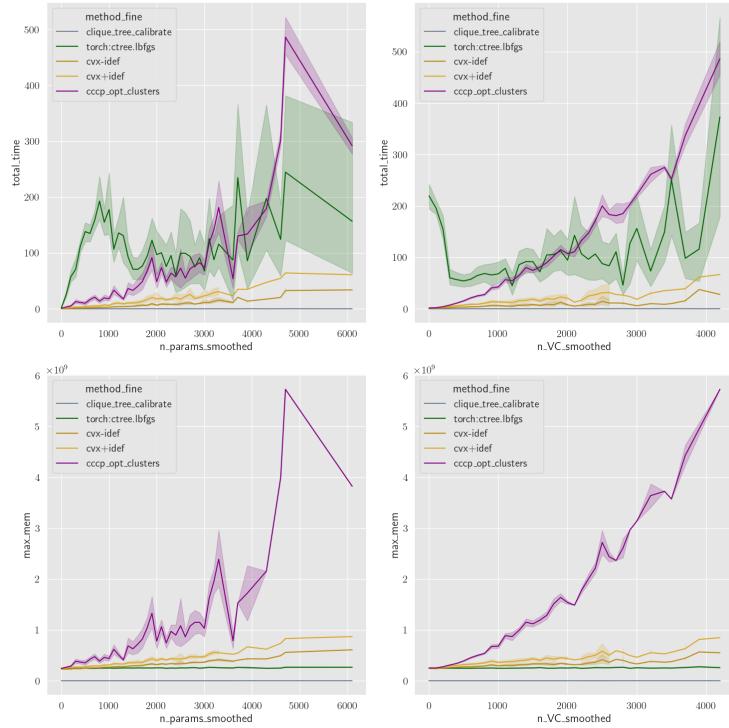
1. Choose a number of variables  $N \in \{8, \dots, 32\}$ , and a treewidth  $k \in \{1, \dots, 4\}$  uniformly at random. Then draw a random  $k$ -tree and corre-



*Figure 8.C.4:* A graph of the gap (the difference between the attained objective value, and the best objective value obtained across all methods for that value of  $\gamma$ ), as  $\gamma$  varies. The x-axis is  $\log_{10}(\gamma + 10^{-15})$ . As before, colors indicate the optimization method, and the size of the circle illustrates the number of optimization variables (i.e., the number of possible worlds). `cvx-idef` corresponds to just solving (8.4), and `cvx+idef` corresponds to then solving problem (8.9) afterwards. The CCCP runs are split into regimes where the entire problem is convex ( $\gamma \leq 1$ , labeled `cccp-VEX`), and the entire problem is concave ( $\gamma > 1$ , labeled `cccp-CAVE`). The optimization approaches `opt_dist` are split into three different optimizers: LBFGS, Adam, and also a third one that performs relatively poorly: accelerated gradient descent. Note that for small  $\gamma$ , the exponential-cone based methods significantly outperform the gradient-based ones.



*Figure 8.C.5:* An analogue of [Figure 8.1](#), for the cluster setting. Note that there is even more separation between the exponential-cone based approaches, and the black-box optimization based ones. The new grey points on the bottom correspond to belief propagation, which is both faster and typically the most accurate.



*Figure 8.C.6:* Resource costs for the cluster setting. Once again, the *OInc*-optimizing exponential cone methods are in gold, the small-gamma and CCCP is in violet, and the baselines are in green. The bottom line is belief propagation, which is significantly faster and requires very little memory, but also only gives the correct answer under very specific circumstances.

sponding tree of clusters  $(\mathcal{C}, \mathcal{T})$ , as follows:

- (a) Initialize  $G \leftarrow K_{k+1}$  to a complete graph on  $k + 1$  vertices, and  $\mathcal{C} \leftarrow \{K_{k+1}\}$  to be set containing a single cluster, and  $\mathcal{T} \leftarrow \emptyset$ .
  - (b) Until there are  $N$  vertices: add a new vertex  $v$  to  $G$ , then randomly select a size  $k$ -clique (fully-connected subgraph)  $U \subset G$ , and add edges between  $v$  and every vertex  $u \in U$ . Add  $U \cup \{v\}$  to  $\mathcal{C}$ , and add edges to every other cluster  $C \in \mathcal{C}$  such that  $U \subset C$ .
2. Draw the same parameters  $V_X \in \{2, 3\}$ ,  $A \in \{8, \dots, 120\}$ ,  $N_a^S \in \{0, 1, 2, 3\}$ , and  $N_a^T \in \{1, 2\}$  as in [Section 8.C.1](#) uniformly at random. While  $N_a^S + N_a^T > k + 1$ , for any  $a$ , resample  $N_a^S$  and  $N_a^T$ .
  3. Form a PDG whose structure  $\mathcal{A}$  can be decomposed by  $(\mathcal{C}, \mathcal{T})$ , as follows: for each edge  $a \in \mathcal{A}$ , sample a cluster  $C \in \mathcal{C}$  uniformly at random; then select  $N_a^S$  nodes from that cluster without replacement as sources, and  $N_a^T$  nodes as targets; this is possible because each cluster has  $k + 1$  nodes, and  $N_a^S + N_a^T \leq k + 1$  by construction.
  4. Fill in the probabilities by drawing uniform random numbers and re-normalizing, just as before, to form a PDG  $\mathcal{M}$
  5. The black-box optimization baselines work in much the same way also, although now the optimization variables include not one distribution  $\mu$  but a collection  $\boldsymbol{\mu}$  of them; this time, we use only the simplex representation of  $\boldsymbol{\mu}_\theta$ . More importantly, we want these clusters to share appropriate marginals; to encourage this, we add a terms to the loss function, so overall, it is

$$\ell(\theta) := \llbracket \mathcal{M} \rrbracket_\gamma(\boldsymbol{\mu}_\theta) + \sum_{C-D \in \mathcal{T}} \exp \left( \sum_{w \in V(C \cap D)} (\mu_C(C \cap D=w) - \mu_D(C \cap D=w))^2 \right) - 1.$$

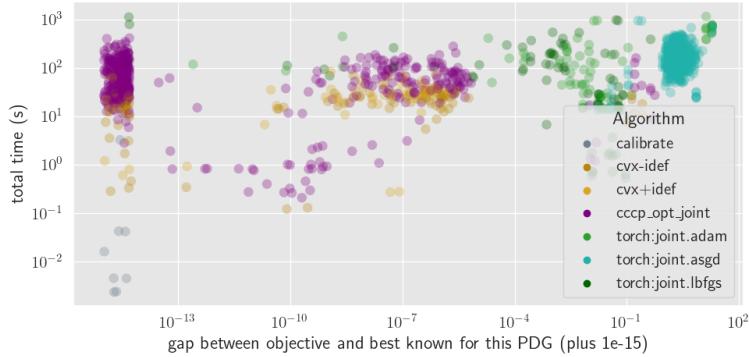


Figure 8.C.7: Gap vs inference time for the small PDGs in the `bnlearn` repository

This is admittedly pretty ad-hoc; the point is just that it is zero and does not contribute to the gradient if  $\mu_\theta$  is calibrated, and otherwise quickly becomes overwhelmingly important.

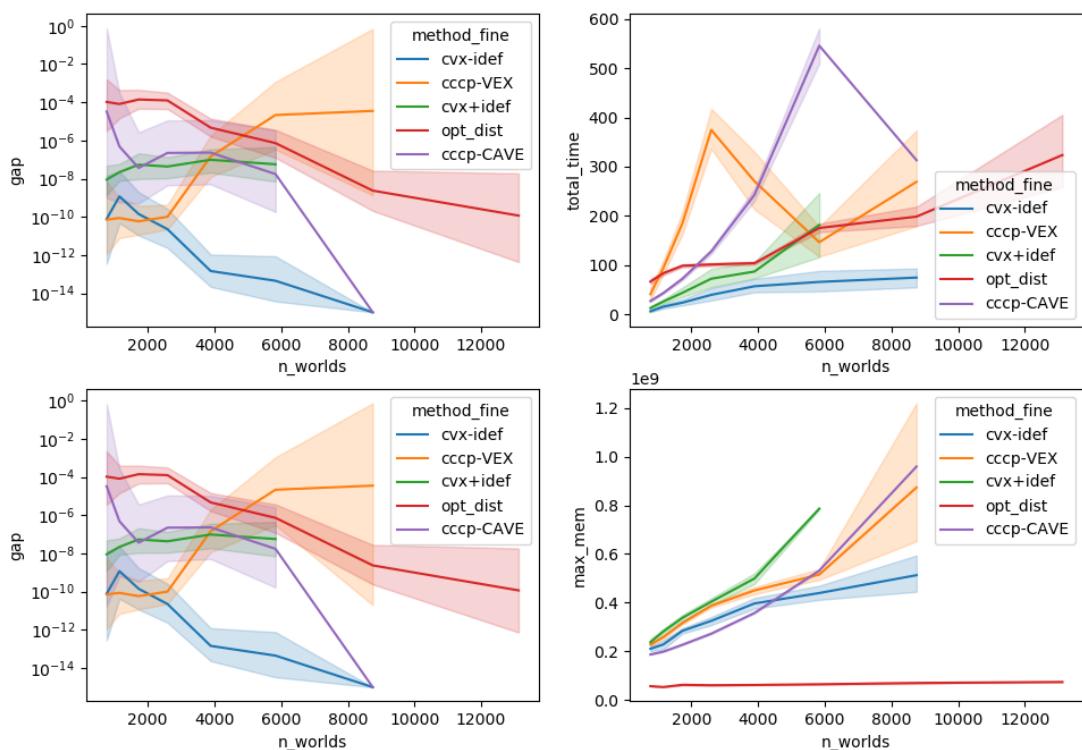
**Analyzing the Results.** Observe in Figure 8.2 that the separation between the tree marginal convex solver and the black-box algorithms is even more distinct. This is because, in this case, the penalty for violating constraints was too small, and the optimization effort was largely wiped out by the calibration before evaluation.

This illustrates another general advantage that the convex solver has over black-box optimizers: it is much less brittle and reliant and exactly tuning parameters correctly. Note that even in this minimal example, there were many hyper-parameters that require tuning: the regularization strengths that enforce soft constraints (tree marginal calibration, normalization), as well as learning rate, not to mention various other structural choices: the optimizer, the representation of the distribution, and the maximum number of iterations, none of which are clear-cut choices, but rather require first being tuned to the data. While the convex solver does have internal parameters (tolerances and such) these do not

need to be tuned to the problem under normal circumstances.

### 8.C.3 Comparing to Belief Propagation, on Tree Marginals.

Since PDGs generalize other graphical models, one might wonder how our method stacks up against algorithms tailored to the more traditional models. In brief: our algorithm is much slower, and only handle much smaller networks. Concretely, our methods can handle all of the “small” networks, and some of the “medium” ones, from the `bnlearn` repository. In these cases, we have verified that the two methods yield the same results. [Figure 8.C.7](#) contains the analogue of [Figures 8.1](#) and [8.2](#) for the Bayesian Nets. This graph looks qualitatively quite similar to the other graphs we’ve seen, suggesting that the results in our synthetic experiments hold more broadly for small real-world models as well.



*Figure 8.C.8:* A variant of Figure 8.C.1, with gap (accuracy) information on the left, and slightly different parameter settings.

# CHAPTER 9

## LOWER BOUNDS, AND THE DEEP CONNECTION BETWEEN INCONSISTENCY AND INFERENCE

On one hand, PDGs are powerful modeling tools, and representations of uncertainty. On the other, PDGs are universal loss functions. This chapter describes a deep connection between two roles of PDGs. We start with a suggestive observation: that the semantics by which PDGs specify a distribution is equivalent to the semantics that simply evaluates a PDG’s degree of inconsistency ([Section 9.1](#)). We then build on this semantic connection by describing a surprisingly strong algorithmic one: an inconsistency oracle can be used to perform inference ([Section 9.3](#)) in almost-linear time. Thus, the two views of PDGs are essentially equivalent—not only semantically, but algorithmically as well. This leads to a philosophically profound result: precisely quantifying one’s degree of inconsistency is no easier than optimally resolving those inconsistencies.

To get there, we first need to formally describe the two computational problems of interest ([Section 9.2](#)) as approximate inference problems, which allows us to also formally finish the upper bounds developed in [Chapter 8](#).

### 9.1 A Semantic Connection

Recall that we defined the inconsistency  $\langle\!\langle \cdot \rangle\!\rangle$  of a PDG is the score of its best-scoring distribution. Although we defined  $\langle\!\langle \cdot \rangle\!\rangle$  in terms of the scoring-function semantics ([9.1, left](#)), it worth noting that the reverse is also possible: the scoring function can be defined in terms of inconsistency ([9.1, right](#)).

$$\langle\!\langle m \rangle\!\rangle_\gamma = \inf_{\mu \in \Delta_{VX}} [m]_\gamma(\mu) \quad [m]_\gamma = \langle\!\langle m + \mu! \rangle\!\rangle_\gamma. \quad (9.1)$$

Although the technical details are not surprising, this fact seems quite surprising if we take a few steps back and reflect on its meaning at a higher level. In general, a function  $f : X \times Y \rightarrow \mathbb{R}$  typically contains strictly more information than its minimizers ( $x \mapsto \arg \min_y f(x, y) \subseteq Y$ ) :  $X \rightarrow 2^Y$ , which in turn one typically thinks of as containing more information than its minimum value

$$\min_Y f = (x \mapsto \min_y f(x, y)) : X \rightarrow \mathbb{R}.$$

In this case, however, when we take  $X$  to be the set of PDGs,  $Y$  to be the set of joint distributions, and  $f$  to be the scoring function,  $\min_Y f$  is equivalent to  $f$  itself, in the sense that either can be derived from the other.

So, semantically, the PDG inconsistency measure is equivalent to the full scoring function. Despite returning only a single number, the PDG inconsistency  $\langle\!\langle \cdot \rangle\!\rangle_\gamma$  encodes so much information that the optimal (set of) distribution(s) of a PDG [Section 3.3](#) can be derived from it.

When a single number encodes so much information, it is often because that information has been heavily compressed, and the drawback is that a lot of calculation is necessary to recover the original information. But, perhaps surprisingly, in this case it appears that the compressed form is not only more simpler and more compact, but also easier to use. To calculate  $\langle\!\langle \cdot \rangle\!\rangle_\gamma$  from  $\llbracket \cdot \rrbracket_\gamma$  requires solving an optimization problem over an enormous space, while the calculating  $\llbracket \cdot \rrbracket_\gamma$  from  $\langle\!\langle \cdot \rangle\!\rangle_\gamma$  requires only passing the function argument to the PDG and a single evaluation [\(9.1\)](#). Yet recovering the scoring function from inconsistency is only the beginning. Once we define the computational problems at hand precisely (which we do in [Section 9.2](#)), it will turn out that even PDG inference can be done in linear time with an inconsistency oracle [Section 9.3](#). All of this suggests at a theoretical level something we illustrated at length in

[Chapter 6](#): that calculating the degree of inconsistency is a fundamental problem, that can be used to solve other problems.

Indeed, as previously mentioned (e.g., in [Section 6.6](#)), a factor graph (a special case of a PDG, by [Theorem 3.6](#)) is associated with a similarly important number, called its *free energy* or *log partition function* (which is the inconsistency of this factor graph when viewed as a PDG, by [Proposition 6.14](#)). Calculating this number is computationally difficult, but if it is known, then inference becomes far easier ([Ma et al. 2013](#); [Koller and Friedman 2009](#)). This chapter will clarify the nature of this relationship, generalize it to a far wider class of probabilistic models, and significantly strengthen the reduction. More precisely, we will show that an inconsistency oracle on its own can be used to perform inference in time linear in the length of its required output length.

This will make precise a rather profound fact: not only is it difficult to see one's inconsistencies, but in fact precisely quantifying them is no easier than optimally resolving them.

## 9.2 The Computational Complexity of (Approximate) Inference and Inconsistency Calculation

Most work in graphical models works with abstract representations of real numbers, assuming that arithmetic (e.g., addition and multiplication of numbers) can be done in constant time. This assumption has several enormous advantages.

1. It simplifies things a great deal, and puts the focus where it should be: on the important aspects of the algorithm, rather than the details of numerical

representations.

2. It is essentially the right model of computational complexity in practice, given the we typically cache out numerical computations with floating point representations in hardware, which approximately have this complexity.
3. Finally, the assumption typically does not impact theoretical results, because (with a great deal of difficult and uninteresting work) one can often assume that all parameters are rational numbers, and then show that the calculations can be done just as before with a small overhead.

Unfortunately, and for a very superficial reason, the final point does not apply to PDGs. While [Theorem 8.10](#) gives us a way of doing inference to machine precision in polynomial time, which is the typical use case of an exact algorithm, it is not technically an exact inference algorithm. Indeed, if we use binary representations of numbers, exact inference for PDGs is technically not possible in finite time: in a PDG, the exact answer to an inference query may be an irrational number (even if all components of  $\mathbb{P}$ ,  $\alpha$ , and  $\beta$  are rational). Therefore, in order to say something precise about the computational complexity of inference, the problem we formulate precisely must be that of *approximate* inference.

**Definition 9.1** (approximate PDG inference). An instance of problem APPROX-PDG-INFER is a tuple  $(\mathcal{M}, \gamma, Q, \epsilon)$ , where  $\mathcal{M}$  is a PDG with variables  $\mathcal{X}$ ,  $\gamma \in \{0^+\} \cup [0, \infty]$  is the relative importance of structural information,  $Q$  is a conditional probability query of the form “ $\Pr(Y=y|X=x) = ?$ ”, where  $X, Y \subseteq \mathcal{X}$  and  $(x, y) \in \mathcal{V}(X, Y)$ , and  $\epsilon > 0$  is the precision desired for the answer. A solution to

this problem instance is a pair of numbers  $(r^-, r^+)$  such that

$$r^- \leq \inf_{\mu \in [\mathbf{m}]_\gamma^*} \mu(Y=y|X=x) \leq r^- + \epsilon$$

and       $r^+ \geq \sup_{\mu \in [\mathbf{m}]_\gamma^*} \mu(Y=y|X=x) \geq r^+ - \epsilon.$        $\square$

The problem we solved in [Sections 8.3](#) and [8.4](#) is the special case in which  $\mathbf{m}$  is assumed to be proper and  $\gamma \in \{0^+\} \cup (0, \min_a \frac{\beta_a}{\alpha_a})$ . This is enough to ensure there is a unique optimal distribution  $\mu^* \in [\mathbf{m}]_\gamma^*$ , with respect to which we must answer all queries. In this case, the definition above essentially amounts to providing a single number  $p$  such that  $p - \epsilon \leq \mu^*(Y=y|X=x) \leq p + \epsilon$ . We call this easier subproblem APPROX-INFER-CVX. We will also be interested in the unconditional variants of both inference problems, in which no additional evidence is supplied (i.e.,  $X = \emptyset$ ). We now define the analogous problem of approximately calculating a PDG's degree of inconsistency.

**Definition 9.2** (approximate inconsistency calculation). An instance of problem APPROX-CALC-INC is a triple  $(\mathbf{m}, \gamma, \epsilon)$ , where  $\mathbf{m}$  is a PDG,  $\gamma \geq 0$ , and  $\epsilon > 0$  is the desired precision. A solution to this problem instance is a number  $r$  such that  $|\langle\langle \mathbf{m} \rangle\rangle_\gamma - r| < \epsilon$ .       $\square$

The interior point method behind [Theorem 8.10](#) solves APPROX-CALC-INC in the process of finding a tree marginal for inference. But, technically, it does not solve APPROX-PDG-INFER. A solution to APPROX-PDG-INFER is a conditional probability, not a calibrated tree marginal. While a calibrated tree marginal does allow us to compute conditional probabilities, an  $\epsilon$ -close tree marginal does not give us  $\epsilon$ -close answers to probabilistic queries, especially those conditioned on improbable events (i.e., finding  $\Pr(Y=y|X=x)$  when  $\Pr(X=x) \approx 0$ ).

Nevertheless, because precision is so cheap, the interior point method behind [Theorem 8.10](#) can still be used as a subroutine to solve APPROX-INFER-CVX.

**Theorem 9.1.** APPROX-INFER-CVX can be solved in

link to  
proof

$$\tilde{O}\left((N+A)^4 V^{4(T+1)} \log \frac{1}{\epsilon \mu^*(x)} \left[ \log \frac{\beta^{\max}}{\beta^{\min}} + \log \frac{1}{\epsilon \mu^*(x)} \right] \right)$$

time,<sup>1</sup> where  $\mu^*(x)$  is the probability of the event  $X=x$  in the optimal distribution  $\mu^*$ ,  $\beta^{\max} := \max_{a \in \mathcal{A}} \beta_a$  is the largest observational confidence,

$$\text{and } \beta^{\min} := \begin{cases} \min_{a \in \mathcal{A}} \{\beta_a : \beta_a > 0\} & \text{if } \gamma = 0^+ \\ \gamma & \text{if } \gamma > 0 \end{cases} .$$

The factor of  $\log(1/\mu^*(x))$  is unusual, but even exact inference algorithms typically must write down  $\mu^*(X=x)$  on the way to calculating  $\mu^*(Y=x|X=x)$ , which implicitly incurs a cost of at least  $\log(1/\mu^*(x))$ . A Bayesian network with  $N$  variables in which cpds are articulated to precision  $k$  can have nonzero marginal probabilities as small as  $2^{-Nk}$ , in which case the additional worst case overhead for small probabilities is linear. We conjecture that it is not possible to form smaller marginal probabilities with PDGs, although the question remains open. Algorithmically speaking, [Theorem 9.1](#) extends [Theorem 8.10](#) in three key ways.

1. We must request additional precision to ensure that the marginal probabilities deviate at most  $\epsilon$  from the true ones. This sense of approximation effectively bounds the  $\ell_1$  norm of  $\mu^* - \mu$ , while [Theorem 8.10](#) (a) bounds its  $\ell_2$  norm and (b) bounds its  $\ell_\infty$  norm.

---

<sup>1</sup>At the cost of substantial overhead and engineering effort, the exponent 4 can be reduced to 2.872, by appeal to [Skajaa and Ye \(2015\)](#) and the current best matrix multiplication algorithm ([Duan et al. 2022](#),  $O(n^{2.372})$ ) to invert  $n \times n$  linear systems.

2. We must introduce a loop to refine precision until we have a suitably precise estimate of  $\Pr(X=x)$ .
3. Rather than directly dividing our estimate of  $\Pr(Y=y, X=x)$  by our estimate of  $\Pr(X=x)$ , we calculate something slightly more stable.

See the proof for details. One immediate corollary of [Theorem 9.1](#) is that  $\text{APPROX-INFER-CVX} \subseteq \text{EXP}$ , without the assumption of bounded treewidth.

It may be worth noting that there is at least one instance in the literature where *approximate* Bayesian Network (BN) inference is tractable for a subclass of models other than those of bounded treewidth: [Dagum and Luby \(1997\)](#) give a randomized algorithm for the special class of Bayesian Networks that do not have extreme conditional probabilities. Specifically they show that, assuming a network with  $N$  nodes, the inference problem is in  $\text{RP}(N, 1/\epsilon)$ . In addition to restricting to a different class of models (bounded conditional probabilities, but not bounded treewidth), their approximation algorithm is another significant respect: it is polynomial in  $1/\epsilon$ , rather than in  $\log(1/\epsilon)$ . Thus, the time it requires is exponential with respect to the number of requested digits, while our algorithm takes linear time.

Because PDGs generalize BNs, approximate inference for PDGs is at least as hard as it is for BNs.

**Proposition 9.2** ([Roth 1996](#)). *APPROX-PDG-INFER is #P-hard.*

Thus, the exponential time of [Theorem 9.1](#) is the best we could have hoped for, in the general case. The argument is due to [Roth \(1996\)](#), although we have altered it somewhat.

### 9.3 A Deeper, Computational Connection

Our approach to  $\gamma$ -inference computes  $\langle\!\langle m \rangle\!\rangle_\gamma$  as a side effect. But suppose that we were interested in calculating only this inconsistency. Might there be a more direct, asymptotically easier way to do so? In general, the answer is no.

**Theorem 9.3.** (a) *Determining whether there is a distribution that satisfies all cpds of a PDG is NP-hard.*

[ link to proof ]

(b) *Calculating a PDG's degree of inconsistency (exactly) is #P hard.*

(c) *APPROX-CALC-INC is #P hard, even for fixed  $\gamma \geq 0$  and  $\epsilon > 0$ .*

Historically, our original approach to inferring the probability of  $Y$  in a PDG  $m$  was to minimize their combined inconsistency (Richardson and Halpern 2021). The idea is to add a hypothesis distribution  $h(Y)$  to  $m$ , and adjust  $h$  to minimize the overall inconsistency  $\langle\!\langle m + h \rangle\!\rangle_\gamma$ . Parts (b) and (c) of Theorem 9.3 significantly undermine this approach, because even just calculating  $\langle\!\langle m + h \rangle\!\rangle_\gamma$  is intractable. Typically minimizing a function is more difficult than evaluating it, so one might imagine the intractability of  $\langle\!\langle m + h \rangle\!\rangle_\gamma$  to be merely the first of many difficulties—yet it turns out to be the only one. There is a strong sense in which being able to calculate inconsistency is enough to perform inference efficiently. Specifically, with oracle access to the inconsistency  $\langle\!\langle m + h \rangle\!\rangle_\gamma$ , our original approach gives right answer with the best possible asymptotic time complexity. Thus, while it may not be a practical inference algorithm, it is a powerful reduction from inference to inconsistency calculation.

**Theorem 9.4.** (a) *There is an  $O(\log^{1/\epsilon})$ -time reduction from unconditional APPROX-INFER-CVX to the problem of determining which of two PDGs is more inconsistent,*

[ link to proof ]

using  $O(\log^{1/\epsilon})$  subroutine calls.

- (b) There is an  $O\left(\log \frac{\langle\!\langle m \rangle\!\rangle_\gamma}{\gamma \epsilon \mu^*(x)} \cdot \log \frac{1}{\epsilon \mu^*(x)}\right)$  time reduction from **APPROX-INFERENCE-CVX** to **APPROX-CALC-INC** using  $O(\log(1/\epsilon) \log \log^{1/\mu^*(x)})$  calls to the inconsistency subroutine.
- (c) There is also an  $O(|VC|)$  reduction from **APPROX-CALC-INC** to **APPROX-INFERENCE-CVX**. With the additional assumption of bounded treewidth, this is linear in the number of variables in the PDG.

Recall that the runtime of  $O(\log(1/\epsilon))$  achieved by part (a) is optimal, because it is the complexity of writing down an answer, which in general requires  $\log(1/\epsilon)$  bits. While it is a clean result, part (a) is unsatisfying as a complexity result because it relies heavily on being able to compare the two inconsistencies in constant time. Part (b) fleshes out the algorithm of part (a) more precisely by reducing to inconsistency approximation (which we now know is computable), and also extends the procedure to handle to conditional probability queries. This leads to a significantly more complex analysis, and a more expensive reduction, although it is possible that much of the difference in the costs is due to loose bounds in our analysis. Part (c) is a straightforward observation in light of the results in [Section 8.4](#).

To summarize: in the range of  $\gamma$ 's in which we have an (approximate) inference algorithm for PDGs, (approximately) calculating a PDG's degree of inconsistency is at least as difficult. For PDGs of bounded treewidth, the two problems are equivalent, and can be solved in polynomial time.

## 9.4 The Reductions

### 9.4.1 Hardness

We now turn to [Theorem 9.3](#). We begin by proving parts (a) and (b) directly by reduction to SAT and #SAT, respectively.

**Theorem 9.3.**

- (a) *Determining whether there is a distribution that satisfies all cpds of a PDG is NP-hard.*
- (b) *Calculating a PDG's degree of inconsistency (exactly) is #P hard.*
- (c) *APPROX-CALC-INC is #P hard, even for fixed  $\gamma \geq 0$  and  $\epsilon > 0$ .*

*Proof.* (a). We can directly encode SAT problems in PDGs. Choose any CNF formula

$$\varphi = \bigwedge_{j \in \mathcal{J}} \bigvee_{i \in \mathcal{I}(j)} (X_{j,i})$$

over binary variables  $\mathbf{X} := \bigcup_{j,i} X_{j,i}$ , and let  $n := |\mathbf{X}|$  denote the total number of variables in  $\varphi$ . Let  $\mathcal{M}_\varphi$  be the PDG containing every variable  $X \in \mathbf{X}$  and a binary variable  $C_j$  (taking the value 0 or 1) for each clause  $j \in \mathcal{J}$ , as well as the following edges, for each  $j \in \mathcal{J}$ :

- a hyperedge  $\{X_{j,i} : i \in \mathcal{I}(j)\} \rightarrowtail C_j$ , together with a degenerate cpd encoding the boolean OR function (i.e., the truth of  $C_j$  given  $\{X_{j,i}\}$ );
- an edge  $\mathbb{1} \rightarrowtail C_j$ , together with a cpd asserting  $C_j$  be equal to 1.

First, note that the number of nodes, edges, and non-zero entries in the cpds are polynomial in the  $|\mathcal{J}|$ ,  $|\mathbf{X}|$ , and the total number of parameters in a simple matrix representation of the cpds is also polynomial if  $\mathcal{I}$  is bounded (e.g., if  $\varphi$  is a 3-CNF formula). A satisfying assignment  $\mathbf{x} \models \varphi$  of the variables  $\mathbf{X}$  can be regarded as a degenerate joint distribution  $\delta_{\mathbf{x}=\mathbf{x}}$  on  $\mathbf{X}$ , and extends uniquely to a full joint distribution  $\mu_{\mathbf{x}} \in \Delta\mathcal{V}(\mathbf{m}_{\varphi})$  consistent with all of the edges, by

$$\mu_{\mathbf{x}} = \delta_{\mathbf{x}} \otimes \delta_{\{C_j = \vee_i x_{j,i}\}}$$

Conversely, if  $\mu$  is a joint distribution consistent with the edges above, then any point  $\mathbf{x}$  in the support of  $\mu(\mathbf{X})$  must be a satisfying assignment, since the two classes of edges respectively ensure that  $1 = \mu(C_j = 1 \mid \mathbf{X} = \mathbf{x}) = \bigvee_{i \in \mathcal{I}(j)} \mathbf{x}_{j,i}$  for all  $j \in \mathcal{J}$ , and so  $\mathbf{x} \models \varphi$ .

Thus,  $\{\mathbf{m}_{\varphi}\} \neq \emptyset$  if and only if  $\varphi$  is satisfiable, so an algorithm for determining if a PDG is consistent can also be adapted (in polynomial space and time) for use as a SAT solver, and so the problem of determining if a PDG consistent is NP-hard.

**(b) Hardness of exact computation.** We prove this by reduction to #SAT. Again, let  $\varphi$  be some CNF formula over  $\mathbf{X}$ , and construct  $\mathbf{m}_{\varphi}$  as in [the proof of Theorem 9.3](#). Furthermore, let  $[\![\varphi]\!] := \{\mathbf{x} : \mathbf{x} \models \varphi\}$  be the set of assignments to  $\mathbf{X}$  satisfying  $\varphi$ , and  $\#\varphi := |[\![\varphi]\!]|$  denote the number such assignments. We now claim that

$$\#\varphi = \exp \left[ -\frac{1}{\gamma} \langle\!\langle \mathbf{m}_{\varphi} \rangle\!\rangle_{\gamma} \right]. \quad (9.2)$$

Once we do so, we will have a reduced the #P-hard problem of computing  $\#\varphi$  to the problem of computing  $\langle\!\langle \mathbf{m} \rangle\!\rangle_{\gamma}$  (exactly).

We now prove (9.2). By definition, we have

$$\langle\!\langle \mathbf{m}_\varphi \rangle\!\rangle_\gamma = \inf_{\mu} \left[ OInc_{\mathbf{m}_\varphi}(\mu) + \gamma SInc_{\mathbf{m}_\varphi}(\mu) \right].$$

We start with a claim about first term.

**Claim 9.4.1.**  $OInc_{\mathbf{m}_\varphi}(\mu) = \begin{cases} 0 & \text{if } \text{Supp } \mu \subseteq [\![\varphi]\!] \times \{1\} \\ \infty & \text{otherwise.} \end{cases}$

*Proof.* Writing out the definition explicitly, the first can be written as

$$OInc_{\mathbf{m}_\varphi}(\mu) = \sum_j \left[ D\left(\mu(C_j) \parallel \delta_1\right) + \mathbb{E}_{\mathbf{x} \sim \mu(\mathbf{X}_j)} D\left(\mu(C_j \mid \mathbf{X}_j = \mathbf{x}) \parallel \delta_{\vee_i x_{j,i}}\right) \right], \quad (9.3)$$

where  $\mathbf{X}_j = \{X_{ij} : j \in \mathcal{I}(j)\}$  is the set of variables that appear in clause  $j$ , and  $\delta_{(-)}$  is the probability distribution placing all mass on the point indicated by its subscript. As a reminder, the relative entropy is given by

$$D\left(\mu(\Omega) \parallel \nu(\Omega)\right) := \mathbb{E}_{\omega \sim \mu} \log \frac{\mu(\omega)}{\nu(\omega)},$$

and in particular, if  $\Omega$  is binary,

$$D\left(\mu(\Omega) \parallel \delta_\omega\right) = \begin{cases} 0 & \text{if } \mu(\omega) = 1; \\ \infty & \text{otherwise.} \end{cases}$$

Applying this to (9.3), we find that either:

1. Every term of (9.3) is finite (and zero) so  $OInc_{\mathbf{m}_\varphi}(\mu) = 0$ , which happens when  $\mu(C_j = 1) = 1$  and  $\mu(C_j = \vee_i x_{j,i}) = 1$  for all  $j$ . In this case,  $\mathbf{c} = \mathbf{1} = \{\vee_i x_{j,i}\}_j$  so  $\mathbf{x} \models \varphi$  for every  $(\mathbf{c}, x) \in \text{Supp } \mu$ ;
2. Some term of (9.3) is infinite, so that  $OInc_{\mathbf{m}_\varphi}(\mu) = \infty$ , which happens if some  $j$ , either

(a)  $\mu(C_j \neq 1) > 0$  — in which case there is some  $(\mathbf{x}, c) \in \text{Supp } \mu$  with  $\mathbf{c} \neq 1$ ,

or

(b)  $\text{Supp } \mu(\mathbf{C}) = \{\mathbf{1}\}$ , but  $\mu(C_j \neq \vee_i x_{j,i}) > 0$  — in which case there is some  $(\mathbf{x}, 1) \in \text{Supp } \mu$  for which  $1 = c_j \neq \vee_i x_{j,i}$ , and so  $\mathbf{x} \not\models \varphi$ .

Condensing and rearranging slightly, we have shown that

$$OInc_{m_\varphi}(\mu) = \begin{cases} 0 & \text{if } \mathbf{x} \models \varphi \text{ and } \mathbf{c} = \mathbf{1} \text{ for all } (\mathbf{x}, \mathbf{c}) \in \text{Supp } \mu \\ \infty & \text{otherwise} \end{cases}.$$

□

Because  $SInc$  is bounded, it follows immediately that  $\langle\!\langle m_\varphi \rangle\!\rangle_\gamma$  is finite if and only if there is some distribution  $\mu \in \Delta\mathcal{V}(\mathbf{X}, \mathbf{C})$  for which  $OInc_{m_\varphi}(\mu)$  is finite, or equivalently, by [Claim 9.4.1](#), iff there exists some  $\mu(\mathbf{X}) \in \Delta\mathcal{V}(\mathbf{X})$  for which  $\text{Supp } \mu(\mathbf{X}) \subseteq \llbracket \varphi \rrbracket$ , which in turn is true if and only if  $\varphi$  is satisfiable.

In particular, if  $\varphi$  is not satisfiable (i.e.,  $\#\varphi = 0$ ), then  $\langle\!\langle m_\varphi \rangle\!\rangle_\gamma = +\infty$ , and

$$\exp\left[-\frac{1}{\gamma} \langle\!\langle m_\varphi \rangle\!\rangle_\gamma\right] = \exp[-\infty] = 0 = \#\varphi,$$

so in this case (9.2) holds as promised. On the other hand, if  $\varphi$  is satisfiable, then, again by [Claim 9.4.1](#), every  $\mu$  minimizing  $\llbracket m_\varphi \rrbracket_\gamma$ , (i.e., every  $\mu \in \llbracket m_\varphi \rrbracket_\gamma^*$ ) must be supported entirely on  $\llbracket \varphi \rrbracket$  and have  $OInc_{m_\varphi}(\mu) = 0$ . As a result, we have

$$\langle\!\langle m_\varphi \rangle\!\rangle_\gamma = \inf_{\mu \in \Delta[\llbracket \varphi \rrbracket \times \{\mathbf{1}\}]} \gamma SInc_{m_\varphi}(\mu).$$

A priori, by the definition of  $SInc_{m_\varphi}$ , we have

$$SInc_{m_\varphi}(\mu) = -H(\mu) + \sum_j \left[ \alpha_{j,1} H_\mu(C_j \mid \mathbf{X}_j) + \alpha_{j,0} H_\mu(C_j) \right],$$

where  $\alpha_{j,0}$  and  $\alpha_{j,1}$  are values of  $\alpha$  for the edges of  $m_\varphi$ , which we have not specified because they are rendered irrelevant by the fact that their corresponding cpds are deterministic. We now show how this plays out in the present case. Any  $\mu \in \Delta[\llbracket \varphi \rrbracket \times \{\mathbf{1}\}]$  we consider has a degenerate marginal on  $C$ . Specifically, for every  $j$ , we have  $\mu(C_j) = \delta_1$ , and since entropy is non-negative and never increased by conditioning,

$$0 \leq H_\mu(C_j \mid \mathbf{X}_j) \leq H_\mu(C_j) = 0.$$

Therefore,  $SInc_{m_\varphi}(\mu)$  reduces to the negative entropy of  $\mu$ . Finally, making use of the fact that the maximum entropy distribution  $\mu^*$  supported on a finite set  $S$  is the uniform distribution on  $S$ , and has  $H(\mu^*) = \log |S|$ , we have

$$\begin{aligned} \langle\!\langle m_\varphi \rangle\!\rangle_\gamma &= \inf_{\mu \in \Delta([\varphi] \times \{\mathbf{1}\})} \gamma SInc_{m_\varphi}(\mu) \\ &= \inf_{\mu \in \Delta([\varphi] \times \{\mathbf{1}\})} -\gamma H(\mu) \\ &= -\gamma \sup_{\mu \in \Delta([\varphi] \times \{\mathbf{1}\})} H(\mu) \\ &= -\gamma \log(\#_\varphi), \end{aligned}$$

giving us

$$\#_\varphi = \exp \left[ -\frac{1}{\gamma} \langle\!\langle m_\varphi \rangle\!\rangle_\gamma \right],$$

as desired. We have now reduced #SAT to computing  $\langle\!\langle m \rangle\!\rangle_\gamma$ , for  $\gamma > 0$  and an arbitrary PDG  $m$ , which is therefore #P-hard.

To show the same for  $\gamma = 0$ , it suffices to add an additional hyperedge pointing to all variables, and associate it with a joint uniform distribution, and confidence 1, resulting in a new PDG  $m'_\varphi$ . Because this new edge's contribution to  $OInc_m$  equals  $D(\mu \parallel \text{Unif}(\mathcal{X})) = \log |\mathcal{V}\mathcal{X}| - H(\mu)$ , we have

$$\llbracket m'_\varphi \rrbracket_0(\mu) = OInc_{m'_\varphi}(\mu) = \llbracket m_\varphi \rrbracket(\mu) + \log |\mathcal{V}\mathcal{X}| - H(\mu) = \llbracket m_\varphi \rrbracket_1(\mu) + \log |\mathcal{V}\mathcal{X}|.$$

Since this is true for all  $\mu$ , we can take the of this equation over  $\mu$ , and so conclude that

$$\begin{aligned}\langle\!\langle \mathbf{m}'_\varphi \rangle\!\rangle_0 &= \langle\!\langle \mathbf{m}_\varphi \rangle\!\rangle_1 + \log |\mathcal{V}\mathcal{X}| = \log(|\mathcal{V}\mathcal{X}|/\#\varphi) \\ \implies \#\varphi &= |\mathcal{V}\mathcal{X}| \exp(-\langle\!\langle \mathbf{m}'_\varphi \rangle\!\rangle_0)\end{aligned}$$

Thus, the number of satisfying assignments can be found through via an oracle for  $\langle\!\langle - \rangle\!\rangle_0$ , as well. This shows that calculating this purely observational inconsistency is #P-hard as well.

**(c) Hardness of approximation.** To calculate  $\#\varphi$  exactly, it turns out that we do not need to know  $\langle\!\langle \mathbf{m}_\varphi \rangle\!\rangle_\gamma$  exactly. Instead, we claim it suffices to approximate it to within  $\epsilon < \gamma \log(1 + 2^{-(n+1)})$ .

Suppose that  $|r - \langle\!\langle \mathbf{m}_\varphi \rangle\!\rangle_\gamma| < \epsilon$ . Then

$$\begin{aligned}\exp\left(-\frac{r}{\gamma}\right) &\in \exp\left[-\frac{1}{\gamma}(\langle\!\langle \mathbf{m}_\varphi \rangle\!\rangle_\gamma \pm \epsilon)\right] \\ &= \exp\left[-\frac{1}{\gamma}\langle\!\langle \mathbf{m}_\varphi \rangle\!\rangle_\gamma\right] \cdot \exp(\pm\epsilon/\gamma) \\ &= \#\varphi \cdot \exp(\pm\epsilon/\gamma) \\ &= [\#\varphi \exp(-\epsilon/\gamma), \#\varphi \exp(+\epsilon/\gamma)].\end{aligned}$$

Since  $\#\varphi$  is a natural number and at most  $2^n$ , If we can get a relative approximation of it to within a factor of  $2^{-(n+1)}$ , then rounding that approximate value to the nearest whole number gives the exact value of  $\#\varphi$ . Thus, it suffices to choose  $\epsilon$  small enough that

$$\exp(-\epsilon/\gamma) > 1 - 2^{-(n+1)} \quad \text{and} \quad \exp(+\epsilon/\gamma) < 1 + 2^{-(n+1)};$$

this is satisfied any choice of  $\epsilon < \gamma \log(1 + 2^{-(n+1)})$ . Thus, being able to approximate  $\langle\!\langle \mathbf{m}_\varphi \rangle\!\rangle_\gamma$  sufficiently closely will tell us whether or not  $\varphi \in \text{SAT}$ . Note that

for large  $n$ , the maximum value of  $\epsilon$  for which this is true is on the order of  $\epsilon_{\max} \in \Theta(\gamma 2^{-n})$ . It follows that  $\log(1/\epsilon_{\max}) \in O(n)$ , and so values of  $\epsilon$  small enough to determine the satisfiability of a formula  $\varphi$  with  $n$  variables can be specified in time  $O(n)$ . Thus, the problem APPROX-CALC-INC is #P hard (in the size of its input).  $\square$

#### 9.4.2 Inference via Inconsistency Minimization.

We now address [Theorem 9.4](#), which is closely related to our original idea for an inference algorithm, via inconsistency minimization. While that idea does not yield an efficient inference algorithm, it does yield an efficient reduction from inconsistency minimization to inference. In order to prove this, we first need another construction with PDGs. A probability over a (set of) variables can be viewed as a vector whose elements sum to one. It turns out that it is possible to use the machinery of PDGs to, effectively, give only one value of such a probability vector. That is, for any  $p \in [0, 1]$ , we can construct a PDG that represents the belief that  $\Pr(Y=y) = p$ , but say nothing about how the probability splits between other values of  $y$ . A generalization of this construction was presented in [Section 4.2.4](#); for the reader's convenience, we now [repeat](#) the special case of it that will matter for the reduction.

explicitly give

We first introduce an auxiliary binary variable  $Y_y$ , with  $\mathcal{V}(Y_y) = \{y, \neg y\}$ , and takes the value  $y$  if  $Y = y$ , and  $\neg y$  if  $Y \neq y$ . Note that this variable is a function of the value of variable  $Y$  (although we will need to enforce this with an additional arc), and therefore there is a unique way to extend a distribution over variables including  $Y$  to also include the variable  $Y_y$ .

With this definition, there is now an obvious way to add a hyperarc with no source and target  $Y_y$ , together with a asserting that  $\Pr(Y=y) = p$ . This cpd is written as a vector  $\hat{p}$  on the right of the figure below. The PDG we have just constructed is illustrated on the left of the figure below. In addition to  $\hat{p}$  and the new variable, this PDG includes the structural constraint  $s$  needed to define the variables  $Y_y$  in terms of  $Y$ ; it is a deterministic function, drawn with a double-headed gray arrow.

$$\begin{array}{ccc}
 \xrightarrow{\hat{p}} & \boxed{Y_y} & s(Y_y|Y) := \begin{cases} y & \text{if } Y = y \\ \neg y & \text{if } Y \neq y \end{cases} \\
 \uparrow\!\!\! \nearrow s & & \\
 \boxed{Y} & \hat{p}(Y_y) := \begin{bmatrix} y & \neg y \\ p & 1-p \end{bmatrix}
 \end{array}$$

So, when we add  $\Pr(Y = y) = p$  to a PDG  $m$ , what we really mean is: first convert construct a widget as above, and add that structure (i.e., the new variable  $Y_y$ , its definition  $s$ , and the cpd  $\hat{p}$ ) to  $m$ . In what sense does this “work”? The first order of business is to prove that it behaves as we should expect, semantically, in the case we’re interested in.

**Lemma 9.5.** *Suppose that  $m$  is a PDG with variables  $\mathcal{X}$  and  $\beta \geq 0$ . Then, for all  $Y \subseteq \mathcal{X}$ ,  $y \in \mathcal{V}Y$ ,  $p \in [0, 1]$  and  $\gamma \geq 0$ , we have that:*

$$\langle\!\langle m + \Pr(Y=y) = p \rangle\!\rangle_\gamma \geq \langle\!\langle m \rangle\!\rangle_\gamma,$$

*with equality if and only if there exists  $\mu \in \llbracket m \rrbracket_\gamma^*$  such that  $\mu(Y=y) = p$ .*

*Proof.* The inequality is immediate; it is an instance of monotonicity of inconsistency (Lemma 6.1). We now prove that equality holds iff there is a minimizer with the appropriate conditional probability.

( $\Leftarrow$ ). Suppose that there is some  $\mu \in \llbracket m \rrbracket_{\gamma}^*$  with  $\mu(Y=y) = p$ . Because  $\mu \in \llbracket m \rrbracket_{\gamma}^*$ , we know that  $\llbracket m \rrbracket_{\gamma}(\mu) = \langle\!\langle m \rangle\!\rangle$ . Let  $\hat{\mu}$  be the extension of  $\mu$  to the new variable “ $Y_y$ ”, whose value is a function of  $Y$  according to  $s$ . Then

$$\begin{aligned}\langle\!\langle m + \Pr(Y=y) = p \rangle\!\rangle_{\gamma} &\leq \llbracket m + \Pr(Y=y) = p \rrbracket_{\gamma}(\hat{\mu}) \\ &= \llbracket m \rrbracket_{\gamma}(\mu) + \mathbb{E}_{\mu} \left[ \log \frac{\hat{\mu}(Y_y)}{\hat{p}(Y_y)} \right] \\ &= \llbracket m \rrbracket_{\gamma}(\mu) + \mu(Y=y) \log \frac{\mu(Y=y)}{p} + \mu(Y \neq y) \log \frac{\mu(Y \neq y)}{1-p} \\ &= \llbracket m \rrbracket_{\gamma}(\mu) + \mu(Y=y) \log(1) + \mu(Y \neq y) \log(1) \\ &= \llbracket m \rrbracket_{\gamma}(\mu) = \langle\!\langle m \rangle\!\rangle_{\gamma}.\end{aligned}$$

To complete this direction of the proof, it suffices to observe that we already knew the inequality held in the opposite direction (by monotonicity), so the two terms are equal.

( $\Rightarrow$ ). Suppose that the two inconsistencies are equal, i.e.,  $\langle\!\langle m + \Pr(Y=y) = p \rangle\!\rangle_{\gamma} = \langle\!\langle m \rangle\!\rangle_{\gamma}$ . This time, choose  $\hat{\mu} \in \llbracket m + \Pr(Y=y) = p \rrbracket_{\gamma}^*$ , and define  $\mu$  to be its marginal on the variables of  $m$  (which contains the same information as  $\hat{\mu}$  itself). Let  $q := \mu(Y=y)$ . Then

$$\begin{aligned}\langle\!\langle m \rangle\!\rangle_{\gamma} &= \langle\!\langle m + \Pr(Y=y) = p \rangle\!\rangle_{\gamma} \\ &= \llbracket m + \Pr(Y=y) = p \rrbracket_{\gamma}(\hat{\mu}) \\ &= \llbracket m \rrbracket_{\gamma}(\mu) + \mu(Y=y) \log \frac{\mu(Y=y)}{p} + \mu(Y \neq y) \log \frac{\mu(Y \neq y)}{1-p} \\ &= \llbracket m \rrbracket_{\gamma}(\mu) + \left[ q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p} \right] \\ &= \llbracket m \rrbracket_{\gamma}(\mu) + D(q \parallel p) \\ &\geq \langle\!\langle m \rangle\!\rangle_{\gamma} + D(q \parallel p)\end{aligned}$$

Therefore  $0 \geq D(q \parallel p)$ . But relative entropy is non-negative (Gibbs inequality);

see any introductory text on information theory, such as MacKay (2003)), so we actually know that  $D(q \parallel p) = 0$ , and thus  $p = \mu(Y=y)$ . In addition, the algebra above shows that  $\mu \in [\![\mathbf{m}]\!]_\gamma^*$ , as its score is  $\langle\!\langle \mathbf{m} \rangle\!\rangle_\gamma$ . Thus, we have found  $\mu \in [\![\mathbf{m}]\!]_\gamma^*$  such that  $\mu(Y=y) = p$ , completing the proof.  $\square$

We next show that the overall inconsistency is strictly convex in the parameter  $p \in [0, 1]$ . The result is simpler to state (and equally easy to prove) this result in the general case.

**Lemma 9.6.** Fix  $Y \subseteq \mathcal{X}$ , and  $\gamma \in (0, \min_a \frac{\beta_a}{\alpha_a})$ . As  $h = h(Y)$  ranges over  $\Delta \mathcal{V}Y$ , the function  $h \mapsto \langle\!\langle \mathbf{m} + h \rangle\!\rangle_\gamma$  is strictly convex. [ link to proof ]

*Proof.* We start by expanding the definitions. If  $h$  is a cpd on  $Y$  given  $X$ , then

$$\begin{aligned}\langle\!\langle \mathbf{m} + h \rangle\!\rangle_\gamma &= \inf_{\mu} [\![\mathbf{m} + h]\!]_\gamma(\mu) \\ &= \inf_{\mu} \left[ [\![\mathbf{m}]\!]_\gamma(\mu) + D\left(\mu(Y) \parallel h(Y)\right) \right].\end{aligned}$$

Fix  $\gamma \leq \min_a \frac{\beta_a}{\alpha_a}$ . Then we know that  $[\![\mathbf{m}]\!]_\gamma(\mu)$  is a  $\gamma$ -strongly convex (so, in particular, strictly convex) function of  $\mu$ , and hence there is a unique joint distribution which minimizes it. We now show that the overall inconsistency is strictly convex in  $h$ .

Suppose that  $h_1(Y)$  and  $h_2(Y)$  are two distributions over  $Y$ . Let  $\mu_1, \mu_2$  and  $\mu_\lambda$  be the joint distributions that minimize  $[\![\mathbf{m} + h_1]\!]_\gamma$  and  $[\![\mathbf{m} + h_2]\!]_\gamma$ , respectively. For every  $\lambda \in [0, 1]$ , define  $h_\lambda := (1 - \lambda)h_1 + \lambda h_2$ ,  $\mu_\lambda := (1 - \lambda)\mu_1 + \lambda\mu_2$ , and  $\mu_\lambda^*$  to be a minimizer of  $[\![\mathbf{m} + h_\lambda]\!]_\gamma$ . The following is a simple consequence of these

definitions:

$$\begin{aligned}
\langle\langle \mathbf{m} + h_\lambda \rangle\rangle_\gamma &= [\![\mathbf{m} + h_\lambda]\!]_\gamma(\mu_\lambda^*) \\
&\leq [\![\mathbf{m} + h_\lambda]\!]_\gamma(\mu_\lambda) \\
&= [\![\mathbf{m}]\!]_\gamma(\mu_\lambda) + \mathbf{D}\left(\mu_\lambda(Y) \parallel h_\lambda(Y)\right).
\end{aligned}$$

By the convexity of  $[\![\mathbf{m}]\!]_\gamma$  and  $\mathbf{D}$ , we have

$$[\![\mathbf{m}]\!]_\gamma(\mu_\lambda) \leq (1 - \lambda)[![\mathbf{m}]\!]_\gamma(\mu_1) + \lambda[\![\mathbf{m}]\!]_\gamma(\mu_2) \quad (9.4)$$

$$\begin{aligned}
\text{and } \mathbf{D}\left(\mu_\lambda(Y) \parallel h_\lambda(Y)\right) &\leq (1 - \lambda)\mathbf{D}\left(\mu_1(Y) \parallel h_1(Y)\right) \\
&\quad + \lambda \mathbf{D}\left(\mu_2(Y) \parallel h_2(Y)\right). \quad (9.5)
\end{aligned}$$

If  $\mu_1 \neq \mu_2$  then since  $[\![\mathbf{m}]\!]$  is strictly convex, (9.4) must be a strict inequality. On the other hand, if  $\mu_1 = \mu_2$ , then since  $\mu_\lambda = \mu_1 = \mu_2$  and  $\mathbf{D}$  is strictly convex in its second argument when its first argument is fixed, (9.5) must be a strict inequality. In either case, the sum of the two inequalities must be strict. Combining this with the first inequality, we get

$$\begin{aligned}
\langle\langle \mathbf{m} + h_\lambda \rangle\rangle_\gamma &\leq [\![\mathbf{m}]\!]_\gamma(\mu_\lambda) + \mathbf{D}\left(\mu_\lambda(Y) \parallel h_\lambda(Y)\right) \\
&< (\lambda - 1) \left[ [\![\mathbf{m}]\!]_\gamma(\mu_1) + \mathbf{D}\left(\mu_1(Y) \parallel h_1(Y)\right) \right] \\
&\quad + \lambda \left[ [\![\mathbf{m}]\!]_\gamma(\mu_2) + \mathbf{D}\left(\mu_2(Y) \parallel h_2(Y)\right) \right] \\
&= (\lambda - 1)\langle\langle \mathbf{m} + h_1 \rangle\rangle + \lambda \langle\langle \mathbf{m} + h_2 \rangle\rangle,
\end{aligned}$$

which shows that  $\langle\langle \mathbf{m} + h \rangle\rangle$  is *strictly* convex in  $h$ , as desired.  $\square$

Let  $\mathbf{m}$  be a PDG with  $\beta \geq \mathbf{0}$  and variables  $\mathcal{X}$ , and fix  $Y \subseteq \mathcal{X}$ ,  $y \in \mathcal{V}Y$ , and  $\gamma \in (0, \min_a \frac{\beta_a}{\alpha_a})$ . For  $p \in [0, 1]$ , define

$$f(p) := \langle\langle \mathbf{m} + \Pr(Y=y) \rangle\rangle_\gamma. \quad (9.6)$$

The next several results (Corollary 9.6.1 and Lemmas 9.7, 9.8 and 9.10 to 9.12) are properties of this function  $f(p)$ .

**Corollary 9.6.1.** *The function  $f$  defined in (9.6) is strictly convex.*

*Proof.* Simply take  $h$  to be the cpd  $\hat{p}$ , absorb the other components of (the PDG representation of)  $\Pr(Y=y) = p$  into  $\mathbf{m}$ , and then apply Lemma 9.6.  $\square$

The results from this point until the proof of Theorem 14 are all technical results that support the more precise analysis of part (b). We recommend returning to these results as needed, after first reading the proof of part (a).

**Lemma 9.7.** *For  $p \in (0, 1]$ , let  $\mu_p^*$  be the unique element of  $[\![\mathbf{m} + \Pr(Y=y) = p]\!]_\gamma^*$ . Then  $f'(p) = \frac{p - \mu_p^*(Y=y)}{p(1-p)}$ .*

*Proof.* First, suppose  $\mu_p^*$  is in the interior of the simplex. Since it minimizes the differentiable function

$$[\![\mathbf{m} + \Pr(Y=y) = p]\!]_\gamma = \mu \mapsto [\![\mathbf{m}]\!]_\gamma + D(\mu(Y=y) \parallel p),$$

the gradient of that function at  $\mu_p^*$  must be zero:

$$\nabla [\![\mathbf{m} + \Pr(Y=y) = p]\!]_\gamma(\mu_p^*) = \nabla_\mu \left[ D(\mu(Y=y) \parallel p) \right]_{\mu=\mu_p^*} + \nabla [\![\mathbf{m}]\!](\mu_p^*) = 0. \quad (9.7)$$

What is the derivative of  $f$ ? Observe that  $f$  is the sum of two compositions of differentiable maps:

$$\begin{aligned} f_m := & \quad p \mapsto \mu_p^* \mapsto [\![\mathbf{m}]\!]_\gamma(\mu_p^*) \\ \text{and} \quad f_{\Pr} := & \quad p \mapsto (p, \mu_p^*) \mapsto D(\mu_p^*(Y=y) \parallel p). \end{aligned}$$

Thus, we can use the multivariate chain rule. for any differentiable functions  $h : \mathbb{R}^m \rightarrow \mathbb{R}^k$ , and  $g : \mathbb{R}^k \rightarrow \mathbb{R}^n$ , their composition  $g \circ h : \mathbb{R}^m \rightarrow \mathbb{R}^n$  is also a differentiable map whose Jacobian is  $\mathbf{J}_{g \circ h}(x) = \mathbf{J}_g(h(x))\mathbf{J}_h(x)$ . In our case,  $g$  will be a scalar map ( $n = 1$ ), so  $\mathbf{J}_g(h(x)) = [\dots, \frac{\partial g}{\partial x_j}, \dots](h(x)) = \nabla g(h(x))^\top$ . Let  $\mathbf{J}_{\mu_p^*}(p)$  be the Jacobian of the map  $p \mapsto \mu_p^*$ . Then

$$\begin{aligned} f'(p) &= f'_m(p) + f'_{\text{Pr}}(p) \\ &= (\nabla \llbracket \mathbf{m} \rrbracket(\mu_p^*))^\top \mathbf{J}_{\mu_p^*} + \left( \nabla_\mu \left[ \mu(x) \mathbf{D}(\mu(y|x) \parallel p) \right]_{\mu=\mu_p^*} \right)^\top \mathbf{J}_{\mu_p^*} + \frac{\partial}{\partial p} \left[ \mathbf{D}(\mu_p^*(Y=y) \parallel p) \right] \\ &\quad (\text{Alternatively, the line above can be derived from the law of total derivative.}^2) \\ &= \underbrace{\left( \nabla_\mu \left[ \mathbf{D}(\mu(Y=y) \parallel p) \right] \right)_{\mu=\mu_p^*}}_{= \frac{\partial}{\partial p} \mathbf{D}(\mu_p^*(Y=y) \parallel p)} + \underbrace{\nabla \llbracket \mathbf{m} \rrbracket(\mu_p^*)^\top}_{\text{by (9.7)}} \mathbf{J}_{\mu_p^*}(p) + \frac{\partial}{\partial p} \mathbf{D}(\mu_p^*(Y=y) \parallel p) \\ &= \frac{\partial}{\partial p} \mathbf{D}(\mu_p^*(Y=y) \parallel p) \end{aligned}$$

Finally,

$$\begin{aligned} \frac{d}{dp} \left[ \mathbf{D}(q \parallel p) \right] &= \frac{d}{dp} \left[ q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p} \right] \\ &= q \left( \frac{p}{q} \right) \frac{d}{dp} \left[ \frac{q}{p} \right] + (1-q) \left( \frac{1-p}{1-q} \right) \frac{d}{dp} \left[ \frac{1-q}{1-p} \right] \\ &= pq \left( \frac{-1}{p^2} \right) + (1-p)(1-q) \left( \frac{-1}{(1-p)^2} \right) (-1) \\ &= -\frac{q}{p} + \frac{1-q}{1-p} \\ &= \frac{-(1-p)q + p(1-q)}{p(1-p)} \\ &= \frac{pq - q + p - pq}{p(1-p)} = \frac{p - q}{p(1-p)}. \end{aligned}$$

Thus, we find

$$f'(p) = \frac{p - \mu_p^*(Y=y)}{p(1-p)}, \quad \text{as promised.} \quad \square$$

---

<sup>2</sup>Law of total derivative:  

$$\frac{df}{dp} = \sum_{w \in \mathcal{VX}} \frac{\partial \mathbf{D}(\mu(y|x) \parallel p)}{\partial \mu_p^*(w)}(\mu_p^*, p) \frac{\partial \mu_p^*(w)}{\partial p} + \frac{\partial}{\partial p} \mathbf{D}(\mu_p^*, p) + \sum_{w \in \mathcal{VX}} \frac{\partial \llbracket \mathbf{m} \rrbracket_\gamma}{\partial \mu_p^*(w)}(\mu_p^*) \frac{\partial \mu_p^*(w)}{\partial p}.$$

**Lemma 9.8.** *If  $0 < p_1 < p_2 < 1$  and  $\mu_1^*, \mu_2^*$  are respective minimizing distributions, then  $f(p_2) \geq f(p_1) + f'(p_1)(p_2 - p_1) + \frac{1}{2}\gamma\|\mu_1^* - \mu_2^*\|_1^2$ .*

*Proof.* The general approach is to adapt and strengthen the proof of Lemma 9.6, to show something like strong convexity, in this special case. Define

$$\mathbf{m}_1 := \mathbf{m} + \Pr(Y=y) = p_1 \quad \text{and} \quad \mathbf{m}_2 := \mathbf{m} + \Pr(Y=y) = p_2.$$

Choose  $\mu_1 \in [\![\mathbf{m}_1]\!]_\gamma^*$  and  $\mu_2 \in [\![\mathbf{m}_2]\!]_\gamma^*$ . As before, let  $m_1 := \mu_1(Y=y)$  and  $m_2 := \mu_2(Y=y)$ . Then

$$f(p_1) = \langle \langle \mathbf{m}_1 \rangle \rangle_\gamma = [\![\mathbf{m}_1]\!]_\gamma(\mu_1) = [\![\mathbf{m}]\!]_\gamma(\mu_1) + D(m_1 \parallel p_1)$$

and       $f(p_2) = \langle \langle \mathbf{m}_2 \rangle \rangle_\gamma = [\![\mathbf{m}_2]\!]_\gamma(\mu_2) = [\![\mathbf{m}]\!]_\gamma(\mu_1) + D(m_2 \parallel p_2),$

where, as before,  $D(m \parallel p) = m \log \frac{m}{p} + (1-m) \log \frac{1-m}{1-p}$  is the relative entropy between Bernoulli distributions with respective parameters  $m$  and  $p$ .

For each  $\lambda \in [0, 1]$ , define

$$p_\lambda := (1-\lambda)p_1 + \lambda p_2,$$

$$\mathbf{m}_\lambda := \mathbf{m} + \Pr(Y=y) = p_\lambda,$$

and       $\mu_\lambda := (1-\lambda)\mu_1 + \lambda\mu_2 .$

We now provide stronger analogues of (9.4) and (9.5). Since  $[\![\mathbf{m}]\!]_\gamma$  is not just convex but also  $\gamma$ -strongly convex, with respect to the 1-norm (Lemma 9.9, below), we can strengthen (9.4) to

$$[\![\mathbf{m}]\!]_\gamma(\mu_\lambda) \leq (1-\lambda)[\![\mathbf{m}]\!]_\gamma(\mu_1) + \lambda[\![\mathbf{m}]\!]_\gamma(\mu_2) - \frac{\gamma}{2}(1-\lambda)\lambda\|\mu_1 - \mu_2\|_1^2.$$

Adding this inequality to the analogous one describing joint convexity of  $D$  in

its two arguments (9.5), we find that

$$\begin{aligned}
& \llbracket \mathbf{m} \rrbracket_\gamma(\mu_\lambda) + \mathbf{D}(m_\lambda \| p_\lambda) \\
& \leq (1 - \lambda) \left( \llbracket \mathbf{m} \rrbracket_\gamma(\mu_1) + \mathbf{D}(m_1 \| p_1) \right) + \lambda \left( \llbracket \mathbf{m} \rrbracket_\gamma(\mu_2) + \mathbf{D}(m_2 \| p_2) \right) \\
& \quad - \frac{\gamma}{2} (1 - \lambda) \lambda \|\mu_1 - \mu_2\|_1^2 \\
& = (1 - \lambda)f(p_1) + \lambda f(p_2) - \frac{\gamma}{2} (1 - \lambda) \lambda \|\mu_1 - \mu_2\|_1^2. \tag{9.8}
\end{aligned}$$

Putting it all together, we find that

$$\begin{aligned}
f(p_\lambda) &= \langle \langle \mathbf{m}_\lambda \rangle \rangle_\gamma \\
&\leq \llbracket \mathbf{m}_\lambda \rrbracket_\gamma(\mu_\lambda) \\
&= \llbracket \mathbf{m} \rrbracket_\gamma(\mu_\lambda) + \mathbf{D}(m_\lambda \| p_\lambda) \\
&\leq (1 - \lambda)f(p_1) + \lambda f(p_2) - \frac{\gamma}{2} (1 - \lambda) \lambda \|\mu_1 - \mu_2\|_1^2 \tag{9.8}.
\end{aligned}$$

Since this is true for all  $\lambda \in [0, 1]$ , we can divide by  $\lambda$ , rearrange, and take the limit as  $\lambda \rightarrow 0$ , to find:

$$\begin{aligned}
\frac{\gamma}{2} (1 - \lambda) \|\mu_1 - \mu_2\|_1^2 &\leq \frac{(1 - \lambda)f(p_1) - f(p_1 + \lambda(p_2 - p_1))}{\lambda} + f(p_2) \\
&= \frac{f(p_1) - f(p_1 + \lambda(p_2 - p_1))}{\lambda} + f(p_2) - f(p_1) \\
\implies f(p_2) - f(p_1) &\geq \lim_{\lambda \rightarrow 0} f(p_1) + \frac{f(p_1 + \lambda(p_2 - p_1)) - f(p_1)}{\lambda} \frac{\gamma}{2} (1 - \lambda) \|\mu_1 - \mu_2\|_1^2 \\
&= f'(p_1)(p_2 - p_1) + \frac{\gamma}{2} \|\mu_1 - \mu_2\|_1^2,
\end{aligned}$$

as desired.  $\square$

**Lemma 9.9.** *Negative entropy is 1-strongly convex with respect to the L1 norm, i.e.,  $f(\mathbf{p}) = \sum_i p_i \log p_i$  satisfies*

$$f(\mathbf{q}) \geq f(\mathbf{p}) + \nabla f(\mathbf{p})^\top (\mathbf{q} - \mathbf{p}) + \frac{1}{2} \|\mathbf{p} - \mathbf{q}\|_1^2.$$

*Proof.* First,  $\nabla f(\mathbf{p}) = \log(\mathbf{p}) + 1$ . Thus,

$$\begin{aligned}
& f(\mathbf{q}) - f(\mathbf{p}) - \nabla f(\mathbf{p})^\top (\mathbf{q} - \mathbf{p}) \\
&= \sum_i [q_i \log q_i - p_i \log p_i - (\log p_i + 1)(q_i - p_i)] \\
&= \sum_i [q_i \log q_i - p_i \log p_i - q_i \log p_i + p_i \log p_i] \\
&= \sum_i q_i \log \frac{q_i}{p_i} \\
&= D(q \| p) \\
&\geq 2\delta(\mathbf{p}, \mathbf{q})^2 \quad [\text{by Pinsker's inequality (Tsybakov 2009)}] \\
&= 2\left(\frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_1\right)^2 \\
&= \frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_1^2
\end{aligned}$$

where  $\delta(\mathbf{p}, \mathbf{q})$  is the total variation distance between  $\mathbf{p}$  and  $\mathbf{q}$  as measures, and  $\|\mathbf{p} - \mathbf{q}\|_1 = \sum_i |p_i - q_i|$  is the L1 norm of their difference, as points on a simplex.  $\square$

[Lemma 9.8](#) guarantees that if the optimal distributions corresponding to adding  $p_1$  and  $p_2$  to the PDG ( $\mu_1$  and  $\mu_2$ , respectively) are far apart, then so are  $f(p_1)$  and  $f(p_2)$ . But what if these optimal distributions are close together? It turns out that if  $\mu_1 \approx \mu_2$  then it's still the case that  $f(p_1)$  and  $f(p_2)$  are far apart, provided that  $p_1$  and  $p_2$  are. However, showing this requires an entirely approach, which we pursue in [Lemma 9.11](#). But first, we need two intermediate technical results.

**Lemma 9.10.** *For all  $p_1, p_2 \in [0, 1]$ ,*

$$[\![\mathbf{m}]\!]_\gamma(\mu_1^*) - [\![\mathbf{m}]\!]_\gamma(\mu_2^*) \geq (m_2 - m_1) \log \frac{m_2}{p_2} \frac{1 - p_2}{1 - m_2},$$

where  $m_1 = \mu_1^*(Y=y)$  is the marginal of  $\mu_1^* \in [\![\mathbf{m} + \Pr(Y=y) = p_1]\!]_\gamma^*$ , and  $m_2 = \mu_2^*(Y=y)$  is defined symmetrically.

*Proof.* For  $p, q \in [0, 1]$ ,  $D(p \parallel q) := p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$  is the relative entropy between the Bernoulli distributions described by their parameters. First, we calculate

$$\begin{aligned}\frac{d}{dp} D(p \parallel q) &= \log \frac{p}{q} + \frac{p}{q} q \frac{d}{dp} \left[ \frac{p}{q} \right] + (-1) \log \frac{1-p}{1-q} + (1-p) \frac{d}{dp} \left[ \frac{1-p}{1-q} \right] \\ &= \log \frac{p}{q} + 1 - \log \frac{1-p}{1-q} - 1 \\ &= \log \left( \frac{p}{q} \frac{1-q}{1-p} \right).\end{aligned}$$

Thus,

$$\begin{aligned}\nabla_{\mu} [D(\mu(Y=y) \parallel q)] &= \nabla_{\mu} [\mu(Y=y)] \log \left( \frac{\mu(Y=y)}{q} \frac{1-q}{1-\mu(Y=y)} \right) \\ &= \mathbb{1}[Y=y] \log \left( \frac{\mu(Y=y)}{q} \frac{1-q}{1-\mu(Y=y)} \right).\end{aligned}$$

Recall the stationary conditions, which state that, since  $\mu_2^* \in [\mathbf{m} + \Pr(Y=y) = p_2]$ , we have

$$\nabla_{\mu} [D(\mu(Y=y) \parallel p_2)]_{\mu=\mu_2^*} = -\nabla_{\mu} [\mathbf{m}]_{\gamma}(\mu_2^*).$$

Now use the above to compute the directional derivative of interest:

$$\begin{aligned}\nabla_{\mu} [\mathbf{m}]_{\gamma}(\mu_2^*)^T (\mu_1^* - \mu_2^*) \\ &= -(\mu_1^* - \mu_2^*)^T \nabla_{\mu} [D(\mu(Y=y) \parallel p_2)]_{\mu=\mu_2^*} \\ &= (\mu_2^*(Y=y) - \mu_1^*(Y=y)) \log \left( \frac{\mu_2^*(Y=y)}{p_2} \frac{1-p_2}{1-\mu_2^*(Y=y)} \right) \\ &= (m_2 - m_1) \log \frac{m_2}{p_2} \frac{1-p_2}{1-m_2}.\end{aligned}$$

Finally, since  $[\mathbf{m}]_{\gamma}$  is convex, we have

$$\begin{aligned}[\mathbf{m}]_{\gamma}(\mu_1^*) - [\mathbf{m}]_{\gamma}(\mu_2^*) &\geq \nabla_{\mu} [\mathbf{m}]_{\gamma}(\mu_2^*)^T (\mu_1^* - \mu_2^*) \\ &= s_1(m_2 - m_1) \log \frac{m_2}{p_2} \frac{1-p_2}{1-m_2}\end{aligned}$$

as promised. □

**Lemma 9.11.** Suppose that  $0 < b < z < p^* < 1$ , and let

$$\mu_z^* \in [\![m + \Pr(Y=y) = z]\!]_\gamma^* \quad \text{and} \quad \mu_b^* \in [\![m + \Pr(Y=y) = b]\!]_\gamma^*$$

be the respective optimal distributions for their corresponding PDGs. If  $\|\mu_b - \mu_z\|_1 \leq \delta$ , then

$$f(b) - f(z) \geq (z - b)^2 - \left( \frac{2}{b} + \log \frac{1-b}{1-z} + \log \frac{1}{b} \right) \delta.$$

*Proof.* Because we have assumed  $b < z < p^*$  and  $f$  is strictly convex (Lemma 9.6), it follows from Lemma 9.7 that  $m_z := \mu_z^*(Y=y) \geq z$  and  $m_b := \mu_b^*(Y=y) \geq b$ .

Now

$$\begin{aligned} f(b) - f(z) &= (\llbracket m \rrbracket_\gamma(\mu_b^*) + D(m_b \| b)) - (\llbracket m \rrbracket_\gamma(\mu_z^*) + D(m_z \| z)) \\ &= (\llbracket m \rrbracket_\gamma(\mu_b^*) - \llbracket m \rrbracket_\gamma(\mu_z^*)) + D(m_b \| b) - D(m_z \| z). \end{aligned}$$

Lemma 9.10 will give us a lower bound for the first half; we now investigate the second half. The first step is some algebraic manipulation.

$$\begin{aligned} D(m_b \| b) - D(m_z \| z) &= -m_z \log \frac{m_z}{z} - (1-m_z) \log \frac{1-m_z}{1-z} + m_b \log \frac{m_b}{b} + (1-m_b) \log \frac{1-m_b}{1-b} \\ &= m_z \log \frac{z}{m_z} + (1-m_z) \log \frac{1-z}{1-m_z} + (m_b + m_z - m_z) \log \frac{m_b}{b} \\ &\quad + ((1-m_b) + (1-m_z) - (1-m_z)) \log \frac{1-m_b}{1-b} \quad (\text{add zero}) \\ &= m_z \left( \log \frac{z}{m_z} + \log \frac{m_b}{b} \right) + (1-m_z) \left( \log \frac{1-z}{1-m_z} + \log \frac{1-m_b}{1-b} \right) \\ &\quad + (m_b - m_z) \log \frac{m_b}{b} + ((1-m_b) - (1-m_z)) \log \frac{1-m_b}{1-b} \\ &\quad \quad \quad (\text{collect } z\text{-marginal terms}) \\ &= m_z \log \frac{z}{b} \frac{\cancel{m_b}}{\cancel{m_z}} + (1-m_z) \log \frac{1-z}{1-b} \frac{\cancel{1-m_b}}{\cancel{1-m_z}} \\ &\quad + (m_b - m_z) \log \frac{m_b}{b} \frac{1-b}{1-m_b} \\ &=: \blacksquare_0 + \blacksquare_1 + \blacksquare_2, \end{aligned}$$

where, to be explicit, we have defined

$$\begin{aligned}\blacksquare_0 &= m_z \log \frac{z}{b} + (1 - m_z) \log \frac{1 - z}{1 - b} \\ \blacksquare_1 &= m_z \log \frac{m_b}{m_z} + (1 - m_z) \log \frac{1 - m_b}{1 - m_z} = -D(m_z \parallel m_b) \\ \blacksquare_2 &= (m_b - m_z) \log \frac{m_b}{b} \frac{1 - b}{1 - m_b}.\end{aligned}$$

There is one final quantity that will play a similar role. Let

$$\blacksquare_4 := (m_z - m_b) \log \frac{m_z}{z} \frac{1 - z}{1 - m_z}$$

be the lower bound on  $\llbracket M \rrbracket_\gamma(\mu_b^*) - \llbracket M \rrbracket_\gamma(\mu_z^*)$  obtained by applying Lemma 9.10 with  $p_1 = b$  and  $p_2 = z$ . With these definitions, we have  $f(b) - f(z) \geq \blacksquare_0 + \blacksquare_1 + \blacksquare_2 + \blacksquare_4$ .

Observe that if  $m_z$  were equal to  $z$ , then  $\blacksquare_0$  would equal  $D(z \parallel b)$ . But in fact we know that  $m_z > z$ . It is easy to see that that  $\blacksquare_0$  is linear in  $m_z$  with positive slope, since  $z > b$  and  $1 - b > 1 - z$ . It follows that  $\blacksquare_0 > D(z \parallel b)$ .

Let's step back for a moment. Lemma 9.8 shows that  $f(b)$  and  $f(z)$  cannot be too close, provided that  $\mu_z^*$  and  $\mu_b^*$  are far apart. In the equations above, we can see the beginnings of a complementary argument: if  $\mu_z^*$  and  $\mu_b^*$  are close together, then  $m_b \approx m_b$ , and so all terms apart from  $\blacksquare_0$  (i.e.,  $\blacksquare_1 + \blacksquare_2 + \blacksquare_4$ ) go to zero. And yet, because of  $\blacksquare_0$ ,  $f(b)$  and  $f(z)$  remain far apart. To make this argument precise,

we now merge  $\blacksquare_1$ ,  $\blacksquare_2$ , and  $\blacksquare_4$  back together, calculating

$$\begin{aligned}
& \blacksquare_2 + \blacksquare_4 + \blacksquare_1 \\
&= (m_b - m_z) \log \frac{m_b}{b} \frac{1-b}{1-m_b} \frac{z}{m_z} \frac{1-m_z}{1-z} - D(m_z \| m_b) \\
&= (m_b - m_z) \log \frac{z}{b} \frac{1-b}{1-z} + (m_b - m_z) \log \frac{m_b}{m_z} \frac{1-m_z}{1-m_b} + m_z \log \frac{m_b}{m_z} + (1-m_z) \log \frac{1-m_b}{1-m_z} \\
&= (m_b - m_z) \log \frac{z}{b} \frac{1-b}{1-z} + (m_b - \cancel{m_z} + \cancel{m_z}) \log \frac{m_b}{m_z} + (1 - \cancel{m_z} + \cancel{m_z} - m_b) \frac{1-m_b}{1-m_z} \\
&= (m_b - m_z) \log \frac{z}{b} \frac{1-b}{1-z} + D(m_b \| m_z) \\
&\geq (m_b - m_z) \log \frac{z}{b} \frac{1-b}{1-z}.
\end{aligned}$$

Suppose that  $\|\mu_z^* - \mu_b^*\|_1 \leq \delta$ . Because the total variation distance  $\text{TV}(p, q)$  is half the L1-norm  $\|p - q\|_1$  for discrete distributions,

$$\delta \geq \|\mu_z^* - \mu_b^*\|_1 = 2 \text{TV}(\mu_z^*, \mu_b^*) \geq 2 |\mu_z^*(Y=y) - \mu_b^*(Y=y)| = 2|m_z - m_b|.$$

Thus,

$$\blacksquare_1 + \blacksquare_2 + \blacksquare_4 \geq -\frac{\delta}{2} \log \frac{z}{b} \frac{1-b}{1-z}.$$

All that remains is  $\blacksquare_0$ . To put things in a convenient form, we apply Pinsker's inequality. The total variation distance between two Bernoulli distributions (i.e., binary distributions) with respective positive probabilities  $p$  and  $q$  is just  $|p - q|$ . Pinsker's inequality (Tsybakov 2009) in this case says:  $\frac{1}{2}D(p \| q) \geq (p - q)^2$ . Thus,  $\blacksquare_0 > D(z \| b) \geq 2(z - b)^2$ , and so we have

$$f(b) - f(z) \geq 2(z - b)^2 - \left( \frac{1}{2} \log \frac{z}{b} \frac{1-b}{1-z} \right) \delta \quad \text{as promised. } \square$$

**Lemma 9.12.** Suppose that  $b < z < p^*$ . Furthermore, suppose that  $|\log \frac{z}{b} \frac{1-b}{1-z}| \leq k$ . Not only is it the case that  $f(b) > f(z)$ , but also

$$f(b) - f(z) \geq \frac{k^2}{32\gamma} \log^2 \left( 1 + 16 \frac{\gamma}{k^2} (z - b)^2 \right).$$

*Proof.* We now have two bounds that work in different regimes. If  $\delta = \|\mu_b^* - \mu_z^*\|_2$  is large, then the argument of Lemma 9.8 is effective, as it shows that a separation between  $f(b)$  and  $f(z)$  that scales with  $\delta^2$ . On the other hand, if  $\delta$  is small, we saw in Lemma 9.11 a very different approach that still gets us a separation of  $2(z - b)^2$  even if  $\delta = 0$ . We now combine the two cases to eliminate the (unknown) parameter  $\delta$  from our complexity analysis. (Our algorithm is no different in the two cases; all that differs is the analysis.)

Taken together, we know that we attain the maximum of the two lower bounds, which is weakest when they coincide. We can then solve for the worst-case value of  $\delta$ , which leads to the smallest possible separation between  $f(z)$  and  $f(b)$ . Setting the two bounds equal to one another:

$$\begin{aligned} \frac{1}{2}\gamma\delta_{\text{worst}}^2 &= 2(z - b)^2 - \left(\frac{1}{2}\log\frac{z}{b}\frac{1-b}{1-z}\right)\delta \\ \iff \quad \frac{1}{2}\delta_{\text{worst}}^2 + \left(\frac{1}{2}\log\frac{z}{b}\frac{1-b}{1-z}\right)\delta_{\text{worst}} - 2(z - b)^2 &= 0. \end{aligned}$$

The quadratic equation then tells us that

$$\delta_{\text{worst}} = \frac{1}{\gamma} \left( -B + \sqrt{B^2 + 4\gamma(z - b)^2} \right), \quad \text{where} \quad B := \frac{1}{2} \log \frac{z}{b} \frac{1-b}{1-z}.$$

It is easily verified that this expression for  $\delta_{\text{worst}}$  is decreasing in  $B$ . Therefore, we get a lower bound on it by plugging in our upper bound  $\frac{k}{2}$  for  $B$ . Thus  $\delta_{\text{worst}} \geq \frac{1}{\gamma} \left( -k + \sqrt{k^2 + 4\gamma(z - b)^2} \right)$ .

Square roots are not easy to manipulate in general, and this expression in particular has involves a nested subtraction that makes it hard to characterize. To make things clearer, we now begin to loosen this bound to get a quantity that is easier to think about. The first observation is that, for any numbers  $A, B > 0$ ,

$$-B + \sqrt{B^2 + A} = B \left( -1 + \sqrt{1 + \frac{A}{B^2}} \right).$$

This manipulation puts the square root in the standard form  $\sqrt{1+x}$ . Here is the second observation: for all  $x$ ,  $-1 + \sqrt{1+x} \geq \frac{1}{2} \log(1+x)$  (verified in [Lemma 9.13](#) below). Although it gives a looser bound, the logarithm is easier to manipulate and no longer involves subtraction. Applying these two transformations in our case, we find:

$$\begin{aligned}\delta_{\text{worst}} &= \frac{1}{\gamma} \left( -B + \sqrt{B^2 + 4\gamma(z-b)^2} \right) \\ &\geq \frac{1}{\gamma} \left( -\frac{k}{2} + \sqrt{\frac{k^2}{4} + 4\gamma(z-b)^2} \right) \\ &= \frac{k}{2\gamma} \left( -1 + \sqrt{1 + \frac{16}{k^2}\gamma(z-b)^2} \right) \\ &\geq \frac{k}{4\gamma} \log \left( 1 + \frac{16}{k^2}\gamma(z-b)^2 \right).\end{aligned}$$

Finally, since  $f(b) - f(z) \geq \frac{\gamma}{2} \delta_{\text{worst}}^2$ , to get lower bound for  $f(b) - f(z)$ , we simply need to square this lower bound for  $\delta_{\text{worst}}$  and multiply by  $\gamma/2$ . As a result,

$$\begin{aligned}f(b) - f(z) &\geq \frac{\gamma}{2} \frac{k^2}{16\gamma^2} \log^2 \left( 1 + \frac{16\gamma}{k^2}(z-b)^2 \right) \\ &= \frac{k^2}{32\gamma} \log^2 \left( 1 + \frac{16\gamma}{k^2}(z-b)^2 \right),\end{aligned}$$

where  $\log^2(x)$  means  $(\log(x))^2$ . □

**Lemma 9.13.**  $-1 + \sqrt{1+x} \geq \frac{1}{2} \log(1+x)$ .

*Proof.* Apply the well-known inequality  $e^y \geq 1 + y$  with  $y = -1 + \sqrt{1+x}$ , to get

$$\exp(-1 + \sqrt{1+x}) \geq \sqrt{1+x},$$

which implies  $-1 + \sqrt{1+x} \geq \log \sqrt{1+x} = \frac{1}{2} \log(1+x)$ . □

We are now ready to tackle the theorem itself.

### Theorem 9.4.

- (a) *There is an  $O(\log^{1/\epsilon})$ -time reduction from unconditional APPROX-INFER-CVX to the problem of determining which of two PDGs is more inconsistent, using  $O(\log^{1/\epsilon})$  subroutine calls.*
- (b) *There is an  $O\left(\log \frac{\langle\!\langle m \rangle\!\rangle_\gamma}{\gamma \epsilon \mu^*(x)} \cdot \log \frac{1}{\epsilon \mu^*(x)}\right)$  time reduction from APPROX-INFER-CVX to APPROX-CALC-INC using  $O(\log(1/\epsilon) \log \log 1/\mu^*(x))$  calls to the inconsistency subroutine.*
- (c) *There is also an  $O(|VC|)$  reduction from APPROX-CALC-INC to APPROX-INFER-CVX. With the additional assumption of bounded treewidth, this is linear in the number of variables in the PDG.*

*Proof. (a,b).* Suppose that we have access to a procedure that can calculate a PDG's degree of inconsistency. The idea behind the reductions of both (a) and (b) is to perform  $\hat{\gamma}$ -inference on a given PDG  $m$ , using this procedure as a subroutine. The complexity of the reduction depends on the specification of the inconsistency calculation procedure. We will perform two analyses, the second building on the first.

1. First, we assume the procedure simply tells us which of two PDGs is more inconsistent. With this assumption, we get an algorithm that can answer unconditional probability queries with the optimal complexity of part (a) of the theorem.
2. We then provide a refinement of that algorithm that still works if the inconsistency calculation procedure produces only finite-precision binary approximations to inconsistency values—thus reducing the problem of approximate inference that of approximately calculating PDG inconsistency.

This considerably more difficult analysis gives us part (b). In addition, we extend the algorithm, using [Lemma 8.17](#), so that it can also answer conditional queries.

We begin by describing our algorithm, which uses the first variant of the inconsistency procedure (the one that tells us which of two PDGs is more inconsistent) to produce a sequence of approximations  $(p_1, p_2, \dots)$  that converges exponentially to

$$p^* := \llbracket \mathbf{m} \rrbracket_{\gamma}^*(Y=y) \stackrel{\text{(Lemma 9.5)}}{=} \arg \min_p \left\langle \mathbf{m} + (\Pr(Y=y) = p) \right\rangle_{\gamma}$$

through a variant of binary search. More precisely, the algorithm we present below is an extremely close relative of an algorithm known to the competitive programming community as *trinary search* ([tri 2023](#)). In some ways it is slightly more efficient, but its analysis is more complex.

The state of the algorithm consists of three points in an interval  $a, b, c \in [0, 1]$ , where  $a \leq b \leq c$ . Intuitively,  $b$  is our current best guess at  $p^*$ , while  $a$  is a lower bound, and  $c$  is an upper bound. Once again (as in [\(9.6\)](#) and in preceding lemmas), let  $f$  be the function

$$\begin{aligned} f : [0, 1] &\rightarrow \bar{\mathbb{R}} \\ p &\mapsto \left\langle \mathbf{m} + (\Pr(Y=y) = p) \right\rangle_{\gamma}. \end{aligned}$$

Both variants of the inconsistency calculation procedure will be employed for the sole purpose of determining whether or not  $f(p) > f(p')$ , given  $p, p' \in [0, 1]$ . We start with the simpler variant, which can directly determine which of two PDGs has greater inconsistency. In this case, the  $\gg$  and  $\ll$  in the algorithm below should be interpreted simply as  $>$  as  $<$ . This will enable us to approximate the minimizer  $p^*$  of  $f$  arbitrarily closely, with the following algorithm ([Algorithm 3](#)).

---

**Algorithm 3** Unimodal (Trinary) Search

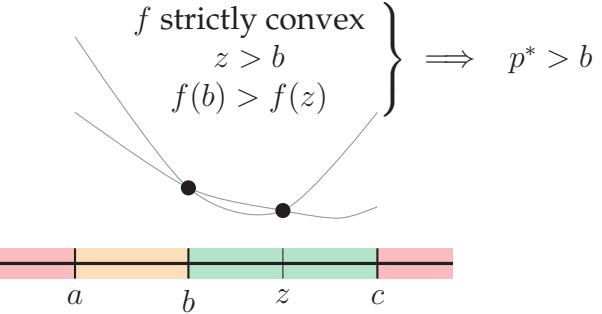
---

```

Initialize  $(a, b, c) \leftarrow (0, \frac{1}{2}, 1)$ ;
while  $|c - a| > \epsilon$  do
    if  $b - a \geq c - b$  then
        Let  $z := \frac{b+a}{2}$ ;
        if  $f(b) \gg f(z)$  then
             $(a, b, c) \leftarrow (a, z, b)$ ;
        else
             $(a, b, c) \leftarrow (z, b, c)$ ;
    else if  $c - b > b - a$  then
        Let  $z := \frac{b+c}{2}$ ;
        if  $f(z) \ll f(b)$  then
             $(a, b, c) \leftarrow (b, z, c)$ ;
        else
             $(a, b, c) \leftarrow (a, b, z)$ ;
    return  $b$ ;

```

---



We begin by proving that [Algorithm 3](#) does indeed output a point within  $\epsilon$  of  $p^*$ . Because  $f$  is convex, this algorithm satisfies an important invariant:

**Claim 9.13.1.** *Both  $b$  and  $p^*$  always lie in the interval  $[a, c]$ .*

*Proof.* We proceed by induction on  $i$ . At the beginning, it is obviously true that  $b$  and  $p^*$  lie in  $[a, b] = [0, 1]$ , which contains the entire domain of  $f$ . Now, suppose inductively that  $p^* \in [a, b]$  at some iteration of the algorithm  $i$ .

(case 1) If  $b - a \geq c - b$ , then  $z \in [a, b]$ .

- Suppose  $f(z) < f(b)$ . Then for all  $y > b$ , it must be the case that  $f(y) > f(b)$  by convexity of  $f$ . (For if  $f(y) < f(b)$ , then segment between  $(z, f(z))$  and  $(y, f(y))$  would lie entirely below  $(b, f(b))$ , which contradicts convexity).

Thus, we can rule out all such  $y$  as possible minimizers of  $f$ , so we can restrict our attention to  $[a, b]$ , which contains  $p^*$  (and  $x$ ).

- On the other hand, if  $f(z) > f(b)$ , then it must be the case that no  $y < z$  can be a minimizer of  $f$  by convexity, with the same reasoning as above. (Namely, if  $f(y) < f(z)$  then the segment between  $(y, f(y))$  and  $(b, f(b))$  lies below  $(z, f(z))$ , contradicting convexity). Thus the true minimizer  $p^*$  lies in  $[z, c]$ , an interval which contains  $b$ .

(case 2) The other case is symmetric; we include it for completeness. Suppose

$$c - b > b - a, \text{ and so } z = \frac{b+c}{2}.$$

- Suppose  $f(z) < f(b)$ . Then  $f(y) > f(b)$  for all  $y < b$  (because if  $f(y) < f(b)$ , then segment between  $(y, f(y))$  and  $(x, f(x))$  would lie below  $(b, f(b))$ ). So  $p^*, z \in [b, c]$ .
- On the other hand, if  $f(z) > f(b)$ , then  $f(y) > f(z)$  for all  $y > z$  (because, if  $f(y) < f(z)$  then the segment between  $(y, f(y))$  and  $(b, f(b))$  lies below  $(z, f(z))$ , contradicting convexity). So  $p^*, b \in [a, z]$ .  $\square$

In every case,  $p^*$  is still in what becomes the interval  $[a, c]$  in the next iteration ( $i + 1$ ). So, by induction,  $p^* \in [a, b]$  at every iteration of the algorithm, proving

**Claim 9.13.1.**  $\square$

We have shown that both  $b$  and  $p^*$  lie within  $[a, c]$ , and we know that, at termination,  $|c - a| < \epsilon$ . Therefore, the final value of  $b$  (i.e., the output of the algorithm) must be within  $\epsilon$  of  $p^*$ .

Next, we analyze the complexity of this algorithm, modulo the complexity of comparing the numbers  $f(z)$  and  $f(b)$ , which we will later bound precisely. Each iteration reduces the size of the interval  $[a, c]$  by a factor of at least  $3/4$ . This is

because in each case we focus on the larger half of the interval, and ultimately discard either half or all of it—so we reduce the size of the interval by at least one quarter. It follows that, after  $n$  iterations, the size of the interval is at most  $(\frac{3}{4})^n$ , and thus the total number of iterations is at most  $\lceil \log(\frac{1}{\epsilon}) / (\log \frac{4}{3}) \rceil$ . Apart from the time needed to compare  $f(b)$  and  $f(z)$ , it is easy to see that each iteration of the algorithm takes constant time. So overall, it requires  $\log(\frac{1}{\epsilon})$  space(enough to track the numbers  $\{a, b, c\}$ , plus a reference to the PDG  $m$ ), and time  $O(\log \frac{1}{\epsilon})$ , (linear in the number of bits returned). This completes the proof of [Theorem 9.4](#) (a).

Although it is common to assume that numbers can be compared in  $O(1)$  time, and this is an assumption well suited to modern computer architecture, it is arguably not appropriate in this context. How do we know a `float64` has enough precision to do the comparison? The obvious approach to implementing the inconsistency calculation subroutine would be to repeatedly request more and more precise estimates of inconsistency, until one is larger than the other—but this procedure does not terminate if the two PDGs have the same inconsistency. So, a priori, it's not even clear that the decision  $f(b) > f(z)$  is computable. It is not hard to show that it is in fact computable. Because  $f$  is strictly convex, it must be the case that  $f(b) > f(z)$ ,  $f(z) > f(b)$ , or  $f(b) > f(\frac{b+z}{2})$ . In the last case, we can act as if  $f(b) > f(z)$ , and the algorithm will be correct, because the argument supporting [Claim 9.13.1](#) still applies. Thus, by running the subroutine on all three questions until one of them returns true, and then aborting the other two calculations, we can see that the problem of interest is decidable. But how long does it take? We now provide a deeper analysis of the comparison between  $f(z)$  and  $f(b)$  when the inconsistency calculation procedure can give us only finite approximations to the true value.

**Part (b): reduction to approximate inconsistency calculation.** Instead of assuming that we have direct access to the numbers  $f(b)$  and  $f(z)$  and can compare them in one step, we now adopt a weaker assumption, that we only have access to finite approximations to them. With this model of computation, it is not obvious that we can determine which of  $f(b)$  and  $f(z)$  is bigger—but fortunately, we do not need to. This is because, when  $f(z)$  and  $f(b)$  are close,  $p^*$  lies between  $z$  and  $b$ , and so both branches of the algorithm maintain the invariant  $p^* \in [a, c]$ . To simplify our analysis, we will default to keeping the “left” branch (with the smaller numbers), if the queried approximations to  $f(z)$  and  $f(b)$  are too close to determine whether one is larger than the other.

More precisely, the test “ $f(z) \ll f(b)$ ” is now shorthand for the following procedure:

- Run the inconsistency calculation procedure to obtain approximations to  $f(z)$  and  $f(b)$  that are correct to within

$$\epsilon' := \frac{1}{16\gamma} \log^2 \left( 1 + 8\gamma(z - b)^2 \right). \quad (\text{This number comes from Lemma 9.12.}) \quad (9.9)$$

Call these approximations  $\tilde{f}(z)$  and  $\tilde{f}(b)$ . By definition, they satisfy  $|f(z) - \tilde{f}(z)| \leq \epsilon'$  and  $|f(b) - \tilde{f}(b)| \leq \epsilon'$ . If  $|\tilde{f}(z) - \tilde{f}(b)| > \epsilon'$  (so that we know for sure which of  $f(z)$  and  $f(b)$  is larger based on these approximations), then return TRUE. If  $\tilde{f}(z) < \tilde{f}(b)$ , and FALSE otherwise.

- On the other hand, if  $|\tilde{f}(z) - \tilde{f}(b)| \leq \epsilon'$ , return TRUE if  $z < b$  and FALSE if  $b > z$ .

The remainder of the proof of correctness demonstrates that this level of precision is enough to never mistakenly eliminate the branch containing  $p^*$ .

We begin by proving a series of three additional invariants about the values  $(a, b, z, c)$  in each iteration, which are required for our analysis. The first property is that  $b$  and  $z$  are not too close to the boundary or each other.

**Claim 9.13.2.** *At the beginning of each iteration,*

$$b \in [\frac{\epsilon}{2}, 1 - \frac{\epsilon}{2}], \quad z \in [\frac{\epsilon}{4}, 1 - \frac{\epsilon}{4}], \quad \text{and} \quad |b - z| \geq \frac{\epsilon}{4}.$$

We prove this by contradiction. Initially,  $b = \frac{1}{2}$  so it's neither the case that  $b < \frac{\epsilon}{2}$  nor that  $b > 1 - \frac{\epsilon}{2}$  for any  $\epsilon < 1$ . (The procedure terminates immediately if  $\epsilon \geq 1$ .) In search of a contradiction, suppose that either  $b < \frac{\epsilon}{2}$  or  $b > 1 - \frac{\epsilon}{2}$  later on. Specifically, suppose it first occurs in the  $(t + 1)^{\text{st}}$  iteration, and let  $(a_{t+1}, b_{t+1}, c_{t+1})$  to refer to the values of  $(a, b, c)$  in that iteration. Let  $(a_t, b_t, z_t, c_t)$  denote the values of the variables in the previous iteration. We know that  $b_{t+1} \notin [\frac{\epsilon}{2}, 1 - \frac{\epsilon}{2}]$  and  $b_t \in [\frac{\epsilon}{2}, 1 - \frac{\epsilon}{2}]$ . In particular,  $b_t \neq b_{t+1}$ , which means the procedure cannot have taken the second or fourth branches in the  $t^{\text{th}}$  iteration. There are two remaining cases, corresponding to the first and third branches.

- **(branch 1)** In this case,  $b_t - a_t \geq c_t - b_t$  and  $z_t = (a_t + b_t)/2$ . Furthermore, as a result of the assignment in this branch, we have  $c_{t+1} = b_t$  and  $b_{t+1} = z_t = (a_t + b_t)/2$ .

– If  $b_{t+1} < \frac{\epsilon}{2}$ , this means  $a_t + b_t < \epsilon$ . As  $a_t \geq 0$ , this implies  $b_t = c_{t+1} < \epsilon$ . But then  $|c_{t+1} - a_{t+1}| \leq c_{t+1} < \epsilon$ , so the algorithm must have already terminated! This is a contradiction.

– If  $b_{t+1} > 1 - \frac{\epsilon}{2}$ , then  $1 - \frac{\epsilon}{2} < b_{t+1} = z_t = (a_t + b_t)/2 < b_t$ , contradicting our assumption that  $b_t \leq 1 - \frac{\epsilon}{2}$ .

- **(branch 3)** In this case,  $c_t - b_t > b_t - a_t$  and  $z_t = (b_t + c_t)/2$ . The assignment at the end of this branch ensures that  $a_{t+1} = b_t$  and  $b_{t+1} = z_t$ .

- If  $b_{t+1} < \frac{\epsilon}{2}$ , then  $b_t = a_{t+1} < b_{t+1} < \frac{\epsilon}{2}$ . which is a contradiction.
- If  $b_{t+1} = (b_t + c_t)/2 > 1 - \frac{\epsilon}{2}$ , then, since  $c_t \leq 1$ , we know  $b_t + 1 > 2 - \epsilon$ , so  $b_t = a_{t+1} > 1 - \epsilon$ . But now  $|c_{t+1} - a_{t+1}| \leq 1 - a_{t+1} < 1 - (1 - \epsilon) = \epsilon$ . So the algorithm must have already terminated.

Thus, it cannot be the case that  $b < \frac{\epsilon}{2}$  or  $b > 1 - \frac{\epsilon}{2}$  in any iteration of the algorithm. The fact that  $z \in [\frac{\epsilon}{4}, 1 - \frac{\epsilon}{4}]$  follows immediately from the definition of  $z$  in either branch. Finally,  $|z - b| = \frac{1}{2} \max\{c - b, b - a\} \geq \frac{1}{2}(\frac{c-a}{2}) > \frac{\epsilon}{4}$ . This completes the proof of [Claim 9.13.2](#).  $\square$

**Claim 9.13.3.** *It is always the case that  $b \in [\frac{2a+c}{3}, \frac{a+2c}{3}]$ .*

We prove this by induction. It is clearly true at initialization; suppose it is also true at time  $t$ , i.e.,  $\frac{2a_t+c_t}{3} \leq b_t \leq \frac{a_t+2c_t}{3}$ . We now show the same is true at time  $t + 1$  in each of the four cases of the algorithm.

- **(branch 1)** At the end, we assign  $a_{t+1} = a_t$ ,  $b_{t+1} = z = \frac{a_t+b_t}{2}$ , and  $c_{t+1} = b_t$ .

So,

$$\frac{2a_{t+1} + c_{t+1}}{3} = \frac{2a_t + b_t}{3} < \underbrace{\frac{a_t + b_t}{2}}_{= b_{t+1}} < \frac{a_t + 2b_t}{3} = \frac{a_{t+1} + 2c_{t+1}}{3}.$$

- **(branch 2)** In this case, we must make use of the fact that, in the first two branches  $b - a \geq c - b$ , meaning  $a_t + c_t \leq 2b_t$ . As in the first branch, we have  $z = \frac{a_t+b_t}{2}$ . This time, however,  $a_{t+1} = z = \frac{a_t+b_t}{2}$ ,  $b_{t+1} = b_t$ , and  $c_{t+1} = c_t$ . Thus, we find

$$\begin{aligned} \frac{2a_{t+1} + c_{t+1}}{3} &= \frac{a_t + b_t + c_t}{3} \leq \frac{(2b_t) + b_t}{3} = b_t \\ &= b_{t+1} \leq \frac{a_t + 2c_t}{3} < \frac{\frac{a_t+b_t}{2} + 2c_t}{3} = \frac{a_{t+1} + 2c_{t+1}}{3}. \end{aligned}$$

- **(branch 3)** Symmetric with branch 1. Concretely,  $a_{t+1} = b_t$ ,  $b_{t+1} = z = \frac{b_t+c_t}{2}$ ,

and  $c_{t+1} = c_t$ . Thus,

$$\frac{2a_{t+1} + c_{t+1}}{3} = \frac{2b_t + c_t}{3} < \underbrace{\frac{b_t + c_t}{2}}_{= b_{t+1}} < \frac{b_t + 2c_t}{3} = \frac{a_{t+1} + 2c_{t+1}}{3}.$$

- **(branch 4)** Symmetric with branch 2. Concretely,  $a_{t+1} = a_t$ ,  $b_{t+1} = b_t$ ,

$c_{t+1} = z = \frac{b_t+c_t}{2}$ , and we know  $2b_t < a_t + c_t$ . Thus,

$$\begin{aligned} \frac{2a_{t+1} + c_{t+1}}{3} &= \frac{2a_t + \frac{c_t+b_t}{2}}{3} < \frac{2a_t + c_t}{3} \leq b_t \\ &= b_{t+1} = \frac{2b_t + b_t}{3} < \frac{(a_t + c_t) + b_t}{3} = \frac{a_{t+1} + 2c_{t+1}}{3}. \end{aligned}$$

The final result we need is a bound for Lemma 9.12.

**Claim 9.13.4.**  $|\log \frac{z}{b} \frac{1-b}{1-z}| \leq \log 4$  ( $< \sqrt{2}$ ).

*Proof.* Let  $\bar{a} := 1 - a$ ,  $\bar{b} := 1 - b$ , and  $\bar{c} := 1 - c$ . Claim 9.13.3 tells us that  $b \geq \frac{2a+c}{3} \geq \frac{c}{3}$ , and also that  $b \leq \frac{a+2c}{3}$ , which gives us a symmetric fact:  $\bar{b} = 1 - b \geq \frac{2}{3} - \frac{a}{3} - \frac{2c}{3} = \frac{\bar{a}+2\bar{c}}{3} \geq \frac{\bar{a}}{3}$ . Consider two cases, corresponding to the two definitions of  $z$ . Either  $z = (a+b)/2$  or  $z = (b+c)/2$ .

**Case 1.**  $\frac{a+b}{2} = z < b$ . Thus,

$$\begin{aligned} \left| \log \frac{z}{b} \frac{1-b}{1-z} \right| &= \log \frac{b}{z} \frac{1-z}{1-b} \\ &= \log \frac{2b}{a+b} \frac{1-\frac{a+b}{2}}{1-b} \\ &= \log \frac{b}{a+b} \frac{2-a-b}{1-b} \\ &= \log \frac{b}{a+b} + \log \frac{\bar{a}+\bar{b}}{\bar{b}} \\ &\leq \log \frac{\bar{a}+\bar{b}}{\bar{b}} \quad \begin{bmatrix} \text{as first term} \\ \text{is negative} \end{bmatrix} \\ &= \log \left( 1 + \frac{\bar{a}}{\bar{b}} \right) \\ &\leq \log \left( 1 + \frac{\bar{a}}{(\bar{a}/3)} \right) = \log 4. \end{aligned}$$

□

**Case 2.**  $\frac{b+c}{2} = z > b$ . Thus,

$$\begin{aligned} \left| \log \frac{z}{b} \frac{1-b}{1-z} \right| &= \log \frac{z}{b} \frac{1-b}{1-z} \\ &= \log \frac{b+c}{2b} \frac{1-b}{1-\frac{b+c}{2}} \\ &= \log \frac{b+c}{b} \frac{1-b}{2-b-c} \\ &= \log \frac{b+c}{b} + \log \frac{\bar{b}}{\bar{b}+\bar{c}} \\ &\leq \log \frac{b+c}{b} \quad \begin{bmatrix} \text{as second term} \\ \text{is negative} \end{bmatrix} \\ &= \log \left( 1 + \frac{c}{b} \right) \\ &\leq \log \left( 1 + \frac{c}{(c/3)} \right) = \log 4. \end{aligned}$$

We are now in a position to prove that we never mistakenly eliminate  $p^*$  when comparing truncated representations. Without loss of generality, suppose that  $z > b$ , as the two cases are symmetric. Since we choose the left branch in the event of a tie, we have made a mistake if we instead needed to have chosen the right branch:  $p^* > z$ . In search of a contradiction, suppose that indeed this is the case. Under these conditions (i.e.,  $b < z < p^*$ ), and in light of [Claim 9.13.4](#), we can apply [Lemma 9.12](#) with  $k = \sqrt{2} > \ln 4$ , which tells us that

$$f(b) - f(z) > \frac{1}{16\gamma} \log^2 \left( 1 + 8\gamma(z-b)^2 \right).$$

The definition of  $\epsilon'$  in [\(9.9\)](#) is constructed precisely to make sure that this is never true. Therefore, the algorithm cannot choose the wrong branch.

**Complexity Analysis.** We now provide a more careful analysis of the runtime of the algorithm. We already have a bound on the number of iterations required; what is missing is a bound on how long it takes to compute  $\ll$ , i.e., to compare the approximations  $\tilde{f}(z)$  and  $\tilde{f}(b)$ . Assuming these numbers are in binary format,

$\tilde{f}(z)$  is of the form  $A.B$ , and  $\tilde{f}(b)$  is of the form  $A'.B'$ , where  $\{A, A', B, B'\}$  are binary sequences.

Without loss of generality, assume that  $|A| \leq |A'|$ . (Otherwise, swap their labels.) The complexity of comparing the two numbers  $\tilde{f}(z)$  and  $\tilde{f}(b)$  does not depend on  $|A'|$ , the longer of the two sequences to the left of the radix point. This is because once we see the radix point in one number but not the other, we can immediately conclude the former is smaller. In the first iteration,  $|A|$  is at most

$$\begin{aligned} |A| &\leq \max \left\{ 0, \log_2 \langle\!\langle \mathbf{m} + \Pr(Y=y)=\frac{1}{2} \rangle\!\rangle_\gamma \right\} \\ &\leq \max \left\{ 0, \log \left( \langle\!\langle \mathbf{m} \rangle\!\rangle_\gamma + D(p^* \parallel .5) \right) \right\} \\ &\leq \log_2 \left( \max\{0, \langle\!\langle \mathbf{m} \rangle\!\rangle_\gamma\} + 1 \right) \quad [ \text{since } D(p \parallel .5) \leq 1 ] \\ &\in O(\log \langle\!\langle \mathbf{m} \rangle\!\rangle_\gamma). \end{aligned}$$

Furthermore,  $|A|$  cannot increase as the algorithm progresses, because whichever of  $\{z, b\}$  leads to a smaller value of  $f$  becomes the new value of  $b$  in the following iteration.

Next, we derive an upper bound on the number of bits of  $B$  and  $B'$  that we must compare. Taking the (base-2) logarithm of (9.9), we find that we need to examine at most

$$\begin{aligned} |B| &\leq 4 + \log_2(\gamma) - 2 \log_2 \log (1 + 8\gamma(z - b)^2) \\ &\leq 4 + \log_2(\gamma) - 2 \log_2 \log \left( 1 + \frac{1}{2}\gamma\epsilon^2 \right) \quad [ \text{since } (z - b) \geq \frac{\epsilon}{4} ] \end{aligned}$$

bits to the right of the radix point in order to eliminate the possibility that  $f(b) < f(z)$ . This expression is still not very friendly; we now loosen it even further to provide a bound of a simpler, more recognizable form. When  $x \geq 0$ , we know that  $\log(1 + x) \geq 1 - \frac{1}{1+x} = \frac{x}{x+1}$ ; it follows that  $-\log_2 \log(1 + x) \leq$

$-\log_2\left(\frac{x}{x+1}\right) = \log_2(1 + \frac{1}{x})$ . Thus,

$$\begin{aligned} |B| + |A| &\leq 4 + \log_2(\gamma) + 2\log_2\left(1 + \frac{2}{\gamma\epsilon^2}\right) + \log_2(\langle\langle m\rangle\rangle_\gamma + 1) \\ &\in O\left(\log\frac{1}{\epsilon} + |\log\gamma| + \log\langle\langle m\rangle\rangle_\gamma\right). \end{aligned}$$

Recall that the process takes at most  $O(\log\frac{1}{\epsilon})$  iterations—but in the process, produces the same number of bits of output, since  $\log(1/\epsilon)$  is the number of bits need to encode the final approximation to  $p^*$ . So, accounting for the time needed to compare  $f(z)$  and  $f(b)$ , the algorithm runs in time

$$O\left(\log\frac{1}{\epsilon} \cdot \left(\log\frac{1}{\gamma} + \log\frac{1}{\epsilon} + \log\langle\langle m\rangle\rangle_\gamma\right)\right).$$

At this point, we have shown reduced unconditional inference to inconsistency calculation. To extend the reduction to conditional queries, we can apply Lemma 8.17 with  $k = 2$ ,  $K_0 = 0$ ,  $K_1 = \log\frac{1}{\gamma} + \log(1 + \langle\langle m\rangle\rangle_\gamma)$ ,  $K_2 = 1$ , and  $\Phi = 1$ , to get an algorithm that runs in

$$O\left(\left(\log\frac{1}{\gamma} + \log\langle\langle m\rangle\rangle_\gamma\right) \cdot \log\frac{1}{\epsilon\mu^*(X=x)} + \log^2\frac{1}{\epsilon\mu^*(X=x)}\right) \text{ time.}$$

It uses  $O(\log\log\frac{1}{\mu^*(X=x)} \cdot \log\frac{1}{\epsilon})$  calls to the inconsistency calculation procedure.

Finally, we remark that typically we are interested in doing inference up to floating point precision. In this case, by selecting  $\epsilon \leq 10^{-78}$ , the procedure above runs in constant time, making at most 1555 inconsistency procedure calls before outputting the 64-bit float that is closest to  $p^*$ .

**The other direction: reducing inconsistency calculation to inference.** This reduction is much simpler, shares more techniques with the primary thrust of the paper. First find a tree decomposition  $(\mathcal{C}, \mathcal{T})$  of the PDG's structure, and then query the marginals of each clique. Because of the work we've already done,

we know this information is enough information to simply evaluate the scoring function, including the joint entropy, by (8.12).

## CHAPTER 10

### REASONING WITH PDGS

We have just seen how we can reason about PDGs from the outside, but a lot of probabilistic reasoning can be done *internally* within PDGs. In [Chapter 6](#) we saw how many important inequalities in the literature are instances of what we called *monotonicity of inconsistency*: believing more things can only make you more inconsistent, not less.

Related monotonicity properties hold of many natural epistemic representations. One classical representation of knowledge is a list of formulas  $[\phi_1, \phi_2, \dots, \phi_n]$  that one knows to be true. This representation has an analogous property: learning an additional formula  $\phi_{n+1}$  can only narrow the set of worlds one considers possible. The same is true of both kinds of information in a PDG, and also of QIM-compatibility.

#### 10.1 Observational (Quantitative) Monotonicity and Equivalence

#### 10.2 Structural (Qualitative) Monotonicity and Equivalence

##### 10.2.1 Qualitative Monotonicity

In this section, we develop a related principle for QIM-compatibility.

Here is a direct analogue, that does not turn out to be terribly useful:

**Proposition 10.1.** *If  $\mathcal{A} \subseteq \mathcal{A}'$  and  $\mu \models \Diamond \mathcal{A}'$ , then  $\mu \models \Diamond \mathcal{A}$ .*

After all, if  $\mu$  is consistent with a set of independent causal mechanisms,

then surely it is consistent with a causal picture in which only a subset of those mechanisms are present and independent. There is a sense in which BNs and MRFs are also monotonic, but in the opposite direction: adding edges to a graph results in a weaker independence statement. We will soon see why.

Since we use *directed* hypergraphs, there is actually a finer notion of monotonicity at play. Inputs and outputs play opposite roles, and they are naturally monotonic in opposite directions. If there is an obvious way to regard an element of  $B$  as an element of  $B'$  (abbreviated  $B \hookrightarrow B'$ ), and  $A' \hookrightarrow A$ , then a function  $f : A \rightarrow B$  can be regarded as one of type  $A' \rightarrow B'$ . This is depicted to the right. The same principle applies in our setting. If  $X$  and  $Z$  are sets of variables and  $X \subseteq Z$ , then  $\mathcal{V}(Z) \hookrightarrow \mathcal{V}(X)$ , by restriction. It follows, for example, that any mechanism by which  $X$  determines  $(Y, Y')$  can be viewed as a mechanism by which  $(X, X')$  determines  $Y$ . The general phenomenon is captured by the following.

$$\begin{array}{ccc} A & \xrightarrow{f} & B \\ \uparrow & \swarrow & \downarrow \\ A' & \dashrightarrow & B' \end{array}$$

**Definition 10.1.** If  $\mathcal{A} = \{S \xrightarrow{a} T\}_a$ ,  $\mathcal{A}' = \{S' \xrightarrow{a'} T'\}_{a'}$ , and there is an injective map  $\iota : \mathcal{A}' \rightarrow \mathcal{A}$  such that  $T'_a \subseteq T_{\iota(a)}$  and  $S'_a \supseteq S_{\iota(a)}$  for all  $a \in \mathcal{A}'$ , then  $\mathcal{A}'$  is a *weakening* of  $\mathcal{A}$  (written  $\mathcal{A} \rightsquigarrow \mathcal{A}'$ ).  $\square$

**Theorem 10.2.** *If  $\mathcal{A} \rightsquigarrow \mathcal{A}'$  and  $\mu \models \Diamond \mathcal{A}$ , then  $\mu \models \Diamond \mathcal{A}'$ .*

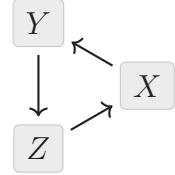
Theorem 10.2 is strictly stronger than Proposition 10.1 because a hyperarc with no targets is vacuous, so removing all targets of a hyperarc is equivalent to deleting it.

Theorem 10.2 explains why BNs and MRFs are arguably *anti-monotonic*: adding  $X \rightarrow Y$  to a graph  $G$  means adding  $X$  to the *sources* the hyperarc whose target is  $Y$ , in  $\mathcal{A}_G$ .

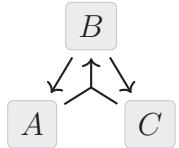
As mentioned in ??, the far more important consequence of this result is that it helps us begin to understand what QIM-compatibility means for cyclic hypergraphs. For the reader's convenience, we now restate the examples in the main text, which are really about monotonicity.

**Example 5.3.** Every  $\mu(X, Y)$  is compatible with  $X \rightleftarrows Y$ . This is because this cycle is weaker than a hypergraph that can already represent any distribution, i.e.,  $\rightarrow X \rightarrow Y \rightsquigarrow X \rightleftarrows Y$ .  $\triangle$ .

**Example 5.4.** What  $\mu(X, Y, Z)$  are compatible with the 3-cycle shown, on the right? By monotonicity, among them must be all distributions consistent with a linear chain  $\rightarrow X \rightarrow Y \rightarrow Z$ . Thus, any distribution in which two variables are conditionally independent given the third is compatible with the 3-cycle. Are there any distributions that are *not* compatible with this hypergraph? It is not obvious. We return to this in Section 5.4.  $\triangle$



Because QIM-compatibility applies to cyclic structures, one might wonder if it also captures the independencies of undirected models. Undirected edges  $A-B$  are commonly identified with a (cyclic) pair of directed edges  $\{A \rightarrow B, B \rightarrow A\}$ , as we have implicitly done in defining  $\mathcal{A}_G$ . In this way, undirected graphs, too, naturally correspond to directed hypergraphs. For example,  $G = A-B-C$  corresponds to the hypergraph  $\mathcal{A}_G$  shown on the left. Compatibility with  $\mathcal{A}_G$ , however, does not coincide with any of the standard Markov properties corresponding to  $G$  (?). This may appear to be a flaw in Definition 5.1 (QIM-compatibility), but it is unavoidable. While both BNs and MRFs are monotonic, it is impossible to capture both classes with a monotonic definition.



**Theorem 10.3.** *It is possible to define a relation  $\models^\bullet$  between distributions  $\mu$  and directed hypergraphs  $\mathcal{A}$  satisfying any two, but not all three, of the following.*

(monotonicity) *If  $\mu \models^\bullet \mathcal{A}$  and  $\mathcal{A} \rightsquigarrow \mathcal{A}'$ , then  $\mu \models^\bullet \mathcal{A}'$ .*

(positive BN capture) *If  $\mu$  satisfies the independencies  $\mathcal{I}(G)$  of a dag  $G$ , then  $\mu \models^\bullet \mathcal{A}_G$ .*

(negative MRF capture) *If  $\mu \models^\bullet \mathcal{A}_G$  for an undirected directed graph  $G$ , then  $\mu$  has one of the Markov properties with respect to  $G$ .*

The proof is a direct and easy-to-visualize application of monotonicity (Theorem 10.2). Assume monotonicity and positive BN capture. Let  $\mu_{xor}(A, B, C)$  be the joint distribution in which  $A$  and  $C$  are independent fair coins, and  $B = A \oplus C$  is their parity. We then have:

$$\mu_{xor} \models \begin{array}{c} B \\ \downarrow \quad \uparrow \\ A \quad C \end{array} \rightsquigarrow \begin{array}{c} B \\ \swarrow \quad \searrow \\ A \quad C \end{array} = \mathcal{A}_{A-B-C}.$$

But  $\mu_{xor} \not\models A \perp\!\!\!\perp C \mid B$ . □

We emphasize that Theorem 10.3 has implications for the qualitative semantics of *any* graphical model (even if one were to reject the definition QIM-compatibility). We reconcile Theorem 10.3 with PDGs and Dependency Networks (DNs), which may at first appear to defy the theorem.

**Probabilistic Dependency Graphs** PDGs capture BNs and MRFs Theorems 3.5 and 3.6, with a monotonic scoring function. But keep in mind that, while the independencies and observational information can be easily separated for a BN, the same is not true of a factor graph. Capturing a quantitative MRF is not the same as capturing its independencies. The independence property exploited by the PDG inference procedure Theorem 8.6 is the same as that of a MRF with the same structure. The subtlety here is that the independencies exploited for

PDG inference only need to be sound, not complete. That is, a PDG's underlying hypergraph  $\mathcal{A}$  may well imply more independencies than the inference procedure exploits.

**Dependency Networks** To readers familiar with *dependency networks* (DNs) (Heckerman et al. 2000), Theorem 10.3 may raise some conceptual issues. When  $G$  is an undirected graph,  $\mathcal{A}_G$  is the structure of a consistent DN. The semantics of such a DN, which intuitively describe an independent mechanism on each hyperarc, coincide with the MRFs for  $G$  (at least for positive distributions). In more detail, DN semantics are given by the fixed point of a markov chain that repeatedly generates independent samples along the hyperarcs of  $\mathcal{A}_G$  for some (typically cyclic) directed graph  $G$ . The precise definition requires an order in which to do sampling. Although this choice doesn't matter for the "consistent DNs" that represent MRFs, it does in general. With a fixed sampling order, the DN is monotonic and captures MRFs, but can represent only BNs for which that order is a topological sort.

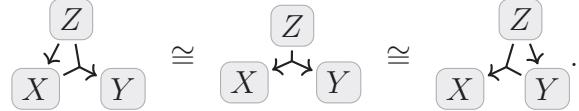
### 10.2.2 QIM Equivalence

We now begin to explore what it means for two hypergraphs to be *equivalent*, at least as far as QIM-compatibility (Definition 5.1) is concerned. An obvious candidate definition would be to call two hypergraphs  $\mathcal{A}_1$  and  $\mathcal{A}_2$  are QIM-equivalent if they are compatible with same distributions ( $\mu \models \Diamond \mathcal{A}_1$  iff  $\mu \models \Diamond \mathcal{A}_2$ ), but this isn't quite right. We have to deal with context. It is easy to see that every hypergraph consisting of just one hyperarc is QIM-compatible with all probability measures  $\mu \in \Delta \mathcal{V}\mathcal{X}$ , even though different hyperarcs intuitively mean

different things. Moreover, every hypergraph is a sum of one-arc hypergraphs, and we have already seen that not all hypergraphs are equivalent. To distinguish between hypergraphs that are not interchangable, we clearly need a stronger notion of equivalence.

Given hypergraphs  $\mathcal{A}_1$  and  $\mathcal{A}_2$ , we can form the combined hypergraph  $\mathcal{A}_1 + \mathcal{A}_2$  that consists of the disjoint union of the two sets of hyperarcs, and the union of their nodes. We say that  $\mathcal{A}$  and  $\mathcal{A}'$  are (*QIM*-)equivalent ( $\mathcal{A} \cong \mathcal{A}'$ ) if for every context  $\mathcal{A}''$  and distribution  $\mu$ , we have that  $\mu \models \mathcal{A} + \mathcal{A}''$  iff  $\mu \models \mathcal{A}' + \mathcal{A}''$ . By construction, structural equivalence ( $\cong$ ) is itself invariant to additional context: if  $\mathcal{A} \cong \mathcal{A}'$  then  $\mathcal{A} + \mathcal{A}'' \cong \mathcal{A}' + \mathcal{A}''$ . Our next result is a simple, intuitive, and particularly useful equivalence.

**Proposition 10.4.** *The following hypergraphs are equivalent:*



These three hypergraphs correspond, respectively, to equivalent factorizations of a conditional probability measure

$$P(X|Z)P(Y|X, Z) = P(X, Y|Z) = P(X|Y, Z)P(Y|Z).$$

Proposition 10.4 provides a simple and useful way to relate QIM-compatibility of different hypergraphs. If we restrict to acyclic structures, for instance, we find:

**Theorem 10.5** ([Chickering 2013](#)). *Any two qualitative Bayesian Networks that represent the same independencies can be proven equivalent using only instances of Proposition 10.4 (in which  $X, Y, Z$  may be sets of variables).*

Theorem 10.5 is essentially a restatement of main result of [Chickering \(2013\)](#),

but it is simpler to state in terms of directed hypergraph equivalences. To state the result in its original form, one has to first define an edge  $X \rightarrow Y$  to be *covered* in a graph  $G$  iff  $\text{Pa}_G(Y) = \text{Pa}_G(X) \cup \{X\}$ ; then, the result states that all equivalent BN structures are related by a chain of reversed covered edges. Observe that this notion of covering is implicit in [Theorem 10.5](#). [Theorem 10.5](#) is one demonstration of the usefulness of [Proposition 10.4](#), but the latter applies far more broadly, to cyclic structures and beyond. It becomes even more useful in tandem with the definition of monotonicity presented in ??, which is an analogue of implication.

### 10.3 A Logic Based on PDGs

So far in this chapter, we have treated the two aspects of PDGs separately, and developed separate reasoning principles for each aspect. We saw in [Section 10.1](#) an exploration of *observational* the proof language based on monotonicity that powered Then, in [Section 10.2.1](#) we saw a different monotonicity principle that applies to directed structural information. In this section, we put the two together. The result is an inconsistency-tolerant logic.

#### 10.3.1 A Natural Preorder on PDGs

**Definition 10.2.** We say that  $m_1$  entails  $m_2$  and write  $m_1 \models m_2$  iff

$$\forall \gamma \geq 0. \forall m'. \quad \langle\!\langle m_1 + m' \rangle\!\rangle_\gamma \geq \langle\!\langle m_2 + m' \rangle\!\rangle_\gamma.$$

□

Let's imagine that we can interact with a PDG only by altering it, and by measuring its degree of inconsistency.  $m_1 \models m_2$  means  $m_1$  is a “stronger

statement": in every context (so long as it's the same context added to  $m_1$  and  $m_2$ ),  $m_1$  has is at least as inconsistent as  $m_2$ . It may be helpful to think of the extremes: an outright contradiction, which has infinite inconsistency no matter the context, entails everything. Here are a few more examples.

**Proposition 10.6.** 1. Marginals are entailed by joint distributions, e.g.,  $p(X, Y) \models p(X)$ .

[ link to proof ]

2. Weaker constraints are entailed by larger ones, e.g.,  $(XY=xy) \models (X=x)$ .

Some more properties of the entailment relation, for general PDGs.

**Proposition 10.7.** 1. Reflexivity:  $m \models m$ .

[ link to proof ]

2. Transitivity:  $m_1 \models m_2$  and  $m_2 \models m_3$  imply that  $m_1 \models m_3$ .

3. If  $\mathcal{A}_1 \supseteq \mathcal{A}_2$  or if  $\beta_1 \geq \beta_2$  but otherwise the two are identical, then  $m_1 \models m_2$ .

4. The same is true if  $m_1$  and  $m_2$  are identical except for  $\alpha_1 \geq \alpha_2$ .<sup>1</sup>

If both  $m_1 \models m_2$  and  $m_2 \models m_1$ , then we say the two PDGs  $m_1$  and  $m_2$  are *equivalent*, and write  $m_1 \equiv m_2$ . Two PDGs are equivalent if they respond the same way to all probes  $m'$  and all choices of  $\gamma$ . Intuitively,  $m_1 \equiv m_2$  iff  $m_1$  and  $m_2$  are equally (in)consistent no matter how one alters them, so long as both are altered in the same way.

**Proposition 10.8.** 1. For all distributions  $p(X)$ ,  $p(X) + p(Y|X) \equiv p(X, Y)$ .

[ link to proof ]

2. If  $m_1 \equiv m_2$ , then their mixtures are also equivalent:  $m_1 \equiv (1-\alpha)m_1 + (\alpha)m_2 \equiv m_2$ .

3.  $m + m \equiv 2m$ . Compared to  $m$ , the PDG  $m + m$  contains two copies of each arc, while  $2m$  has all confidences doubled.

---

<sup>1</sup>So far, this result only holds for PDGs with discrete variables; It is not true for the approach I have been using to formalize *SDef* for continuous variables.

From its definition, it is clear that  $\equiv$  is symmetric. Thus transitivity and reflexivity of  $\models$  (above), imply that  $\equiv$  is an equivalence relation. Moreover, it is the same equivalence relation as equality of scoring function semantics!

**Theorem 10.9.**  $m_1 \equiv m_2$  iff  $\llbracket m_1 \rrbracket = \llbracket m_2 \rrbracket$ . That is,  $m_1 \equiv m_2$  if and only if  $m_1$  and  $m_2$  have the same set of variables  $\mathcal{X}$ , and  $\llbracket m_1 \rrbracket_\gamma(\mu) = \llbracket m_2 \rrbracket_\gamma(\mu)$  for all  $\gamma \geq 0$  and  $\mu \in \Delta \mathcal{V} \mathcal{X}$ .

[ link to proof ]

It is worth stressing an implication of this theorem: equivalent PDGs always have the same set of variables. In some ways this makes perfect sense: postulating the existence of  $X$  (or forgetting about  $X$ ) does in fact distinguish your mental state, even if you know nothing about  $X$ . However, it may also be too strict a requirement. When  $m_1 + \mathcal{X}_2 \equiv m_2 + \mathcal{X}_1$ , we call  $m_1$  and  $m_2$  *congruent modulo variables*, and write  $m_1 \cong_{\mathcal{X}} m_2$ . This definition places the two PDGs on the same set of variables before comparison. All PDGs with  $\alpha = \beta = 0$  are congruent to the empty PDG, modulo variables.

We can also add entailments together.

**Proposition 10.10.** If  $m_1 \models m_3$  and  $m_2 \models m_4$ , then  $m_1 + m_2 \models m_3 + m_4$ .

[ link to proof ]

*Proof.* Suppose  $m_1 \models m_3$ , and  $m_2 \models m_4$ . Then  $\forall m''$ , we have  $\langle\!\langle m_1 + m'' \rangle\!\rangle_\gamma \geq \langle\!\langle m_3 + m'' \rangle\!\rangle$ . In particular, for any  $m'$ , we can set  $m'' := m' + m_2$ , and thus

$$\langle\!\langle m_1 + m_2 + m' \rangle\!\rangle_\gamma \geq \langle\!\langle m_3 + m_2 + m' \rangle\!\rangle_\gamma \geq \langle\!\langle m_3 + m_4 + m' \rangle\!\rangle_\gamma. \quad \square$$



**Corollary 10.10.1.** If  $m_1 \models 0$  and  $m_2 \models 0$ , then  $m_1 + m_2 \models 0$ . This is because 0 is an idempotent PDG:  $0 \equiv 0 + 0$ .

Thus the PDGs that entail the trivial PDG are closed under addition. Do all PDGs have this property? As a point of comparison, in propositional logic, every proposition entails true. So, in this setting, is it the case that  $m \models 0$  for all  $m$ —that is, do we have  $\langle\langle m + m' \rangle\rangle_\gamma \geq \langle\langle m' \rangle\rangle_\gamma$  for all  $m$  and  $m'$ ? This is very close to a special case of the monotonicity of Lemma 6.1.

In general, the answer is no. Nevertheless, it does hold for PDGs  $m$  that have  $\beta \geq 0$  and are in a sense “qualitatively complete” which includes many PDGs of interest.

**Definition 10.3.** A weighted hypergraph  $(\mathcal{A}, \alpha)$  is *qualitatively complete* iff its structural deficiency is non-negative. More precisely:  $(\mathcal{A}, \alpha)$  is qualitatively complete iff for all interpretations  $\mathcal{X}$  of its nodes as variables, all and joint distributions  $\mu \in \Delta \mathcal{V} \mathcal{X}$ , the structural deficiency  $SDef_{(\mathcal{A}, \alpha)}(\mu) \geq 0$  of  $\mu$  with respect to  $(\mathcal{A}, \alpha)$  is non-negative.  $\square$

**Proposition 10.11.** Let  $m$  be a PDG with structure  $(\mathcal{A}, \alpha)$ , and observational confidences  $\beta \geq 0$ . Then the following are equivalent:

1.  $(\mathcal{A}, \alpha)$  is qualitatively complete.
2.  $m \models 0$ .
3.  $\alpha$  can be written as a mixture of BNs plus a positive residual. That is, there exist BN structures  $B_1, \dots, B_n$  (which correspond to vectors  $\alpha_1, \dots, \alpha_n$ ), mixture coefficients  $w_1, \dots, w_n \in [0, 1]$  with  $\sum_{i=1}^n w_i = 1$ , and a vector  $\alpha' \geq 0$  such that  $\alpha = \alpha' + \sum_{i=1}^n w_i \alpha_i$ .

We can scale a PDG  $m$  by  $r \in \mathbb{R}$ , by multiplying all of its constituent confidences  $\alpha$  and  $\beta$  by  $r$ . Concretely, this means  $m' = (r)m$  has confidences

$\beta'_a = r \cdot \beta_a$  and  $\alpha'_a = r \cdot \alpha_a$ , where  $\beta_a$  and  $\alpha_a$  are the confidences of arc  $a$  in  $m$ .

Can we get a result similar to [Proposition 10.10](#) for scalar multiplication? No, we cannot.

**Falsity 10.12.** *If  $m_1 \models m_2$  then  $(r)m_1 \models (r)m_2$ , for all  $r \geq 0$ .*

For when  $r = 0$ , this means But here are some variants that might be true:

**Conjecture 10.13.** 1. *If  $\mathcal{X}_1 \subseteq \mathcal{X}_2$  and  $m_1 \models m_2$ , then  $(r)m_1 \models (r)m_2$  for all  $r \geq 0$ .*

2. *If  $m_1 \models m_2$  then  $(r)m_1 \models (r)m_2$  for all  $r \geq 1$ .*

While new beliefs cannot make you any less inconsistent, new variables are additional degrees of freedom, and cannot make you any more inconsistent.

**Proposition 10.14.** *If  $\mathcal{X}_1 \subseteq \mathcal{X}_2$ , then  $\mathcal{X}_1 \models \mathcal{X}_2$ .*

It follows, for example, that the empty PDG entails any PDG that only contains variables, i.e.,  $0 \models \mathcal{X}$  for any set of variables  $\mathcal{X}$ . (It might be worth noting that, as PDGs,  $\mathcal{X}_1 + \mathcal{X}_2 = \mathcal{X}_1 - \mathcal{X}_2 = \mathcal{X}_1 \cup \mathcal{X}_2$  are all the same.)

The bottom (or initial) element(s) of the entailment order are easily identified; like in propositional logic, they are the (most extreme) contradictions contradictions. Every such PDG is equivalent to  $\text{False} := \frac{\delta_0}{\delta_1} \Rightarrow [X] \Leftarrow$ , which has  $\langle\langle \text{False} \rangle\rangle = \infty$ , and satisfies  $\text{False} \models m$  for all  $m$ .

Are there largest PDGs in this order? What is the analogue of “true”?—is it the empty PDG? Clearly not, because  $m \models 0$  only if  $m$  is qualitatively complete. It turns out there is such a “most annodyne” PDG if and only if the collection of

all possible variables is a PDG. Concretely, let  $\mathcal{U}$  be a “universe of all possible variables”. Then every PDG  $m$  with variables  $\mathcal{X} \subseteq \mathcal{U}$  satisfies  $m \models \mathcal{U}$ .

### 10.3.2 “Propositional” PDG Logic

Starting with the entailment operation and introducing connectives, we can obtain a propositional logic whose atomic formulas are PDGs. We adopt the usual convention of writing “ $\models \varphi$ ” if  $m \models \varphi$  for all PDGs  $m$ .

**Exploring the Set-of-Distribution Semantics.** Because  $\{\cdot\}$  is a feature of just the observational/quantitative half of a PDG, let’s ignore *SDef* for now, and assume  $\gamma = 0$ .

Taking only the binary aspect of this logic into account, we might want it to be the case that if  $\{m_1\} \subseteq \{m_2\}$ , then  $m_1 \models m_2$ . This is not true,<sup>2</sup> but it is true, in this case, that  $(\infty)m_1 \models (\infty)m_2$ . Similarly, if

### Defining Classical Connectives

As usual, we can introduce conjunction by defining  $m \models \varphi \wedge \varphi'$  iff  $m \models \varphi$  and  $m \models \varphi'$ . So, for example,  $m_1 \models (m_2 \Rightarrow m_3)$  states that if  $m_1 \models m_2$ , then  $m_1 \models m_3$ . This lets us prove things like

$$\models p(X, Y) \Leftrightarrow p(X) + p(Y|X) \quad \text{and} \quad \models (p(X) + q(X)) \Rightarrow p(X).$$

We can also continue with  $\neg$  and then define other connectives ( $\vee, \Rightarrow, \Leftrightarrow$ ) in terms of  $(\wedge, \neg)$ . But the usual definition of  $m \models \neg\varphi$  iff  $m \not\models \varphi$ , may not be so well-behaved. For instance, ...

---

<sup>2</sup>For instance,  $\{m\} = \{(2)m\}$ , but  $m \not\models (2)m$ .

Often outside of classical logic, it is more natural to start with implication instead of negation. Here is an alternate conception of implication that more closely follows the intuition we had for entailment. Define

$$m \models \varphi \stackrel{\bullet}{\Rightarrow} \varphi' \quad \text{if and only if} \quad m + \varphi \models m + \varphi'.$$

In words: a formula  $\varphi$  implies another formula  $\varphi'$  in context  $m$  iff  $\varphi$  with context  $m$  entails  $\varphi'$  with context  $m$ . One nice property of this definition:  $m_1 \models m_2$  becomes equivalent to  $\models m_1 \stackrel{\bullet}{\Rightarrow} m_2$ . This definition has a very significant drawback, however: it only makes sense if  $\varphi$  and  $\varphi'$  are themselves PDGs (or at least, can be added to  $m$  to form a new PDG). This would rule out “synthetic” logical connectives such as  $\wedge$  and  $\neg$  inside an implication. Thus, the logic inside a connective  $\stackrel{\bullet}{\Rightarrow}$  collapses to ordinary constructions that can be done with PDGs themselves.

PDGs also have their own operations, most notably  $+$  and  $\sqcup$ . One interesting question is how these interact with the logical connectives we have just defined. Here are some interactions between connectives  $\wedge$ ,  $\vee$ ,  $\neg$ , and PDG operations:

**Conjecture 10.15.** 1. If  $m \models 0$ , and  $m \models m_1 + m_2$ , then  $m \models m_1 \wedge m_2$ .

(unproven!)

2.  $m_1 + m_2 \models m_1 \sqcup m_2$

3. For all  $\alpha \in [0, 1]$ , we have:  $\models (m_1 \wedge m_2) \Rightarrow ((1 - \alpha)m_1 + (\alpha)m_2)$ .



⟨ INCOMPLETE ⟩

### 10.3.3 Epistemic Logic

Given a PPDG  $\mathbf{m}(\Theta)$ , and a set  $I$  of agents, each of which has an attention mask  $Attn_i(\theta) : 2\mathcal{A} \rightarrow \mathbb{R}$ , we can define the local state of agent  $i$  at  $\theta$  to be  $\mathbf{m}_i(\theta) := Attn_i(\theta) \odot \mathbf{m}(\theta)$ . If it's always the case that  $\mathbf{m}_i(\theta) \equiv \mathbf{m}(\theta)$  (which happens iff  $Attn_i(\theta)(a) = 1$ ) we say that  $i$  has uniform attention over  $\mathbf{m}$ . We can form a Kripke structure with states  $\Theta$ , and accessibility relation  $\theta \sim_i \theta'$  iff  $\mathbf{m}_i(\theta) \equiv \mathbf{m}_i(\theta')$ . Because  $\equiv$  is an equivalence relation, the resulting logic satisfies S5. We can also take primitive propositions  $\Phi$  to be the set of PDGs, using the entailment relation.

Some basic properties of this logic:

- Proposition 10.16.**
1.  $(\mathbf{m}, \theta, I) \models \mathbf{m}(\theta)$ .
  2. For all agents  $i \in I$ ,  $(\mathbf{m}, I) \models K_i(\mathbf{m}_i(\theta))$ , i.e., each agent knows its local state.
  3. In particular, if agent  $i$  has uniform global attention over  $\mathbf{m}$ , then  $(\mathbf{m}, \theta, I) \models K_i(\mathbf{m}(\theta))$ .
  4. Introducing variables, have  $(\mathbf{m}, I) \models K_i(\mathbf{m} \& \Theta)$ .
  5. Moreover, the PPDG is common knowledge:  $(\mathbf{m}, I) \models C_I(\mathbf{m} \& \Theta)$ .

Now some other examples.

**Example 10.1.** There are two agents,  $I = \{1, 2\}$ . Each has its own belief about a variable  $X$ , and can only see that belief. We can't say anything interesting here, because the beliefs of the agents don't interact.

Now, suppose both agents observe  $X = x$ . That is,  $\mathbf{m}_1 = \{p, x\}$  and  $\mathbf{m}_2 = \{q, x\}$ . Then  $(\mathbf{m}, I) \models C_{\{1, 2\}}(X = x)$ .

Also,  $(m, I) \models C_{\{1,2\}}((m_1 - m_2) \vee (m_2 - m_1))$ . In English: it's common knowledge that one of the two is more inconsistent than the other.  $\triangle$

## APPENDICES FOR CHAPTER 10

### 10.A Proofs

**Lemma 10.17.**  $\llbracket \mathbf{m}_1 + \mathbf{m}_2 \rrbracket_\gamma = \llbracket \mathbf{m}_1 \rrbracket_\gamma(\mathcal{X}_1) + \llbracket \mathbf{m}_2 \rrbracket_\gamma(\mathcal{X}_2) + \gamma I_\mu(\mathcal{X}_1; \mathcal{X}_2 | \mathcal{X}_1 \cap \mathcal{X}_2)$ .

*Proof.* See appendix of inference paper.  $\square$

**Definition 10.4.** If  $\mathbf{m}_1 + \mathbf{m}_2 \equiv \mathbf{m}_1 \sqcup \mathbf{m}_2$ , then we say  $\mathbf{m}_1$  and  $\mathbf{m}_2$  are independent.  $\square$

**Lemma 10.18.** If  $\mathbf{m}_1$  and  $\mathbf{m}_2$  have disjoint variables  $\mathcal{X}_1 \cap \mathcal{X}_2 = \emptyset$ , then  $\langle\!\langle \mathbf{m}_1 + \mathbf{m}_2 \rangle\!\rangle = \langle\!\langle \mathbf{m}_1 \rangle\!\rangle + \langle\!\langle \mathbf{m}_2 \rangle\!\rangle$ .

**Lemma 10.19.** If  $\mathbf{m}_1$  and  $\mathbf{m}_2$  are such that, for all pairs of distributions  $\mu_1 \in \Delta V \mathcal{X}_1$  and  $\mu_2 \in \Delta V \mathcal{X}_2$  that have the same marginal on their common variables  $\mathcal{X}_1 \cap \mathcal{X}_2$ , it is the case that

$$OInc_{\mathbf{m}_1}(\mu_1) \geq OInc_{\mathbf{m}_2}(\mu_2) \quad \text{and} \quad SDef_{\mathbf{m}_1}(\mu_1) \geq SDef_{\mathbf{m}_2}(\mu_2),$$

then  $\mathbf{m}_1 \models \mathbf{m}_2$ .

*Proof.* Suppose  $\mathbf{m}_1$  and  $\mathbf{m}_2$  satisfy this property. Then clearly  $\llbracket \mathbf{m}_1 \rrbracket_\gamma(\mu_1) \geq \llbracket \mathbf{m}_2 \rrbracket_\gamma(\mu_2)$ , as

$$\llbracket \mathbf{m}_1 \rrbracket_\gamma(\mu_1) = OInc_{\mathbf{m}_1}(\mu_1) + SDef_{\mathbf{m}_1}(\mu_1) \geq OInc_{\mathbf{m}_2}(\mu_2) + SDef_{\mathbf{m}_2}(\mu_2) = \llbracket \mathbf{m}_2 \rrbracket_\gamma(\mu_2).$$

Let  $\mathbf{m}'$  be an arbitrary PDG, with variables  $\mathcal{X}'$ . With some calculation, we find

that

$$\begin{aligned}
\langle\langle \mathbf{m}_1 + \mathbf{m}' \rangle\rangle_\gamma &= \inf_{\mu \in \Delta V(\mathcal{X}_1 \cup \mathcal{X}')} \left( [\mathbf{m}]_\gamma(\mu(\mathcal{X})) + [\mathbf{m}']_\gamma(\mu(\mathcal{X}')) + \gamma I_\mu(\mathcal{X}_1; \mathcal{X}' | \mathcal{X}_1 \cap \mathcal{X}') \right) \\
&= \inf_{\mu \in \Delta V(\mathcal{X}_1 \cup \mathcal{X}' \cup \mathcal{X}_2)} \left( [\mathbf{m}_1]_\gamma(\mu(\mathcal{X}_1)) + [\mathbf{m}']_\gamma(\mu(\mathcal{X}')) \right) \\
&\geq \inf_{\mu \in \Delta V(\mathcal{X}_1 \cup \mathcal{X}' \cup \mathcal{X}_2)} \left( [\mathbf{m}_2]_\gamma(\mu(\mathcal{X}_2)) + [\mathbf{m}']_\gamma(\mu(\mathcal{X}')) \right) \\
&= \inf_{\mu \in \Delta V(\mathcal{X}_2 \cup \mathcal{X}')} \left( [\mathbf{m}_2]_\gamma(\mu(\mathcal{X}_2)) + [\mathbf{m}']_\gamma(\mu(\mathcal{X}')) + \gamma I_\mu(\mathcal{X}_2; \mathcal{X}' | \mathcal{X}_2 \cap \mathcal{X}') \right) \\
&= \inf_{\mu \in \Delta V(\mathcal{X}_2 \cup \mathcal{X}')} [\mathbf{m}_2 + \mathbf{m}']_\gamma(\mu) \\
&= \langle\langle \mathbf{m}_2 + \mathbf{m}' \rangle\rangle_\gamma.
\end{aligned}$$

We can freely remove the term  $I_\mu(\mathcal{X}; \mathcal{X}' | \mathcal{X} \cap \mathcal{X}')$  inside the infimum (line 1), because this quantity is minimized when  $\mathcal{X}$  and  $\mathcal{X}'$  are conditionally independent given their common variables, and we can always find a distribution with this property that has the same marginals on  $\mathcal{X}$  and  $\mathcal{X}'$ .  $\square$

### Proposition 10.6.

1. Marginals are entailed by joint distributions, e.g.,  $p(X, Y) \models p(X)$ .
2. Weaker constraints are entailed by larger ones, e.g.,  $(XY = xy) \models (X = x)$ .

*Proof.* **Marginals.** Implicitly, we have converted  $p(X, Y)$  and  $p(X)$  to PDGs by assigning them a default value of  $\alpha = 1$ . One can easily calculate:

$$\begin{aligned}
&OInc_{p(X, Y)}(\mu(X, Y)) && SDef_{\rightarrow XY}(\mu(X, Y)) \\
&= D(\mu(X, Y) \parallel p(X, Y)) && = 0 \\
&= D(\mu(X) \parallel p(X)) + D(\mu(Y|X) \parallel p(Y|X) \mid \mu(X)) && = SDef_{\rightarrow X}(\mu(X)) \\
&\geq D(\mu(X) \parallel p(X)) = OInc_{p(X)}(\mu(X))
\end{aligned}$$

So, by Lemma 10.19, we have  $p(X, Y) \models p(X)$ .

**Hard Constraints.** First, observe that

$$OInc_{XY=xy}(\mu(XY)) = \infty \mathbb{1}[\mu(XY=xy) < 1] \geq \infty \mathbb{1}[\mu(X=x) < 1] = OInc_{X=x}(\mu(X))$$

In other words, the only distribution attaining a finite  $OInc$  for  $XY = xy$  is the point mass  $\delta_{xy}$ ; thus this is the only candidate distribution  $\mu$  which  $\llbracket X = x \rrbracket_\gamma(\mu)$  could be larger. But this distribution has no entropy, thus  $\llbracket XY = xy \rrbracket_\gamma(\delta_{xy}) = 0 = \llbracket XY = xy \rrbracket_\gamma(\delta_{xy})$ . Thus,  $\llbracket XY = xy \rrbracket_\gamma(\mu) \geq \llbracket X = x \rrbracket_\gamma(\mu)$  for all distributions  $\mu$ , showing  $XY = xy \models X = x$ .

□

### Proposition 10.7.

1. *Reflexivity:*  $m \models m$ .
2. *Transitivity:*  $m_1 \models m_2$  and  $m_2 \models m_3$  imply that  $m_1 \models m_3$ .
3. If  $\mathcal{A}_1 \supseteq \mathcal{A}_2$  or if  $\beta_1 \geq \beta_2$  but otherwise the two are identical, then  $m_1 \models m_2$ .
4. The same is true if  $m_1$  and  $m_2$  are identical except for  $\alpha_1 \geq \alpha_2$ .<sup>3</sup>

*Proof.* **Reflexivity and Transitivity.** Reflexivity is immediate. For transitivity, we have assumed  $m_1 \models m_2$  and  $m_2 \models m_3$ . By definition, we have that for all  $m'$ , we have

$$\langle\!\langle m_1 + m' \rangle\!\rangle \geq \langle\!\langle m_2 + m' \rangle\!\rangle \geq \langle\!\langle m_3 + m' \rangle\!\rangle.$$

**Domination.** If  $m_1$  and  $m_2$  share variables but  $m_1$  has more edges or higher weights, the condition of Lemma 10.19 is satisfied, and so  $m_1 \models m_2$ . □

---

<sup>3</sup>So far, this result only holds for PDGs with discrete variables; It is not true true for the approach I have been using to formalize *SDef* for continuous variables.

**Proposition 10.8.**

1. For all distributions  $p(X)$ ,  $p(X) + p(Y|X) \equiv p(X, Y)$ .
2. If  $\mathbf{m}_1 \equiv \mathbf{m}_2$ , then their mixtures are also equivalent:  $\mathbf{m}_1 \equiv (1-\alpha)\mathbf{m}_1 + (\alpha)\mathbf{m}_2 \equiv \mathbf{m}_2$ .
3.  $\mathbf{m} + \mathbf{m} \equiv 2\mathbf{m}$ . Compared to  $\mathbf{m}$ , the PDG  $\mathbf{m} + \mathbf{m}$  contains two copies of each arc, while  $2\mathbf{m}$  has all confidences doubled.

*Proof.* **Equivalence of conditional/marginal and Joint.**

$$\begin{aligned}
 & OInc_{[p(X)+p(Y|X)]}(\mu) && SDef_{\rightarrow X \rightarrow Y}(\mu) \\
 &= \mathbb{E}_{\mu} \left[ \log \frac{\mu(X)}{p(X)} + \log \frac{\mu(Y|X)}{p(Y|X)} \right] && = H_{\mu}(X) + H_{\mu}(Y|X) - H(\mu) \\
 &= \mathbb{E}_{\mu} \left[ \log \frac{\mu(X)\mu(Y|X)}{p(X)p(Y|X)} \right] && = H_{\mu}(X, Y) - H(\mu) \\
 &= OInc_{p(X,Y)}(\mu); && = SDef_{\rightarrow XY}(\mu).
 \end{aligned}$$

Thus, by applying Lemma 10.19 once in each direction, or by appeal to Theorem 10.9,  $p(X) + p(Y|X) \equiv p(X, Y)$ .

**Mixtures.** Let  $H_{\mu}(\mathbf{m}) := \sum_{a \in \mathcal{A}^m} \alpha_a^m H_{\mu}(T_a | S_a)$ , so that  $SDef_{\mathbf{m}}(\mu) = H_{\mu}(\mathbf{m}) - H_{\mu}(\mathcal{X})$ . Suppose  $\mathbf{m}_1 \equiv \mathbf{m}_2$ . Then, the theorem above tells us

$$\begin{aligned}
 & \llbracket \mathbf{m}_1 \rrbracket_{\gamma}(\mu) = \llbracket \mathbf{m}_2 \rrbracket_{\gamma}(\mu) \\
 \iff & OInc_{\mathbf{m}_1}(\mu) + \gamma SDef_{\mathbf{m}_1}(\mu) = OInc_{\mathbf{m}_2}(\mu) + \gamma SDef_{\mathbf{m}_2}(\mu) \\
 \iff & OInc_{\mathbf{m}_1}(\mu) + \gamma H(\mathbf{m}_1) - \gamma H(\mu) = OInc_{\mathbf{m}_2}(\mu) + \gamma H(\mathbf{m}_2) - \gamma H(\mu) \\
 \iff & OInc_{\mathbf{m}_1}(\mu) + \gamma H(\mathbf{m}_1) = OInc_{\mathbf{m}_2}(\mu) + \gamma H(\mathbf{m}_2)
 \end{aligned}$$

But  $\mathbf{m} \mapsto OInc_m(\mu)$  and  $\mathbf{m} \mapsto H_\mu(\mathbf{m})$  are both linear in  $(\alpha, \beta)$ . Thus, for  $\alpha \in [0, 1]$ , letting  $\mathbf{m}_\alpha := (1 - \alpha)\mathbf{m}_1 + (\alpha)\mathbf{m}_2$ , we have:

$$\begin{aligned} & OInc_{m_\alpha}(\mu) + \gamma H_\mu(\mathbf{m}_\alpha) \\ &= (1 - \alpha)OInc_{m_1} + \alpha OInc_{m_2} + \gamma(1 - \alpha) H_\mu(\mathbf{m}_1) + \gamma\alpha H_\mu(\mathbf{m}_2) \\ &= (1 - \alpha)(OInc_{m_1} + \gamma H_\mu(\mathbf{m}_1)) + \alpha(OInc_{m_2} + \gamma H_\mu(\mathbf{m}_2)) \\ &= OInc_{m_1} + \gamma H_\mu(\mathbf{m}_1). \end{aligned}$$

The last line follows from the equality of the two parenthesized expressions (as shown in the previous block of algebra). Thus following the first argument in reverse, we learn that  $\llbracket \mathbf{m}_\alpha \rrbracket = \llbracket \mathbf{m}_1 \rrbracket = \llbracket \mathbf{m}_2 \rrbracket$ , and so by the previous Theorem,  $\mathbf{m}_\alpha \equiv \mathbf{m}_1 \equiv \mathbf{m}_2$ .  $\square$

**Theorem 10.9.**  $\mathbf{m}_1 \equiv \mathbf{m}_2$  iff  $\llbracket \mathbf{m}_1 \rrbracket = \llbracket \mathbf{m}_2 \rrbracket$ . That is,  $\mathbf{m}_1 \equiv \mathbf{m}_2$  if and only if  $\mathbf{m}_1$  and  $\mathbf{m}_2$  have the same set of variables  $\mathcal{X}$ , and  $\llbracket \mathbf{m}_1 \rrbracket_\gamma(\mu) = \llbracket \mathbf{m}_2 \rrbracket_\gamma(\mu)$  for all  $\gamma \geq 0$  and  $\mu \in \Delta \mathcal{V} \mathcal{X}$ .

*Proof.* ( $\implies$ ). Suppose  $\mathbf{m}_1 \equiv \mathbf{m}_2$ . Let  $\mathcal{X}_1$  be the variables of  $\mathbf{m}_1$ , and  $\mathcal{X}_2$  be the variables of  $\mathbf{m}_2$ . We begin by showing that  $\mathcal{X}_1 = \mathcal{X}_2$  (up to constant variables, that can take only one value). In search of a contradiction, suppose there is a variable  $X$  be a variable present in one PDG but not the other; without loss of generality, suppose  $X \in \mathcal{X}_1 \setminus \mathcal{X}_2$ . Then in particular, we must have  $\langle\!\langle \mathbf{m}_1 \rangle\!\rangle = \langle\!\langle \mathbf{m}_2 \rangle\!\rangle$ , and also  $\langle\!\langle \mathbf{m}_1 + X \rangle\!\rangle = \langle\!\langle \mathbf{m}_2 + X \rangle\!\rangle$ . But  $X \in \mathcal{X}_1$ , so  $\mathbf{m}_1 + X = \mathbf{m}_1$ . Therefore,

$$\langle\!\langle \mathbf{m}_2 \rangle\!\rangle = \langle\!\langle \mathbf{m}_1 \rangle\!\rangle = \langle\!\langle \mathbf{m}_1 + X \rangle\!\rangle = \langle\!\langle \mathbf{m}_2 + X \rangle\!\rangle.$$

Since  $X \notin \mathcal{X}_2$ , the minimizer of  $\llbracket \mathbf{m}_2 + X \rrbracket_\gamma(\mu(X, \mathcal{X}_2))$  is realized when  $X$  and  $\mathcal{X}_2$  are independent (Richardson et al. 2023, Theorem 5). This means that

$\langle\!\langle \mathbf{m}_2 + X \rangle\!\rangle_\gamma = \langle\!\langle \mathbf{m}_2 \rangle\!\rangle_\gamma - \gamma H(\text{Unif}(X))$ . By selecting  $\gamma > 0$ , we discover that find a contradiction, unless  $X$  can only take one value.

So far we have shown that  $\mathbf{m}_1$  and  $\mathbf{m}_2$  have the same set of (nontrivial) variables; we now show that they have the same scoring function. For any  $\gamma \geq 0$  and  $\mu \in \Delta \mathcal{V}\mathcal{X}$ , the definition of  $\equiv$  with a choice of  $\mathbf{m}' = \mu!$  gives us the middle equality in the following.

$$[\![\mathbf{m}_1]\!]_\gamma(\mu) = \langle\!\langle \mathbf{m}_1 + \mu! \rangle\!\rangle_\gamma = \langle\!\langle \mathbf{m}_2 + \mu! \rangle\!\rangle_\gamma = [\![\mathbf{m}_2]\!]_\gamma(\mu).$$

( $\Leftarrow$ ). Suppose  $[\![\mathbf{m}_1]\!] = [\![\mathbf{m}_2]\!]$ . In particular, for the two objects to even have the same type, this means  $\mathbf{m}_1$  and  $\mathbf{m}_2$  have the same set  $\mathcal{X}$  of variables. For every  $\mathbf{m}'$  and  $\gamma \geq 0$ , we have

$$\begin{aligned} \langle\!\langle \mathbf{m}_1 + \mathbf{m}' \rangle\!\rangle_\gamma &= \inf_{\mu \in \Delta \mathcal{V}(\mathcal{X} + \mathcal{X}')} ([\![\mathbf{m}_1]\!]_\gamma(\mathcal{X}) + [\![\mathbf{m}']\!]_\gamma(\mathcal{X}')) \\ &= \inf_{\mu \in \Delta \mathcal{V}(\mathcal{X} + \mathcal{X}')} ([\![\mathbf{m}_2]\!]_\gamma(\mathcal{X}) + [\![\mathbf{m}']\!]_\gamma(\mathcal{X}')) = \langle\!\langle \mathbf{m}_2 + \mathbf{m}' \rangle\!\rangle_\gamma. \end{aligned} \quad \square$$

## 10.B Negative Results and Anti-Conjectures

**Falsity 10.20.** 1. If  $\mathcal{B}$  is the PDG form of a BN, which determines distribution  $\text{Pr}_{\mathcal{B}}$ , then  $\mathcal{B} \models \text{Pr}_{\mathcal{B}}$ . Moreover,  $\text{Pr}_{\mathcal{B}} \models \mathcal{B}$ , and thus  $\mathcal{B} \equiv \text{Pr}_{\mathcal{B}}$ .

1. cannot be true because by selecting  $\text{Pr}_{\mathcal{B}}$  we have lost the ability to control the relative weight of independence and cpds; it's all mashed together, and this will only hold at  $\gamma = 1$ .

**Falsity 10.21.** If  $\mathbf{m} \models 0$ , then  $(\alpha)\mathbf{m} \models 0$  for all  $\alpha \geq 0$ .

This can't be true; for  $\alpha = 0$ , it reduces to  $\mathcal{X}^m \models 0$ , which is false.

Corollary: the following is false.

**Falsity 10.22.** *If  $m_1 \models m_2$ , then  $(r)m_1 \models (r)m_2$ , for all  $r \geq 0$ .*

Choosing  $m_2 = 0$  would imply the previous false result.

We can show that  $\rightarrow X \models 0$ , but not necessarily the reverse. If  $m'$  does not include  $X$ , then  $\langle\!\langle m' + \rightarrow X \rangle\!\rangle_\gamma = \langle\!\langle m' \rangle\!\rangle$ . But what if  $m'$  includes  $X$ ? Then  $\langle\!\langle m' + \rightarrow X \rangle\!\rangle \geq \langle\!\langle m' \rangle\!\rangle$  by domination, and the inequality is strict for “most” choices of  $m'$ . So either way  $\langle\!\langle \rightarrow X + m' \rangle\!\rangle \geq \langle\!\langle 0 + m' \rangle\!\rangle$ , so  $\rightarrow X \models 0$ .

## **Part IV**

# **Foundations**

## CHAPTER 11

### LEARNER'S CONFIDENCE

When we say  $\beta$  is a *confidence*, what exactly do we mean? As we will explore in this chapter, we do not mean something probabilistic— $\beta_a$  is not the probability that the cpd  $\mathbb{P}_a$  is correct, but rather a notion of trust. In this chapter, we illustrate the difference between these concepts, and in so doing, develop a conception of *learner's confidence*, that unifies a number of related concepts in the literature. We will see how a continuum of confidence has two equivalent canonical representations: a multiplicative one that looks on the surface like probability, and an additive one that looks like other standard units. Moreover, the framework gives us a generic way to orderlessly combine information in any situation where the concept of confidence applies.

Finally, we will see how a particularly nice class of learning procedures, is parameterized by a belief space, an observation space, and a loss function. Much of this dissertation has been devoted to establishing PDGs as a universal belief representation, that comes with a particularly natural choice of ??, a direction we will explore in Chapter 12. But before we get ahead of ourselves, we must first develop this general theory of confidence.

#### 11.1 Introduction To Learner's Confidence

What does it mean to have a high degree of confidence in a statement  $\phi$ ? It is often taken to mean that  $\phi$  is likely. We argue that there is a related conception of confidence that arises when learning—one that complements likelihood and, moreover, unifies several different concepts in the literature. This kind of confidence is a measure of *trust*, rather than likelihood. The degree of confidence one

places in a piece of information  $\phi$  quantifies how seriously to take  $\phi$  in updating one's beliefs. So at one extreme, if we observe  $\phi$  but have no confidence in it, we should not change our beliefs at all; at the other, if we have full confidence in  $\phi$ , we should fully incorporate it into our beliefs.

**Example 11.1.** Suppose our prior belief state is a probability measure  $\text{Pr}$ , and  $\phi$  is an event. A full-confidence update then amounts to conditioning on  $\phi$  (i.e., adopting the belief state  $\text{Pr} | \phi$ ), after which  $\phi$  has probability 1 and cannot be further incorporated. Here is one obvious way to describe intermediate degrees of confidence: if we learn  $\phi$  with confidence  $\alpha \in [0, 1]$  and start with prior  $\text{Pr}$ , then we end up with the posterior  $(1 - \alpha) \text{Pr} + \alpha(\text{Pr} | \phi)$ . Thus, having high confidence in  $\phi$  leads to posterior beliefs that give  $\phi$  high probability. The converse is false, however, so confidence and probability can be quite different. If an untrusted source tells us  $\phi$  which we already happen to believe, then our prior assigns  $\phi$  high probability, we learn  $\phi$  with low confidence, and our posterior beliefs still give  $\phi$  high probability.  $\triangle$

Confidence allows us to be uncertain about observations, which is quite different in principle from making observations that are uncertain. *Jeffrey's rule* (Jeffrey 1968) (see Section 11.A.1) is a well-established approach to the latter. An important feature of the former, however, is that it enables learning without fully committing to new observations. Fully certain updates, such as conditioning in Example 11.1, are irreversible: from  $\phi$  and the posterior  $\text{Pr} | \phi$ , it is not possible to recover the prior  $\text{Pr}$ . The same is true for Jeffrey's rule, which, in our view, also prescribes full-confidence updates. The concept we propose here is much closer to the focus of Shafer's *Theory of Evidence* Shafer (1976), although his account is heavily specialized to a specific representation of uncertainty (Dempster-Shafer

belief functions) that have largely fallen out of fashion.

**Example 11.2.** Suppose our belief state is a *belief function*, a generalization of a probability measure over a finite set  $W$  of possible worlds. Like a probability, a belief function  $Bel$  assigns to each event  $U \subseteq W$  a number  $Bel(U) \in [0, 1]$ , with  $Bel(\emptyset) = 0$  and  $Bel(W) = 1$ . It is not necessarily the case that  $Bel(U) + Bel(\bar{U}) = 1$ , but  $Bel$  must satisfy certain axioms (whose details do not matter for our purposes) ensuring that  $Bel(U) + Bel(\bar{U}) \leq 1$ .  $Bel$  can be equivalently represented by its *plausibility function*  $Plaus(U) := 1 - Bel(\bar{U})$ . It is easy to see that  $Bel(U) \leq Plaus(U)$ , and if  $Bel$  is a probability measure, then  $Bel = Plaus$ .

Suppose we come across evidence that supports an event  $\phi \subseteq W$  to a degree  $\alpha \in [0, 1]$ . Together,  $\phi$  and our confidence  $\alpha$  in it can be represented by a belief function  $Bel_{(\alpha, \phi)}$  that Shafer calls a *simple support function*, by

$$Bel_{(\alpha, \phi)}(U) = \begin{cases} 0 & \text{if } U \subseteq \phi \\ \alpha & \text{if } \phi \subseteq U \subsetneq W \\ 1 & \text{if } U = W \end{cases}$$

To combine our prior with the new evidence, Shafer argues for Dempster's *rule of combination*; in this case, that means adopting the posterior belief  $Bel' := Bel \oplus Bel_{(\alpha, \phi)}$ , whose plausibility measure is given by

$$Plaus'(U) = \frac{\alpha Plaus(U \cap \phi) + (1 - \alpha) Plaus(U)}{1 - \alpha + \alpha Plaus(\phi)}. \quad (11.1)$$

It is easy to verify that  $Bel' = Bel$  when  $\alpha = 0$ . At the other extreme, it can be shown<sup>1</sup> that  $Bel'(\phi) = Plaus'(\phi) = 1$  when  $\alpha = 1$ . Thus, confidence  $\alpha \in [0, 1]$  parameterizes a continuous path from ignoring  $\phi$  to fully incorporating it.

In the special case where  $Bel = Plaus$  is a probability measure, a full confidence update ( $\alpha = 1$ ) yields the same conditioned probability  $Plaus' = (Plaus|\phi)$

---

<sup>1</sup>see the appendix for proof

as in [Example 11.1](#). Furthermore, the set of possible posteriors for intermediate  $\alpha \in (0, 1)$  is the same in both cases. However, the two paths are parameterized differently; in fact, for all  $\alpha \in (0, 1)$  the two updates disagree. It follows that the appropriate numerical value of confidence  $\alpha$  must depend on more than just an intuition of “fraction of the way to the update”.

If we use  $\oplus$  to combine two (independent) simple support functions for  $\phi$  with degrees of support  $\alpha_1$  and  $\alpha_2$ , we get another simple support function for  $\phi$ , with combined support  $\alpha_1 + \alpha_2 - \alpha_1\alpha_2$ . We will later see that this is one canonical form of confidence. Is there a way of representing confidence that combines additively? There is; Shafer calls it *weight of evidence*, and proves it must be of the form  $w = -k \log(1 - \alpha)$  for some  $k > 0$  ([Shafer 1976](#), p. 78). This additive form of confidence plays a fundamental role in Shafer’s theory, as well as ours.  $\triangle$

Shafer’s theory aims to address two problematic aspects of the Bayesianism: it prescribes a belief state (belief functions) that can represent ignorance, and enables observations other than those that “establish a single proposition with certainty” ([Shafer 1976](#), Chapter 1: §7,§8). Ironically, in solving the first problem, his solution to the second becomes inaccessible to those who do not work with belief functions. Our notion of confidence directly addresses Shafer’s second concern, but applies far more broadly. To illustrate, we now give a very different example with the same critical elements.

**Example 11.3** (Training a Neural Network). The “belief state” of a neural network may viewed as a setting  $\theta \in \Theta \subseteq \mathbb{R}^d$  of weight parameters. For definiteness, suppose we are talking about a classifier, so that there is a space  $X$  of inputs, a finite set  $Y$  of labels, and a parameterized family of functions  $\{f_\theta : X \rightarrow \Delta Y\}_{\theta \in \Theta}$  mapping inputs  $x \in X$  to distributions  $f_\theta(x) \in \Delta Y$  over labels. In the supervised

setting, an observation  $\phi$  is a pair  $(x, y)$  consisting of an input  $x$  annotated with a label  $y$ .

Suppose we now observe  $\phi = (x, y)$  with some degree of confidence; how should we update the weights  $\theta$ ? In contrast with previous examples, it is not so obvious how to learn  $\phi$  with full confidence. Instead, modern learning algorithms<sup>2</sup> tend to be iterative procedures  $\text{step} : (X \times Y) \times \Theta \rightarrow \Theta$  that make small adjustments  $\theta \mapsto \text{step}(\phi, \theta)$  to the weights. Each step is essentially a low-confidence update. There is no guarantee, for example, that  $f_{\text{step}(\phi, \theta)}(x)$  gives high probability to  $y$ —only that it is higher than it was before. This lower level of confidence is arguably what makes these learning algorithms robust to noisy and contradictory inputs.

Higher levels of confidence can be obtained by applying `step` more than once. Beginning with initial weights  $\theta_0$ , and defining  $\theta_{n+1} = \text{step}(\phi, \theta_n)$ , we get a sequence of weights  $(\theta_0, \theta_1, \theta_2, \dots)$  that converges to some  $\theta_* \in \Theta$ . These limiting weights fully incorporate  $\phi$  into  $\theta_0$  in at least two senses:  $\theta_* = \text{step}(\phi, \theta_*)$  so  $\phi$  cannot be further incorporated by `step`, and  $f_{\theta_*}(x)(y) = 1$ , so the classifier classifies  $x$  as  $y$  with probability 1. Correspondingly, adopting belief  $\theta_*$  is appropriate only if we have complete trust in  $\phi$ , meaning we find it critical that  $x$  be classified as  $y$ . At the opposite extreme, if we have no confidence in  $\phi$ , we should not update  $\theta$  at all. Thus, the number of training iterations  $n$  is a measure of confidence: it interpolates between no confidence (zero iterations of `step`) and full confidence (infinitely many iterations of `step`). It is also additive.

In the simplest settings, training examples do not come with confidence

---

<sup>2</sup>(in contrast to their historical counterparts like conjunction learning ?, and learning algorithms for decision trees)

annotations, in which case one effectively treats them all with the same default confidence (by selecting a learning rate). The number of times that  $\phi = (x, y)$  appears in a dataset is then the de-facto measure of confidence in  $\phi$ . Often, though, these numbers are not our intended confidences, which is why it can be helpful to remove duplicates (?). In richer settings, a more nuanced degree of confidence specific to each training example often arises, such as agreement between annotators (?), or confidence scores in self-training ([Zou et al. 2019](#)).

It is worth emphasizing that confidence in a training example is not merely a matter of accuracy. Suppose, for example, that the classifier is intended to screen job applications, and that we would like to change our current hiring practices to be less discriminatory. In this case, we should have low confidence in training data based on our previous hiring decisions—not because it is inaccurate, but because we don't want it to take it too seriously in forming our new hiring practice.  $\triangle$

Perhaps the most important application of confidence is in treating different sources of information with different degrees of trust. As a result, one might imagine confidence to be relevant for sensor fusion: the problem of combining information from multiple different sensors (of varied reliability). The standard approach to sensor fusion is called a Kalman filter ([Kalman 1960](#); [Brown and Hwang 1997](#))—and, indeed, comes with its own notion of confidence.

**Example 11.4** (1D Kalman Filter). Suppose we are interested in modeling a dynamical system whose state is a real number  $x \in \mathbb{R}$ , and we receive observations  $z$  which we assume is the value of  $x$  plus Gaussian noise. In many engineering disciplines, the standard way to track this information is the [Kalman filter](#) [Kalman \(1960\)](#). It prescribes belief state  $(\hat{x}, \sigma^2)$ , where  $\hat{x} \in \mathbb{R}$  is our current

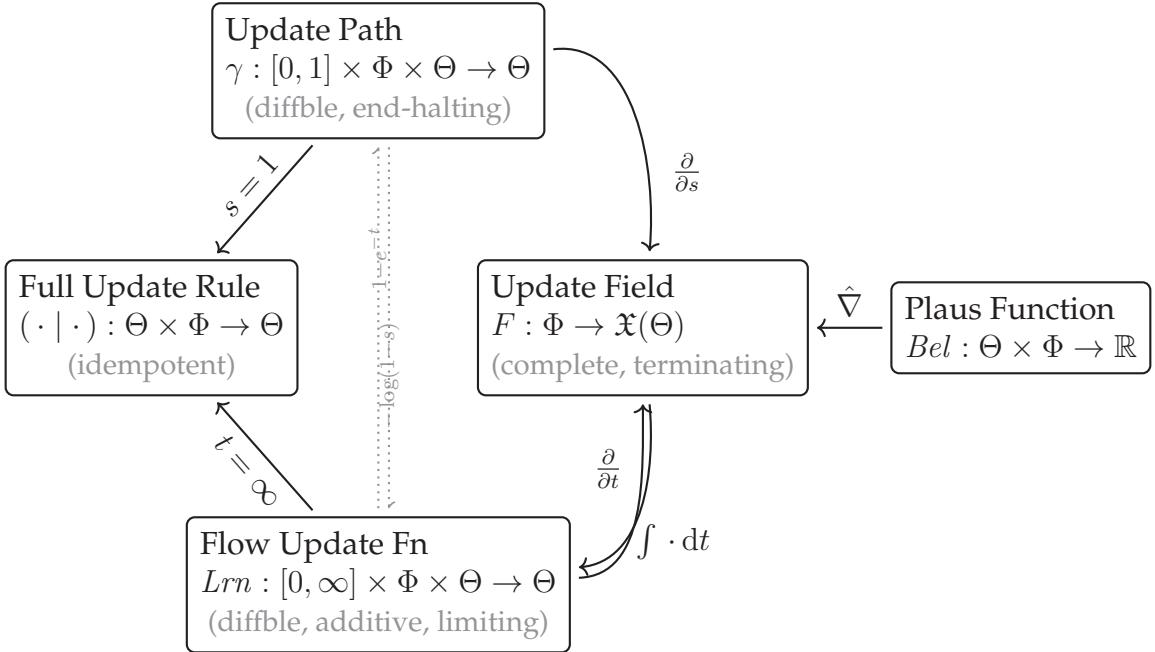


Figure 11.1: Different representations of update functions, and the relationships they have with one another.

estimate of  $x$ , and  $\sigma^2$  is a variance (effectively encoding the probabilistic belief  $x \sim \mathcal{N}(\hat{x}, \sigma^2)$ ). Suppose we now receive an observation  $z \sim \mathcal{N}(x, r^2)$  from a sensor of known variance  $r^2$ . How should we update our beliefs in response to this new information?

The right answer, which ranges from ignoring  $z$  to replacing  $\hat{x}$  with it, depends on how much we trust the sensor. The 1D Kalman filter measures confidence in the observation with a quantity  $K$  called *Kalman gain*, which is used to compute posterior beliefs  $(\hat{x}', \sigma'^2)$  according to:

$$\hat{x}' = \hat{x} + K(z - \hat{x}) = (1 - K)\hat{x} + (K)z; \quad \sigma'^2 = (1 - K)^2\sigma^2 + (K)^2r^2.$$

Like other notions of confidence we have seen,  $K$  interpolates (linearly) between our prior mean  $\hat{x}$  and the new observation  $z$ , and (“quadratically”) between our prior uncertainty  $\sigma^2$  and the sensor variance  $r^2$ . Moreso than in previous examples, we can also say something prescriptive about how best to select a

degree of confidence. This is made possible by two assumptions: (1) we can objectively quantify the reliability of an observation (with the variance  $r^2$ ), and (2) the objective is to minimize uncertainty in our posterior beliefs (the mean squared error of  $\hat{x}$ ). Under these assumptions, the optimal Kalman gain is  $K_{\text{opt}} = \sigma^2 / (\sigma^2 + r^2)$  (Brown and Hwang 1997, p. 146), and so  $K$  is typically chosen this way in practice (Becker 2003). Plugging this value into the update equations, we find that this choice makes  $\hat{x}'$  the average of our prior  $\hat{x}$  and new observation  $z$ , weighted by their respective variances.

With this in mind, let's revisit what happens extreme values of confidence. If  $K = 0$ , which is optimal iff the noise  $\xi$  has unbounded variance, then the update leaves our belief state unchanged. In this case, there is intuitively so much noise in observations  $z$  that we should not trust them at all. At the other extreme, if no noise is added to observations ( $r^2 = 0$ ), then  $K_{\text{opt}} = 1$  and we end up with posterior  $(z, 0)$  that fully trusts the new observation.  $\triangle$

**Example 11.4** features three distinct kinds of (un)certainty:

1. **Learner's Confidence:** a subjective trust in an observation which tells how seriously to take it in updating ( $K$ , in [Example 11.4](#)) ;
2. **Internal Confidence:** a way of quantifying the degree of uncertainty present in a belief state either overall or in a particular proposition ( $\sigma^2$  or the probability density  $\phi \mapsto \mathcal{N}(\phi|\hat{x}, \sigma^2)$  in [Example 11.4](#), respectively);
3. **Statistical Confidence:** an objective measure of the (un)reliability of an observation, based on historical data and/or modeling assumptions about how observations arise (the noise level  $r^2$ , in [Example 11.4](#)).

The three can be closely related, but are very different in nature. Our focus is on

confidence in sense 1, and how learning uses this kind of confidence to update beliefs.

We have tried to contrast learner’s confidence (sense 1) with the more pervasive usage of the word “confidence” (sense 2), such as likelihood, precision, or degree of belief. Other internal confidences include the probability  $\Pr(\phi)$  in Example 11.1, the degree of belief  $Bel(\phi)$  in Example 11.2, and the value of the loss function  $\mathcal{L}(\theta, \phi)$  used to train the classifier in Example 11.3. These two kinds of confidence are related: internal confidences may be thought of as aggregate reflections of learner’s confidence in past observations. We explore these connections in more detail in the coming sections.

We would also like to distinguish our notion of learner’s confidence (sense 1) from statistical confidences (sense 3) such as the variance in readings of a sensor ( $r^2$  in ??) or annotator agreement (from Example 11.3). When readily available, the statistical reliability of a source of information should absolutely play a role in determining how seriously we take it in updating our beliefs—but it can be difficult to come by. We may not always know the variances of our sensors, and that such a quantity is well-defined is a significant assumption on its own. Statistical confidences typically require us to know that observations are drawn independently from a fixed distribution, while learners’s confidence can be meaningful even without this assumption.

We hope that these examples have given the reader an intuitive sense of what confidence is, how ubiquitously it arises, and why it is important. What’s more, it also has a clean mathematical theory. The rest of this chapter develops the general theory of learner’s confidence, characterizing it axiomatically (??),

**Contributions.** We have already motivated the notion of confidence and illustrated many of its most important properties by example. In the remainder of the chapter, we study confidence more formally. We develop a formal framework for talking about confidence. We show that confidence can be measured in several equivalent ways, and classify the ways that In each stage, we make successively stronger assumptions (all of which apply to Examples 11.1 to 11.4), to develop new representations of these learners, which are summarized in Figure 11.1. The final representations we consider (vector fields and loss functions) also enable simultaneous orderless updates, even in settings where it was not previously possible.

## 11.2 A Formal Model of Confidence, Learning, and Belief

Our formalism consists of three parts: a domain  $[\perp, \top]$  of confidence values, a space  $\Theta$  of belief states, and a language  $\Phi$  of possible observations. For instance:

- In Example 11.1,  $\Theta$  is the set of probability measures on some measurable space  $(\Omega, \mathcal{F})$ ,  $\Phi$  is the  $\sigma$ -algebra  $\mathcal{F}$ , and the confidence domain is  $[0, 1]$ .
- In Example 11.2,  $\Theta$  is the set of belief functions over a finite set  $W$ ,  $\Phi = 2^W$  is the set of subsets of  $W$ , and confidence is a degree of support  $\alpha \in [0, 1]$  or a weight of evidence  $w \in [0, \infty]$ .
- In Example 11.3,  $\Theta \subseteq \bar{\mathbb{R}}^d$  is the space of network parameters,  $\Phi = X \times Y$  is the space of input-label pairs, and the confidence domain is the extended natural numbers  $\{0, 1, \dots, \infty\}$  under addition.
- In Example 11.4,  $\Theta = \Phi = \mathbb{R}$ , and  $[\perp, \top] = [0, 1]$  is the domain of  $K$ .

Together, we call the triple  $(\Theta, \Phi, [\perp, \top])$  a *learning setting*. A *learner* in the setting

$(\Theta, \Phi, [\perp, \top])$  is a function  $Lrn : \Phi \times [\perp, \top] \times \Theta \rightarrow \Theta$  that describes the belief update process. Explicitly: from a prior belief  $\theta$ , and a statement  $\phi$  with some degree of confidence  $\chi$ , a learner produces a posterior belief state  $Lrn(\phi, \chi, \theta) \in \Theta$ . We use superscripts and subscripts to fix some arguments of  $Lrn$  and view it as a function of the others, so that  $Lrn(\phi, \chi, \theta)$  can equivalently be written as  $Lrn_\phi(\chi, \theta) = Lrn_\phi^\chi(\theta) = Lrn^\chi(\phi, \theta) = Lrn_{(\theta, \phi)}(\chi)$ . The rest of this section develops axioms and auxiliary concepts that ensure that these functions capture the the update procsss.

We do so in three stages, starting with two fragments of the theory that isolate some critical mathematical properties of confidence. The first an abstract theory of confidence domains  $[\perp, \top]$  themselves (Section 11.2.1); the second is a theory of *commitment functions*, which introduce  $\Theta$  and describe the update process (Section 11.2.2). Finally, we bring in the observations  $\Phi$  (Section 11.2.3).

### 11.2.1 Abstract Confidence Domains

A *confidence domain*  $(D, \leq, \perp, \top, \oplus, \mathbf{g})$  is a set  $D$  of confidence values equipped with a preorder  $\leq$ , a least element  $\perp$  (“no confidence”), a greatest element  $\top$  (“full confidence”), and an operation  $\oplus$  that combines two independent degrees of confidence into a single one. We often abbreviate a confidence domain as  $D = [\perp, \top]$ , leaving  $\leq$  and  $\oplus$  implicit. Because  $\oplus$  represents *independent* combination, we require that it be commutative and associative. We want to ignore independent information we have no confidence in, and, if already fully confident, remain so the face of new independent information. Formally, this amounts to requiring, for all  $\chi, \chi', \chi'' \in D$ :

$$\begin{aligned}
(\chi \oplus \chi') \oplus \chi'' &= \chi \oplus (\chi' \oplus \chi'') && \text{(associativity),} \\
\chi \oplus \chi' &= \chi' \oplus \chi && \text{(commutativity),} \\
\perp \oplus \chi &= \chi && \text{(that } \perp \text{ is neutral),} \\
\top \oplus \chi &= \top && \text{(and that } \top \text{ is absorbing).}
\end{aligned}$$

Finally,  $D$  comes with geometric information  $\mathbf{g}$ , which may optionally include a topology, a differentiable structure. Our work will focus on two particularly important confidence domains describing a continuum of confidence, both of which appear in our examples. The first is the *fractional domain*  $[0, 1]$ , whose elements  $s \in [0, 1]$  represent the “proportion of the way towards complete trust”. If you go proportion  $s$  towards fully trusting something, then  $s'$  of the remaining way, then overall you have gone  $s \oplus s' := s + s'(1 - s) = s + s' - s \cdot s'$  of the way to complete trust. The other confidence domain of particular interest is the *additive domain*  $[0, \infty]$ , which is ideal for analogies of time and weight. We will later see that the two are isomorphic, and have a particularly rich theory.

**Proposition 11.1.** *The fractional domain  $[0, 1]$  the additive domain  $[0, \infty]$  are isomorphic. Furthermore, the space of isomorphisms between them is in natural bijection with  $(0, \infty)$ . Specifically, for each  $\beta \in (0, \infty)$ , there is an isomorphism  $\varphi_\beta : [0, 1] \rightarrow [0, \infty]$  given by*

$$[0, 1] \ni s = 1 - e^{-\beta t} = \varphi_\beta^{-1}(t) \quad \text{and} \quad \varphi_\beta(s) = -\frac{1}{\beta} \log(1 - s) = t \in [0, \infty].$$

### 11.2.2 Belief States and Commitment Functions

We now reintroduce the space  $\Theta$  of beliefs, for the purpose of characterizing how confidence effects belief updates. In this section, we describe the essential

properties of confidence in terms of functions  $Lrn_\phi : [\perp, \top] \times \Theta \rightarrow \Theta$ . Many of the most important aspects of confidence can be characterized purely in terms of such functions, keeping  $\phi$  entirely abstract. We call a function of this type a *commitment function*, if it is obeys certain axioms. First, having no confidence ( $\perp$ ) in an observation means we should ignore it.

$$[\mathbf{L1}] \quad \forall \phi, \theta. \quad Lrn(\phi, \perp, \theta) = \theta$$

Second, we intend for  $Lrn$  only to be used to incorporate information to the extent that it is novel, i.e., information that is not already accounted for in our prior beliefs. Thus, we would like a sequence of two (independent) observations in the same observation  $\phi$  to be equivalent to a single observation of  $\phi$  with their combined degree of confidence.

$$[\mathbf{L2}] \quad \forall \phi, \chi, \chi'. \quad Lrn_\phi(\chi, Lrn_\phi(\chi', \theta)) = Lrn_\phi(\chi \oplus \chi', \theta).$$

The reader is encouraged to verify that Examples 11.1 and 11.4 satisfy [L2](#) for the domain  $[0, 1]$ . For a specific confidence domain, [L2](#) can be quite a strong assumption. However, if we are free to chose the confidence domain, [L2](#) imposes no restrictions at all (see [Proposition 11.15](#) in the appendix). Keep in mind that [L2](#) applies only for two observations of the same statement  $\phi$ . In the language of algebra, [L1](#) and [L2](#) together require  $Lrn_\phi$  to be an *action* of the monoid  $([\perp, \top], \oplus, \perp)$  on  $\Theta$ . Beyond the data of this monoid, a confidence domain  $D$  also has an order ( $\leq$ ), a geometry, and an absorbing top element ( $\top$ ). What should it mean for  $Lrn$  to preserve this additional structure?

**Geometry.** Learner's confidence is meant to interpolate between ignoring new information and defering to it entirely, and the most natural (and useful)

interpolations are continuous and smooth.

**[L3]** If  $[\perp, \top]$  and  $\Theta$  are both topological spaces, then for all  $\theta$  and  $\phi$ , the map

$Lrn_{(\theta, \phi)} = \chi \mapsto Lrn(\theta, \chi, \phi)$  is continuous. If  $[\perp, \top]$  and  $\Theta$  are both manifolds, then  $Lrn_{(\theta, \phi)}$  is differentiable.

Ideally, the update would be continuous in our initial beliefs as well: similar priors typically result in similar posterior beliefs. This suggests a simple strengthening of **L3**: that  $Lrn_\phi : [\perp, \top] \times \Theta \rightarrow \Theta$  be continuous (resp. differentiable). Unfortunately, that is too strong to handle our examples at full confidence. In the probabilistic case, for instance:

**Proposition 11.2.** *There is no extension of conditioning that satisfies ??.* That is, for  $\phi \subsetneq W$ , there is no continuous function  $F_\phi : \Delta W \times [0, 1] \rightarrow \Delta W$  such that  $F_\phi(\mu, 1) = \mu|\phi$  when  $\mu(\phi) > 0$ .

Intuitively, though, this is just an edge case; we can still get continuity if we never observe an event we believe has probability zero. Rather than insisting on this stronger axiom or giving up on it entirely, we can get something in between with the following definition.

**Definition 11.1.** Given  $\phi \in \Phi$ , let  $\Theta_\phi \subseteq \Theta$  be the maximal set open set for which the restriction  $Lrn_\phi|_{\Theta_\phi} : [\perp, \top] \times \Theta_\phi \rightarrow \Theta$  of  $Lrn_\phi$  to  $\Theta_\phi$  is continuous. (unproven!)  $\square$

In each of our examples,  $\Theta_\phi$  consists of those belief states that do not flatly contradict  $\phi$ . In [Example 11.1](#), for instance, [Proposition 11.2](#) and ?? imply that  $\Theta_\phi = \{\mu \in \Delta W : \mu(\phi) > 0\}$  is the set of distributions  $\mu$  on which conditioning on  $\phi$  is well-defined. In [Example 11.3](#), if  $\phi = (x, y)$ , then  $\Theta_\phi$  is the parameter space

in which the gradients  $\nabla_\theta \ell(f_\theta(x), y)$  of the loss function  $\ell$  are finite.

**Order.** For a learner, the characteristic property of  $\chi \leq \chi'$  is that learning with confidence  $\chi$  is more conservative than learning with  $\chi'$ , in the following sense: one can get the effect of learning  $\phi$  with higher confidence by first learning with lower confidence, and then making a second update with some residual degree of confidence.

$$[\mathbf{L4}] \quad \forall \phi, \theta, \chi, \chi'. \quad \chi \leq \chi' \iff \exists \chi'' \leq \chi'. \quad Lrn_\phi^{\chi''} \circ Lrn_\phi^\chi(\theta) = Lrn_\phi^{\chi'}(\theta).$$

$$[\mathbf{L4}^<] \quad \forall \phi, \theta, \chi, \chi'. \quad \chi < \chi' \iff \exists \chi'' < \chi'. \quad Lrn_\phi^{\chi''} \circ Lrn_\phi^\chi(\theta) = Lrn_\phi^{\chi'}(\theta).$$

Observe that, in the presence of [L2](#), the right hand side of [L4](#) is tautology when  $\chi' = \top$  or if  $\chi = \perp$ ; thus our axioms would have implied  $\perp \leq \chi \leq \top$  for all  $\chi$  even had we not assumed this. There is a sense in which [L4](#) can be viewed as a limited form of subtraction for confidences, although the result of that subtraction may be nondeterministic and depend on  $\theta$  and  $\phi$ .

In some sense, the effect of [L4](#) is to rule out cases in which “stopping and resuming” an update results in completely different beliefs than we would have obtained with a single update. [L4](#) also ensures that a high confidence update can be decomposed into several updates of lower confidence.

For the domain  $[0, \infty]$  (and hence, in  $[0, 1]$ , which is isomorphic to it), [L4](#) and its strict analogue  $\mathbf{L4}^<$  are direct consequences of [L2](#).

**Full-confidence updates.** Historically, most of the literature on belief updates is about idempotent updates, which in our framework, correspond to

full-confidence updates. Full updates are quite extreme. An agent that updates by conditioning, for instance, permanently commits to believing everything it ever learns, and gains nothing from making the same observation twice. Clearly humans are not like this; revisiting information improves our learning ([Ausubel and Youssef 1965](#)). Similarly, artificial neural networks are trained with many incremental updates, and benefit from seeing the training data many times. We would like an account that allows for less extreme belief alterations, in which information is only partially incorporated. This is the role of intermediate degrees of confidence.

To understand what a full-confidence update means in our framework, we turn the top element of the confidence domain. How should updates Our axioms already have some implications for it. Because  $Lrn$  preserves monoidal structure ([L2](#)) and  $\top$  is absorbing, it follows that  $Lrn_\phi^\top : \Theta \rightarrow \Theta$  is a projection, i.e.,  $Lrn_\phi^\top \circ Lrn_\phi^\top = Lrn_\phi^\top$ . This suggests that we could use  $Lrn$  to define the belief states in which “ $\phi$  is true” to be image of this projection (i.e., the set of fixed points of  $Lrn_\phi^\top$ ); after all, it is easily shown that learning  $\phi$  with any degree of confidence (i.e., applying  $Lrn_\phi^\chi$ ) has no effect on these states. This illustrates a general point: if the function  $Lrn$  captures the belief updating process, we can use it to understand the relationship between  $\Phi$  and  $\Theta$  at an abstract level. In [Example 11.3](#), for instance, although the network weights  $\Theta$  are an uninterpreted subset of some high dimensional space, the training process  $Lrn$  arguably imbues them with meaning by defining a connection between them and the training examples.

However, to some readers, using  $Lrn$  to define truth may seem backwards. In a given learning setting, we may already have a sense of which belief states  $\theta$  correspond to full belief in  $\phi$ —in [Example 11.1](#), for instance, they are the

measures that give  $\phi$  probability 1. In such cases, we may want additional axioms ensuring that any relationships between  $\Theta$  and  $\Phi$  implicit in  $Lrn$  are compatible with the ones we already have. Our axioms so far have been conditions on the separate functions  $Lrn_\phi : [\perp, \top] \times \Theta \rightarrow \Theta$ , but we have not required that  $Lrn_\phi$  be intrinsically related  $\phi$ . To address this, we will need to assume some additional structure.



### 11.2.3 Modeling Observations: Degree of Belief, and Structural Symmetry

Recall that a *learning setting* is a triple  $(\Theta, \Phi, [\perp, \top])$  consisting of a space  $\Theta$  of beliefs, a language  $\Phi$  of observations, and a confidence domain  $[\perp, \top]$ .

**Belief.** In a learning setting  $(\Theta, \Phi, [\perp, \top])$ , a *believer* is a function  $Bel : \Theta \times \Phi \rightarrow [\perp, \top]$ , and can be thought of as annotating each state  $\theta$  with a function  $Bel_\theta : \Phi \rightarrow [\perp, \top]$  that gives a degree of belief to each observation. Thus the output of a believer is an *internal* confidence—like a probability or a precision—not a learner’s confidence. Our primary reason for defining  $Bel$  is that we would like  $Lrn$  to be monotonic with respect to  $Bel$ . That is, learning  $\phi$  with more confidence should lead to more belief in  $\phi$ .

$$[\text{LB1}] \quad \forall \phi, \theta, \chi, \chi'. \quad \chi \geq \chi' \implies Bel(\phi, Lrn(\phi, \chi, \theta)) \geq Bel(\phi, Lrn(\phi, \chi', \theta)).$$

We cannot ask for strict monotonicity, however: if we already fully believe  $\phi$  (i.e.,  $Bel(\phi, \theta) = \top$ ), there is no way to attain a higher degree of belief, we cannot attain a higher degree of belief by learning  $\phi$ . Instead, if we fully believe  $\phi$ , learning  $\phi$  should have no effect. Perhaps even more importantly, if we learn

something with full confidence, then we ought to fully believe it.

**[LB2]** If  $Bel(\phi, \theta) = \top$ , then  $Lrn(\phi, \chi, \theta) = \theta$ .

**[LB3]**  $Bel(\phi, Lrn(\phi, \top, \theta)) = \top$ .

While **LB3** is certainly desirable, it may not always hold in cases of interest. In [Example 11.3](#), for instance, it is natural to set  $Bel(\theta, (x, y)) = f_\theta(y|x)$ , and there may be a local maximum  $\theta$  of the parameterization  $\theta \mapsto Bel(\theta, (x, y))$ . In this case, there is no continuous monotonic path from  $\theta$  to a global maximum  $\theta^*$  for which  $f_{\theta^*}(y|x) = 1$ , i.e., no way to simultaneously satisfy **LB1**, **LB3** and [L3](#).

**Symmetry.** We would also like update rules to preserve any joint symmetries between the belief space  $\Theta$  and the observation language  $\Phi$ . For instance, in [Example 11.1](#), we would like to require that updates are not sensitive to irrelevant relabelings of points. Concretely, assume we have some set  $\text{Aut}(\Theta, \Phi)$  of structural symmetries (in the form of automorphisms  $\sigma : (\Theta \sqcup \Phi) \rightarrow (\Theta \sqcup \Phi)$ ) that have an action both on belief states ( $\sigma(\theta) \in \Theta$ ) and on observations ( $\sigma(\phi) \in \Phi$ ). For example, in the setting where  $\Theta = \Delta\Omega$  is the set of probability distributions over a finite set  $\Omega$ , and  $\Phi = 2^\Omega$  is the set of events, one might take  $\text{Aut}(\Theta, \Phi) \cong \text{Aut}(\Omega)$  to be the set of permutations of  $\Omega$ , with the obvious relabeling action on both distributions  $\theta \in \Theta$  and on events  $\phi \in \Phi$ .

Once we select a suitable group of automorphisms (which could well just include the identity), the symmetry condition can now be captured by:

**[L5]**  $\forall \theta, \phi, \chi, \sigma \in \text{Aut}(\Theta, \Phi). \quad Lrn(\sigma(\phi), \chi, \sigma(\theta)) = \sigma(Lrn(\theta, \chi, \phi))$ .

### 11.3 Commitment on Confidence Continua

In this section, we focus precisely on the special case in which the confidence domain is a continuum of real numbers. There are two particularly important confidence domains in this setting are the fractional domain  $[0, 1]$  and the additive domain  $[0, \infty]$ .

Most quantities used in science and everyday life can be measured additively: if one starts with seven minutes/meters/votes/dollars, and then gains six more, one has thirteen altogether. To measure confidence in the same way, we must use the domain  $[0, \infty]$ . With this confidence domain, **L2** means  $Lrn$  is *additive*, making it amenable to analogies of weight (e.g., the weight of evidence  $w$  in [Example 11.2](#)) and time (e.g., the number of training iterations  $n$  in [Example 11.3](#)). Indeed, a learner is additive iff it can be implemented so that confidence really does coincide with time: imagine a machine with state space  $\Theta$ , controlled by buttons labeled by  $\Phi$ , that, while  $\phi$  is pressed, evolves from initial state  $\theta_0$  according to  $\theta(t) = Lrn(\phi, t, \theta)$ . Observe that this scenario is coherent if and only if  $Lrn$  is additive; if **L2** did not hold, there would exist  $t_1, t_2$  such that the machine's state after pressing  $\phi$  for  $t_1$  seconds followed by  $t_2$  additional seconds, would be different from the configuration after holding down  $\phi$  for  $t_1 + t_2$  seconds.

Temporal analogies may not always be appropriate (as they may clash with other, truer conceptions of “time”), yet they have such intuitive force that a function  $f : [a, b] \times \Theta \rightarrow \Theta$  (with  $0 \in [a, b] \subseteq \mathbb{R}$ ) satisfying **L1–3** is known generically as a *flow* ([Lee 2013](#)). Recall that **L2** implies **L4** for this domain. Thus, the only non-standard requirement is that  $Lrn_{(\theta, \phi)}$  has a well-defined limit at  $\infty \notin \mathbb{R}$ . The assumptions that confidence lies in  $[0, \infty]$  and combines additively are collectively quite strong, and one might understandably worry that it could

limit the applicability of our formalism. Fortunately, this is not the case. While additivity ([L2](#) for  $[0, \infty]$ ) does pin down how confidence can be measured, it has no effect on what confidence can express.

**Theorem 11.3.** *If  $Lrn$  satisfies [L1](#), [L3](#) and [L4](#) and ?? (but possibly not [L2](#)) then there exist  ${}^+Lrn$  and a continuous function  $g : \Phi \times [\perp, \top] \times \Theta \rightarrow [0, \infty]$  such that*

$$\forall \theta, \phi, \chi. \quad Lrn(\phi, \chi, \theta) = {}^+Lrn(\phi, g(\phi, \chi, \theta), \theta) \quad \text{and} \quad {}^+Bel(\phi, \theta) = g(\phi, Bel(\phi, \theta), \theta).$$

Furthermore,  $({}^+F, g)$  is unique up to a multiplicative factor in the output of  $g$ .

**Corollary 11.3.1.** *There is a unique choice of  $({}^+F, \beta)$  such that  ${}^+F$  and  $F$  have the same effect on observations made with sufficiently low confidence, i.e.,  $\frac{\partial \beta}{\partial \chi}|_{\chi=\perp} = 1$ .*

Thus, updates performed with  $F$  are equivalent to updates performed with  ${}^+F$ , except that the degree of confidence needs to be translated appropriately (via  $\beta$ ). We call  ${}^+F$  the *additive form of  $F$* , and  $\beta(\phi, \chi, \theta)$  the additive form of confidence  $\chi$ . Ideally, the translation  $g$  to an additive scale should not depend on our current beliefs  $\theta$  or observation  $\phi$ .

### 11.3.1 The Vector Fields of Commitment Functions, and Orderless Combination

Suppose we learn  $\phi_1$  (with confidence  $\chi_1$ ), and then  $\phi_2$  (with confidence  $\chi_2$ ). Is this the same as learning them in the opposite order? This is true of belief functions ([Example 11.2](#)) and of conditioning, so we call them *commutative*—but, in general, observing inputs in different orders yields different results. Humans tend to have a recency bias: more recent observations have a stronger influence on

beliefs. Examples 11.1 and 11.4 have this property as well. But, the order matters for our update, what should we do if we receive two pieces of information simultaneously? It turns out that we already have the tools to do this in a natural way, even if  $\phi_1$  and  $\phi_2$  do not commute.

We now turn to an equivalent representation of flow update functions, which, among other things, will ultimately yield a natural way of orderlessly learning  $\phi_1$  and  $\phi_2$  together, and weighted by relative confidence. At a technical level, we show how to extend an arbitrary update function  $F$ , that handles inputs  $\Phi$ , to handle a more expressive set of inputs  $\bar{\Phi} \supseteq \Phi$  closed under new operations of orderless combination ( $\oplus$ ), and rescaling by relative confidence ( $\cdot$ ).

Since  $\Theta$  carries a differentiable structure, it makes sense to talk about its tangent space  $T\Theta$ , which consists of pairs  $(\theta, \mathbf{v})$  where  $\theta \in \Theta$ , and  $\mathbf{v}$ , intuitively, is a direction that one can travel in  $\Theta$  beginning at  $\theta$  (Lee 2013, §3). A *vector field*  $X \in \mathfrak{X}\Theta$  is a differentiable map  $X : \Theta \rightarrow T\Theta$  assigning to each point  $\theta \in \Theta$  a vector  $X(\theta) = (\theta, \mathbf{v}) \in T\Theta$  tangent to  $\theta$ . Additivity (L2) implies that the behavior of flow update functions is determined by the way it handles updates of small confidence. So, in a sense, the only thing we need to know about a flow update function is how it handles infinitesimal confidences, which is to say, its derivative at zero confidence—which can be viewed as a vector field. More precisely, a commitment function  $Lrn_\phi$  has a *vector field representation* given by

$$Lrn'_\phi := \theta \mapsto \frac{\partial}{\partial \chi} Lrn(\theta, \chi, \phi) \Big|_{\chi=0} \quad \in \mathfrak{X}\Theta. \quad (11.2)$$

Moreover, we can recover  $Lrn_\phi$  via the integral curves of  $Lrn'_\phi$ .

**Fact 11.4** (Lee, Thm 9.12). *If  $X \in \mathfrak{X}(\Theta)$ , there is at most one  $f : [0, \infty) \times \Theta \rightarrow \Theta$*

such that for all  $\theta \in \Theta$  and  $a, b \geq 0$ ,

$$f(a, f(b, \theta)) = f(a + b, \theta) \text{ and } \frac{\partial}{\partial \chi} f(\chi, \theta) \Big|_{\chi=0} = X(\theta).$$

**Corollary 11.4.1.** *If  $F_{\phi_1}$  and  $F_{\phi_2} : \Theta \times [0, 1] \rightarrow \Theta$  are distinct, then so are  $F'_{\phi_1}$  and  $F'_{\phi_2}$ .*

Thus, every flow update function  $F$  can be equivalently represented by its differential  $F'$ , a collection of vector fields. It may seem counter-intuitive that  $F'_\phi$ , which no longer explicitly mentions confidence at all, can capture confidence. But it does—in a sense, by specifying everything about the update *except* for the degree of confidence. This vector field representation is useful for two reasons: at a practical level, it gives us a natural extension of  $\Phi$  that allows us deal with “mixtures” of observations and commonly arise. At a deeper level, it will enable us to describe and classify the flow update functions on  $\Theta$ .

One important feature of vector fields is that they can be linearly combined to form new vector fields. Since in the presence of a flow update function, observations correspond to vector fields, observations also inherit this linear structure. There are two aspects of linearity: scalar multiplication, and addition. From scalar multiplication, we get a way of rescaling inputs by a “relative confidence”  $k$ . Concretely, given  $\phi \in \Phi$  and  $k \in [0, \infty)$ , define a new observation  $k \cdot \phi$  and extend  $F$  to a function  $\bar{F}$  that handles it by:

$$F_{k \cdot \phi}^\chi(\theta) := F_\phi^{k\chi}(\theta), \quad \text{or equivalently,} \quad F'_{k \cdot \phi} := kF'_\phi.$$

Note that if  $k > 0$ , the rescaled input  $k \cdot \phi$  behaves the same way that  $\phi$  does for extreme values of confidence, since  $k0 = 0$  and  $k\infty = \infty$ .

From vector field addition, we get a natural way to combine observations. Up to now, we have only been able to combine observations provided they are both of the same input  $\phi$  (e.g., via ??). The vector field representation allows us to do this for distinct inputs. Concretely, given  $\phi_1, \phi_2 \in \Phi$ , we can form a new input  $\phi_1 \oplus \phi_2$  and extend  $F$  to handle it by taking its vector field  $F'_{\phi_1 \oplus \phi_2}$  to be the sum  $F'_{\phi_1} + F'_{\phi_2}$  of the vector fields of  $\phi_1$  and  $\phi_2$ . Unlike before, there is no easy way to describe the flow update function  $F_{\phi_1 \oplus \phi_2}$  directly, but Fact 11.4 implies that there's a unique such function, if it exists. We now prove that it does, except possibly for full confidence.

**Proposition 11.5.** *If  $F$  is a flow update function, and  $\phi_1, \phi_2 \in \Phi$ , then there exists a (unique) function  $F_{\phi_1 \oplus \phi_2} : [0, \infty) \times \Theta \rightarrow \Theta$  such that  $F'_{\phi_1 \oplus \phi_2} = F'_{\phi_1} + F'_{\phi_2}$ .*

The problem is that there may not be any way to continuously extend this function to handle full confidence—that is,  $\lim_{\beta \rightarrow \infty} F_{\phi_1 \oplus \phi_2}^\beta$  may not exist. Thus, it may not be meaningful to observe  $\phi_1 \oplus \phi_2$  with full confidence. For now, we leave  $\phi_1 \oplus \phi_2$  undefined in such cases, but in Section 11.3.2, we will see another representation of certain update rules predicated on a condition sufficient to ensure that these limits do exist, and  $\oplus$  is always defined.

**Proposition 11.6.** *If  $F$  is a flow update function then the following are equivalent:*

1.  $F_{\phi_1}^{\chi_1} \circ F_{\phi_2}^{\chi_2} = F_{\phi_2}^{\chi_2} \circ F_{\phi_1}^{\chi_1}$  for some  $\chi_1, \chi_2 \notin \{\perp, \top\}$ .
2.  $F_{\phi_1}^{\chi_1} \circ F_{\phi_2}^{\chi_2} = F_{\phi_2}^{\chi_2} \circ F_{\phi_1}^{\chi_1}$  for all  $\chi_1, \chi_2 \notin \{\perp, \top\}$ .
3. The vector fields  $F'_{\phi_1}$  and  $F'_{\phi_2}$  commute.
4. For all  $\chi \in \mathbb{R}$ ,  $F_{\phi_1}^\chi \circ F_{\phi_2}^\chi = F_{\phi_1 \oplus \phi_2}^\chi$ .

If this condition holds, then  $\phi_1$  and  $\phi_2$  are said to commute.

Note that  $\phi_1 \oplus \phi_2 = \phi_2 \oplus \phi_1$  when either is defined, so  $\oplus$  provides a way of combining observations orderlessly, even in cases where  $\phi_1$  and  $\phi_2$  do not commute. And, when they do, the combined observation  $\phi_1 \oplus \phi_2$  is equivalent to observing  $\phi_1$  and  $\phi_2$  in either order.

**Proposition 11.7.** *If  $F_{\phi_1}^\chi \circ F_{\phi_2}^\chi = F_{\phi_2}^\chi \circ F_{\phi_1}^\chi$ , then both updates are equal to  $F_{\phi_1 \oplus \phi_2}^\chi$ .*

Intuitively,  $\phi_1 \oplus \phi_2$  is a “mixture observation” containing one part  $\phi_1$  and one part  $\phi_2$ . This intuition is made precise by the following proposition, which shows  $\phi_1 \oplus \phi_2$  is equivalent to an infinitely fine interleaving of  $\phi_1$  and  $\phi_2$  updates.

**Proposition 11.8.** *Let  $\phi_1, \phi_2 \in \Phi$  be inputs. For  $t > 0$  and  $n \in \mathbb{N}$ , let  $u_t := F_{\phi_1}^t \circ F_{\phi_2}^t$  represent a confidence- $t$  update  $\phi_1$  followed by a confidence- $t$  update of  $\phi_2$ , and  $u_t^{(n)}(\theta) := u_t \circ \dots \circ u_t(\theta)$  be denote  $n$  sequential applications of  $u_t$  to  $\theta$ . Then,*

$$F_{\phi_1 \oplus \phi_2}^\chi(\theta) = \lim_{n \rightarrow \infty} u_{\chi/n}^{(n)}(\theta).$$

**What Distinguishes This from Control Theory?** In many ways, our framework at this point has come to resemble a dynamical system. We have a continuous manifold of states  $\Theta$ , a set of inputs (“control signals”)  $\Phi$ , which cause  $\Theta$  to evolve “over time”. However, despite the similar mathematical underpinnings, our framework is different in several important ways:

- Control theory does not require the analogue of a “full-confidence” update; there may be no limit as  $t \rightarrow \infty$ . This allows control theory to talk about a far more general class of dynamical systems without fixed points. But confidence is

- “Time” has a single clear and consistent interpretation in control theory. But the analogue here, additive confidence, is only well-defined up to a multiplicative constant. But time breaks down in other ways as well; by chaining multiple observations together, “time” extends past  $t = \infty$ . It is also sometimes helpful to think in terms of the reparameterized setting of  $[0, 1]$ .

### 11.3.2 Optimizing Learners

We have seen that commitment functions can be represented with vector fields—but how does one select an appropriate vector field? Perhaps the most important way to obtain a vector field is through the gradient of a function. This is especially true in the modern machine learning, where the learning process is typically defined by a loss.

Technically, to view the derivative of a function  $\ell : \Theta \rightarrow \mathbb{R}$  as a vector field  $\nabla \ell \in \mathfrak{X}\Theta$ , one needs more than a manifold structure; we now assume that  $\Theta$  comes equipped with a *Riemannian Metric*. The details, which can be found in the appendix, are unimportant. What matters is only that, (1) for subsets of Euclidean space  $\mathbb{R}^n$ , we can always fall back on the standard Euclidean metric, and (2) for certain special spaces, such as the parameter spaces for probability distributions, there is a different natural geometry.

Our framework allows us to express this as a particularly clean relationship between  $Lrn$  and  $Bel$ :

$$[\text{LB4}] \quad \frac{\partial}{\partial \chi} Lrn(\phi, \chi, \theta) = \nabla_\theta Bel(\phi, \theta)$$

[LB4](#) says learning occurs by gradient ascent: the learning process is about locally increasing degree of belief—no more, and no less.

### 11.3.3 Optimizing Commitment Functions for Probabilistic Beliefs

When  $\Theta$  parameterizes a family of probability distributions, via some function  $\Pr : \Theta \rightarrow \Delta X$ , there is a particularly natural metric on  $\Theta$ , called the Fisher information metric. Chentsov's Theorem [Chentsov \(1982\)](#) tells us that this metric is the only one (up to a multiplicative constant) that is invariant under sufficient statistics. If there are cpds  $p(Y|X)$  and  $q(X|Y)$  such that, for all  $\theta \in \Theta$ , the distribution  $\Pr_\theta(X)$  is unchanged after converting to  $Y$  and back again  $X$  (via  $p$  and  $q$  respectively), as depicted by the following commutative diagram,

$$\begin{array}{ccc} \Theta & \xrightarrow{\Pr} & X \\ \Pr \downarrow & & \uparrow q \\ X & \xrightarrow{p} & Y \end{array}$$

then clearly the family  $\Pr(Y|\Theta) := p \circ \Pr_\theta$  carries the same information about the parameters (and in particular how best to update them) as  $\Pr(X|\Theta)$ . Chentsov's theorem says, that, up to a multiplicative constant, the Fisher information metric is the only metric on  $\Theta$ , as a function of the parameterization  $\Pr$ , which gives identical geometry in both cases.

At each point  $\Theta$ , the components of the Riemannian metric form a matrix—in this case, the Fisher information matrix  $\mathcal{I}(\theta)$ —which allow us to now compute the gradient in the natural geometry from the coordinate derivatives as

$$\text{NGF}[\mathcal{L}]'_\phi(\theta) = -\hat{\nabla}_\theta \mathcal{L}(\theta, \phi) = \mathcal{I}(\theta)^\dagger \nabla \mathcal{L}(\theta, \phi)$$

where  $\mathcal{I}(\theta)^\dagger$  denotes the Moore-Penrose psuedoinverse of the matrix  $\mathcal{I}(\theta)$ , and

$\nabla \mathcal{L}$  is the euclidean gradient of the coordinates, i.e., the vector of partials  $[\frac{\partial \mathcal{L}}{\partial \theta_1}, \dots, \frac{\partial \mathcal{L}}{\partial \theta_n}]^\top$ .

### Expected Utility Maximization Update Rules

Suppose  $\Theta = \Delta X$ , and for each  $\phi \in \Phi$ , we have a utility function  $U_\phi : X \rightarrow \mathbb{R}$  on the underlying set  $X$ . We can then define a learner via exponential decay: which we call the *Boltzmann* leaner for  $U$ :

$$\text{Bolz}[U](\mu, t, \phi) = A \mapsto \frac{1}{\mathbb{E}_\mu[\exp(-\beta U_\phi)]} \int \exp(-\beta U_\phi) \mathbb{1}_A d\mu$$

**Proposition 11.9.** *Boltzmann learners are additive, zero, differentiable, invertable, and commutative.*

link to  
proof

Regarding  $U_\varphi : X \rightarrow \mathbb{R}$  as a potential energy over  $X$ ,  $\text{Bolz}U_\varphi^\beta(\text{Unif})$  is the Boltzmann distribution at inverse temperature (thermodynamic coldness)  $\beta$ . In the thermodynamic analogy, as temperature decreases, one becomes more certain that particles are in their most favorable states. Indeed, using Bolz to update a distribution  $\mu$  given input  $U$  conditions  $\mu$  on the minimizer(s) of  $U$  that have nonzero probability of  $U$ .

**Proposition 11.10.** *The associated vector field is given by  $\text{Boltz}[U]'_\phi \mu = \mu \odot (\mathbb{E}_\mu[U_\phi] - U_\phi)$ .*

link to  
proof

**Proposition 11.11.** *The optimizing update rules for  $\Theta = \Delta X$  whose loss representation is linear (i.e., an expected utility), are precisely the Boltzmann update rules. In particular, the Boltzmann update rule with potential  $U(X)$  is the natural gradient flow update rule for expected value of  $U$ , i.e.,  $\text{Bolz}U = \text{NGF}[\mu \mapsto \mathbb{E}_\mu U]$ .*

**Example 11.5** (Gaussian NGD). Consider the case where  $\Theta = \{(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+\}$  is the half-space of parameters to a Gaussian over some real variable  $X$ , and  $\Phi \cong \mathbb{R}$  consists of possible observations of  $X$ .

One natural loss function is negative log likelihood (differential surprisal) of the observation  $x$  according to your belief state  $\theta = (\mu, \sigma^2)$ :

$$\begin{aligned}\mathcal{L}(x, \mu, \sigma^2) &= -\log \mathcal{N}(x | \mu, \sigma^2) = \frac{1}{2} \log 2\pi\sigma^2 + \frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \\ &= \left\langle \begin{array}{c} \xrightarrow{\mu} \boxed{\mu} \xrightarrow{\mathcal{N}} \boxed{X} \xleftarrow{x} \\ \xrightarrow{\sigma^2} \boxed{\sigma^2} \end{array} \right\rangle.\end{aligned}$$

The Fisher information for a normal distribution is given by

$$\mathcal{I}(\mu, \sigma) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$$

The natural gradient update rule is given by

$$\begin{aligned}F'_x(\mu, \sigma^2) &= -\hat{\nabla}_{\mu, \sigma^2} \mathcal{L}(x, \mu, \sigma^2) \\ &= \mathcal{I}(\mu, \sigma^2)^{-1} \begin{bmatrix} \frac{x - \mu}{\sigma} \\ \frac{-\sigma^2 + (x - \mu)^2}{2\sigma^4} \end{bmatrix} = \begin{bmatrix} \frac{x - \mu}{(x - \mu)^2 - \sigma^2} \\ \frac{(x - \mu)^2 - \sigma^2}{\sigma^2} \end{bmatrix}.\end{aligned}$$

Note that:

- $\mathbb{E}_{x \sim \nu}[F'_x(\mu, \sigma^2)] = \mathbf{0}$  if and only if  $\nu$  has mean  $\mu$  and variance  $\sigma^2$ . Moreover, this point is the unique global attractor. This means that,

1. If observations are drawn from a fixed distribution  $\nu(X)$ , and we repeatedly use  $F$  to update  $\theta = (\nu, \sigma)$  with small confidence  $\epsilon$ , then  $\mu$  will approach the mean  $\mathbb{E}_\nu[X]$  of  $\nu$  and  $\sigma^2$  will approach the variance  $\mathbb{E}_\nu[X^2] - \mathbb{E}_\nu[X]^2$ .
2. If we perform a single high-confidence update on the extended observation  $\varphi \propto \nu$ , in which each  $x$  has relative confidence  $\nu(x)$ , the result will be a

Gaussian with the mean and variance of  $\nu$ , i.e.,

$$\forall \theta. \quad \lim_{c \rightarrow \infty} \Pr_{F_\nu^c(\theta)} = \mathcal{N}(\mathbb{E}_\nu[X], \text{Var}_\nu[X])$$

In this sense, relative confidence acts like probability.

- If we update with the observation  $x = \mu$  of our estimate with confidence  $c$ , the mean is unchanged, and our estimate of the variance becomes the harmonic mean of our previous variance  $\sigma_0^2$  and the inverse confidence  $\frac{1}{c}$ . That is,

$$F_\mu^c(\mu, \sigma_0^2) = \left( \mu, \frac{1}{c + \frac{1}{\sigma_0^2}} \right).$$

Equivalently, the precision of the resulting distribution is the average of the confidence  $c$  and the previous precision  $1/\sigma_0^2$ , which suggests that confidence is of the same type as precision. Note that if  $\sigma_0^2$  is very large, so that our initial beliefs are very uncertain, updating with confidence  $c$  results in variance  $\frac{1}{t}$ . In this sense, the magnitude of confidence acts as the inverse of variance.

△

## 11.4 Further Examples, in Depth

### 11.4.1 Update Rules for Discrete Probabilites

### 11.4.2 Update Rules for Parametric Families

Perhaps the most improtant kind of parametric familiy of functions for the purposes of representing probabilistic information, are exponential families.

Let  $\Theta = \mathbb{R}^n$

Dynamics $F$	Flow $F_A^c(\mu)$	Vector Field $F'_A(\mu)$	Loss $\mathcal{L}^F(\mu, A)$	$F_A^c, F_B^d$ commute?
$LIN$	$(1 - c) \mu + (c) \mu A$	$\mu A - \mu$	$-\log \mu(A)$	if $\mu(A \cap B) > 0$
$LLI$	$\propto \mu^{1-c} (\mu A)^c$ $\propto \mu \mathbb{1}_A$	N/A	N/A	always
$Bolz[1]$	$\propto \mu \cdot \exp(\beta \mathbb{1}_A)$ $\propto \mu \cdot \exp(-\beta \mathbb{1}_{\bar{A}})$	$\mu \odot (\mathbb{1}_A - \mu(A))$ $= \mu(A)(\mu A - \mu)$ $= \mu \odot (\mathbb{1}_A - \mu(A))$	$\mu(\neg A)$	always

Table 11.1: A comparison of different update rules when  $\Theta = \Delta W$  and  $\Phi = 2^W$

Dynamics $F$	Flow $F_q^c(\mu)$	Vector Field $F'_q(\mu)$	Loss $\mathcal{L}^F(\mu, q)$	Properties
$LIN$	$(1 - c) \mu + (c) q$	$q - \mu$	$D(q \parallel \mu)$	
$LLI$	$\propto \mu^{1-c} q^c$	$\mu \odot \left( \log \frac{q}{\mu} + D(\mu \parallel q) \right)$	$D(\mu \parallel q)$	

Table 11.2: Different update rules and their representations when  $\Theta = \Phi = \Delta W$

### 11.4.3 Kalman Filters

## 11.5 Discussion

## APPENDICES FOR CHAPTER 11

### 11.A Further Discussion: Full Confidence, Incremental Confidence, and Independence

#### 11.A.1 Updating with Full Confidence

Since the purpose of  $F_\phi^\top$  is to *fully* incorporate  $\phi$  into our beliefs, two successive updates with the same information ought to have the same effect as a single one. Intuitively, this is because if we have just updated our beliefs to be consistent with the information  $\phi$ , then a second observation of  $\phi$  will require no further alterations of our belief state.

Full-confidence updates are not invertable, and destroy some information about one's prior belief state; because of this, the beliefs that they produce can be easily compressed. Intuitively, this is the benefit of fully trusting information: you can simplify the way you think about things.

**[L6]** Full-confidence updates are idempotent. That is, for all  $\phi \in \Phi$ ,  $F_\phi^\top \circ F_\phi^\top = F_\phi^\top$ .

Once  $\Theta$ ,  $\Phi$ , and any implicit structure in them is specified, there is often a natural choice of full-confidence update rule. To illustrate, we now consider three different rules for different choices of  $\Phi$ . In each case, the possible belief states  $\Theta := \Delta W$  be the set of all probability distributions over a finite set  $W$  of “possible worlds”.

**11.A.1.1 Conditioning.** First, consider the case where observations are events,

i.e.,  $\Phi := 2^W$ . Here, the appropriate rule seems to be conditioning: starting with  $\mu \in \Delta W$ , the conditional measure  $\mu|A$  is given by  $(\mu|A)(B) = \mu(B \cap A)/\mu(A)$ , provided  $\mu(A) > 0$ . Note that  $(\mu|A)|A = \mu|A$ , so conditioning is idempotent. There are well-known issues with conditioning  $\mu$  on  $A$  when  $\mu(A) = 0$ , and so typically this operation is left undefined.

**11.A.1.2 Imaging.** Our second example is the “imaging” approach of Lewis (1976). Suppose that we already have a full-confidence update rule  $f : \Phi \times W \rightarrow W$  on individual worlds, which we interpret as assigning each statement  $\phi \in \Phi$  and  $w \in W$  an element  $f(\phi, w) \in W$  which is the unique “world most similar to  $w$ , in which  $\phi$  is true” (Gardenfors 1982). In this case, idempotence of  $f_\phi : W \rightarrow W$  says the world most similar to  $f(\phi, w)$  in which  $\phi$  is true, is  $f(\phi, w)$  itself. We can then lift  $f$  to a full confidence update rule for  $\Delta W$ , by  $F(\phi, \mu) = A \mapsto \mu(\{w : f(w, \phi) \in A\})$ , intuitively moving the probability of each world  $w$  to  $f(\phi, w)$ . Since  $f$  is idempotent, so is  $F$ .

**11.A.1.3 Jeffrey’s Rule.** Both of the previous approaches establish a single event with certainty. Jeffrey’s rule aims to mitigate this by allowing for “uncertain observations”, in the sense the observations can be probabilistic.

Formally, let  $\Phi$  be the set of pairs  $(X, \pi)$ , where  $X : W \rightarrow S$  is a random variable taking values in a set  $S$ , and  $\pi \in \Delta S$  is a probability on  $S$ . Jeffrey’s rule is then:

$$J((X, \pi), \mu) := \sum_{x \in S} \pi(X=x) \mu|(X=x)$$

When  $\pi$  places all mass on some  $x \in S$ , Jeffrey’s Rule amounts to conditioning on  $X=x$ . For this reason, Jeffrey’s Rule is often thought of as a generalizing conditioning so as to allow for observations of lower confidence. However,  $\pi$  has been fully incorporated into the posterior beliefs (since  $J((X, \pi), \mu)(X) = \pi(X)$ ), while the prior belief  $\mu(X)$  has been destroyed. In addition,  $J_{(X, \pi)}$  is idempotent.

Therefore,  $J$  still establishes its observations with full confidence—it's just that these observations can be probabilistic. We imagine that this mismatch is a result of a historical conflation of confidence with likelihood.

We have seen how full-confidence updates are idempotent; is the reverse true as well? That is, if an update is idempotent, then is it full-confidence?

**Proposition 11.12.** *If  $F$  satisfies ?? L0, L1, L4 and L6, and ??????, while  $\theta, \chi, \phi$  are such that  $F_\phi^\chi(\theta) \neq \theta$  and  $F_\phi^\chi \circ F_\phi^\chi = F_\phi^\chi$ , then  $\chi = \top$ .*

Before we can freely describe the axioms governing intermediate degrees of confidence, we must first clarify a point about sequential observations.

### 11.A.2 Discussion on Incremental Confidence and Independence Assumptions 1

The very idea of our framework is predicated on an implicit assumption:

[L0] There exists a function  $Lrn : \Phi \times [\perp, \top] \times \Theta \rightarrow \Theta$  that captures the updating process.

Historically, ?? L0 has not proved as anodyne as it looks. Some might object that it's not possible to write such a function that is appropriate in all circumstances. For example, Shafer argues for Dempster's rule of combination as a way of incorporating information, but is very careful to emphasize that it ought to be used only on *independent* information, for reasons illustrated below.

**Example 11.6.** You have initial belief state  $\theta_0$ . Now, someone comes up to you and tells you that  $\phi$  is true, a statement that you trust to some intermediate degree of confidence  $c \notin \{\perp, \top\}$ . So, in accordance with ?? L0, you use  $F$  to transform your beliefs, partially incorporating the information to arrive at some belief state  $\theta_1 := F_\phi^c(\theta_0)$ . Immediately afterwards, your friend repeats what they just said:  $\phi$  is true. Your confidence in the statement remains the same, and so according to ?? L0, you again update your beliefs, arriving at  $\theta_2 := F_\phi^c(\theta_1)$ . Except in very special circumstances (e.g., you already know that  $\phi$  is true, or  $c \in \{\perp, \top\}$ ), typically  $\theta_2$ . And yet, it seems your attitude towards  $\phi$  ought to be the same whether you've heard it twice or only once.  $\triangle$

Now, it's important to mention that we're not quite in the same position as Shafer. Shafer was prescribing a concrete representation of  $\Theta$  (a belief function) and a concrete update rule  $F$  (Dempster's rule of combination), and so he needed to defend these choices. We only need to defend something much more modest: we only need to defend the assumption that, if  $\Theta$  and  $\Phi$  properly model the relevant aspects of the scenario at hand, then there exists *some* function  $F$  which performs updates appropriately. Descriptively speaking, we're also in good shape: for synthetic agents, it suffices to point out that learning algorithms represent functions, which given a state, an input, and a number of iterations (confidence), produce an output. And, supposing that  $\Theta$ ,  $\Phi$ , and  $[\perp, \top]$  all capture the relevant respective aspects of a human's belief state, input information, and attitude towards it, how could it be that a human does otherwise? In any case, keeping Example 11.8 in mind, here are three ways to proceed.

**I1. Accept Severe Limitations.** Like Shafer, we could be careful to claim nothing about the belief updating process except in the (unusual) case where

information received is independent. This would be a severe limitation to the theory, and much less necessary than it was for Shafer. Imagine that we are writing code that describes how a synthetic agent updates its beliefs. Shafer's approach is to package any such code with a warning against running it unless assured that observations will always be independent. But independence is notoriously difficult to establish; are we to simply accept that the code will not behave correctly in any realistic scenario?

In practice, many theoretical properties of standard statistical learning algorithms are heavily dependent on independence assumptions (most commonly, that one receives independent, identically distributed samples). This warning label not seem to keep them from being applied in settings where practitioners readily admit samples are not really independent at all—nor indeed performing well empirically in those settings (?).

**12. Appropriately Enrich Domains.** In [Example 11.8](#), it seems obvious that we ought to ignore the second copy of the information, because it has already been accounted for. However, this intuition is highly contingent on the implicit supposition that we *know* the second input to be a replica of the first. Were we ignorant to the nature of the second piece of information, perhaps it would not be so unreasonable to incorporate it again, even without a proof of independence. So, if we would like our agent to make the same decisions that we did, it seems only fair to give it access to the knowledge that we needed to get there. One way of doing this is to extend the belief state so that it also tracks what information has been incorporated.

For [Example 11.8](#) to work, it is critical that we are able to discern that the two inputs were identical. As a result, it seems that the relevant description of the input information was not just  $\phi$ , but a pair  $(\phi, id)$  that also a description

of its identity. It is also critical that we remember the identity of previously incorporated information, so we would also be better off with a belief space  $\Theta$  reflects this. With these two modifications, any commitment function can be straightforwardly modified to avoid the issue in [Example 11.8](#).

We submit that it is always possible to enrich the space of beliefs and observations in this way to track the relevant information, to resolve the issue. With a few more assumptions later on, we will be able to formalize the construction we just alluded to (??).

**I3. An Incremental Interpretation of Confidence.** Finally, we can get around the issue by interpreting a confidence  $c \in [\perp, \top]$  not as an absolute measurement of confidence, but rather an incremental one. This means viewing  $c \in [\perp, \top]$  as the degree of *additional* confidence we have in  $\phi$ , beyond whatever we have already incorporated into our beliefs.

This proposal might be concerning. One might worry that it's harder to make sense of "incremental confidence" than an absolute notion. How ought we to numerically describe the confidence of an update? Suddenly this becomes much more subjective, for to assign a number, not only must we describe how much trust we have in the new information, but we must also take history or current belief state into account. Furthermore, the words "incremental" and "additional" suggest that we will need a formal description of how to aggregate confidences—the very concept of which we will need to defend.

Even modulo these concerns, the incremental interpretation still leaves us in a strictly better place than we were before. To begin, in situations where inputs are independent (i.e., the only cases where we would have been allowed to apply the commitment function according to [Item I1](#)), the two

notions coincide. More explicitly: if the new information  $\phi$  is independent of everything we've previously seen, then an absolute measurement of our confidence in it is no different from a measurement of how much we ought to increment it from having no confidence. Already, though, we can do more. In the situation described by [Example 11.8](#), for instance, the second utterance induce no *additional* confidence ( $\perp$ ), and so applying  $F$  with no confidence clearly gives the desired result of ignoring the new information (per [L1](#)). And even in general, the prospect of having to numerically estimate a fuzzy quantity seems more promising than red tape requiring that  $F$  only be used (in good conscience) on independent information.

We would like to point out that readers who find who find it reasonable to ignore inputs you have no confidence in (per [L1](#)) have implicitly either accepted either [Item I1](#) or [Item I3](#), as the next example shows.

**Example 11.7.** Suppose you first hear  $\phi$  from a partially trusted source, and incorporate it into your beliefs appropriately. Then, the same source sends you a second message, which is obviously spam. In an absolute sense, you now have no confidence ( $\perp$ ) in anything this source tells you, including (in retrospect) both messages. It seems appropriate to excise  $\phi$  from your belief state in response, rather than leaving your belief state unchanged, as [L1](#) would prescribe.

Note that in this scenario, while it seems that we ultimately have no confidence in  $\phi$ , it does not seem to be the case that we have no incremental confidence in  $\phi$ . Rather, the incremental confidence seems to be the inverse of the original confidence.  $\triangle$

We state our results with the incremental interpretation of confidence, with the

understanding that all of our results also admit a more conservative reading, in which confidence is measured absolutely, and also all applications of the function  $F$  are independent.

### 11.A.3 Sequential Observations and Input Independence

Our characterization of full-confidence requires us to think about the effect of making updates with  $F$  in sequence. Does it always make sense to apply  $F$  repeatedly to model sequential observations? ?? L0 says yes. But anodyne as it looks, analogues of ?? L0 have historically been controversial, especially for intermediate values of confidence.

While Shafer endorses Dempster's rule of combination, for example, he is careful to emphasize a limitation: it only applies when combining *independent* pieces information.

**Example 11.8.** We have initial belief state  $\theta_0$ . Now, a friend tells us  $\phi$  is true, which we trust to some intermediate degree  $c \in (\perp, \top)$ . So, in accordance with ?? L0, we use  $F$  to update our beliefs, partially incorporating  $\phi$  arrive at a belief state  $\theta_1 := F_\phi^c(\theta_0)$ . Immediately afterwards, the friend repeats herself: " $\phi$  is true". By ?? L0 we must again update our beliefs with  $F$ , and if our confidence remains the same, then we arrive at  $\theta_2 := F_\phi^c(\theta_1)$ . Unless we were already certain of  $\phi$ , then  $\theta_1 \neq \theta_2$ , yet it seems our beliefs ought to be  $\theta_1$  whether or not we hear her the second time.  $\triangle$

The analogue of Shafer's resolution would be to restrict our endorsement of the  $F$ , so that ?? L0 applies only when receiving information that is, in some sense, independent of our present beliefs. Unfortunately, inputs are seldom

completely independent of our prior beliefs—and worse, there is often no way to know whether they are or not. Yet unknowably entangled information still comes, and we must still choose what to make of it.

In one sense, Shafer’s approach gives us the answer we wanted in [Example 11.8](#), by preventing us from making the second update. But in another, truer sense, it is so restrictive that it says nothing about our final beliefs: the second utterance is not independent, so  $F$  does not apply, and we simply cannot say anything about our beliefs in the end.

An analogous issue about input independence often arises in machine learning, as touched on briefly in [Example 11.3](#). Many theoretical guarantees about the correctness of learning algorithms assume that samples are drawn *independently* from a fixed distribution. Such assumptions underly the standard notion of learnability ([Valiant 1984](#)), and indeed the bulk of statistical learning theory. But at a pragmatic level, many learning theorists take a pragmatic stance: sure, the guarantee only holds for independent inputs, an assumption is almost certainly false—but we use our learning algorithm anyway, and find that in practice inputs are independent enough that there is no need for concern. Obviously, this approach can go awry; in [Example 11.8](#), it amounts to adopting  $\theta_2$  with full acceptance that we will, on occasion, “inappropriately” duplicate updates.

We argue that, by viewing confidence as *incremental* quantity, it is possible to unconditionally endorse ?? L0 and use  $F$  in all cases, without making silly, avoidable mistakes. In [Example 11.8](#), for example, we update using  $F$  upon hearing the second utterance in [Example 11.8](#), (even though it is not independent), but we make this update with zero (incremental) confidence, because it gives us no *additional* information beyond what we already knew. If we restrict our attention

to inputs that are independent of our beliefs, then incremental confidence is no different from absolute confidence. But the real power of this approach comes when we observe an input that is not fully independent of our belief state. For example, what if the second “ $\phi$  is true” in [Example 11.8](#) is not a mistake, but communicates emphasis? It is clearly not independent of what we’ve heard, (one can re-emphasize  $\phi$  only having already articulated  $\phi$ ), but perhaps we can still quantify how much incremental confidence it carries for us.

For the remainder of the chapter, we take this incremental view for convenience, with the understanding that everything we say also admits a conservative reading by restricting ?? L0 only to apply when  $\theta$  and  $(\phi, \chi)$  are independent.

## 11.B More Examples

Next, and example in which we have an additive flow update rule, that is not an optimizing update rule, and hence has no loss representation.

**Example 11.9** (Weighted Average). Suppose we receive vectors  $\phi \in \mathbb{R}^n$ , say estimates of a quantity from different sources. Suppose further that our belief state  $(\mathbf{x}, w) \in \Theta$  consists a current estimate  $\mathbf{x}$  of the quantity of interest, and a weight  $w$  of the total internal confidence in the estimate. In other words:

$$\Theta = \mathbb{R}^n \times [0, \infty]; \quad \text{and} \quad \Phi = \mathbb{R}^n.$$

Updating proceeds by taking a weighted average of the previous estimate and the new input, weighted by their respective confidences, which is captured by:

$$F_y^\beta(\mathbf{x}, w) = \left( \frac{w\mathbf{x} + \beta\mathbf{y}}{w + \beta}, w + \beta \right) \quad \text{and} \quad F_y^\beta(\mathbf{x}, \infty) = (\mathbf{x}, \infty)$$

It is additive, since

$$\begin{aligned}
& F_{\mathbf{y}}^{\beta_2} \circ F_{\mathbf{y}}^{\beta_1}(\mathbf{x}, w) \\
&= \left( \frac{(w + \beta_1)^{\frac{w\mathbf{x} + \beta_1\mathbf{y}}{w + \beta_1}} + \beta_2\mathbf{y}}{(w + \beta_1) + \beta_2}, (w + \beta_1) + \beta_2 \right) \\
&= \left( \frac{w\mathbf{x} + (\beta_1 + \beta_2)\mathbf{y}}{w + (\beta_1 + \beta_2)}, w + (\beta_1 + \beta_2) \right) = F_{\mathbf{y}}^{\beta_1 + \beta_2}(\mathbf{x}, w).
\end{aligned}$$

And it is clearly differentiable, with a simple calculation revealing that  $F'_{\mathbf{y}}(\mathbf{x}, w) = \left(\frac{\mathbf{y}-\mathbf{x}}{w}, 1\right)$ .

Observations:

- The update rule cannot be extended differentiably to states  $\theta = (\mathbf{x}, w)$  with  $w = 0$ . Intuitively, we need to have some estimate with positive confidence to update beliefs in a differentiable way. This is related to the fact that plain empirical risk minimization (ERM) is unstable, but stable with even a small amount of regularization.
- The certainties are given by

$$\lim_{\beta \rightarrow \infty} F_{\mathbf{y}}^{\beta}(\mathbf{x}, w) = (\mathbf{y}, \infty)$$

- $F$  is commutative, invertible, and symmetric with respect to permutation of the dimensions, but it is not conservative: if we had  $U(\mathbf{x}, w, \mathbf{y})$  twice differentiable such that  $\nabla_{\mathbf{x}, w} U = F'$ , then we would have

$$\begin{aligned}
\frac{\partial^2}{\partial w \partial x_i} U &= \frac{\partial}{\partial w} \frac{y_i - x_i}{w} = \frac{x_i - y_i}{w^2}, \quad \text{but} \\
\frac{\partial^2}{\partial x_1 \partial w} U &= \frac{\partial}{\partial x_1} 1 = 0
\end{aligned}$$

violating Clairaut's theorem (which asserts equality of mixed partials). Therefore,  $F'$  cannot be written as the gradient of a function, and so  $F$  is not an optimizing update rule.  $\triangle$

Next, a toy example that showcases an assortment of other features and themes that can be captured with our definition of confidence.

**Example 11.10.** Jugo is an impartial juror. Like the other jurors, she has two buttons in front of her, labeled G and N. Her instructions are to listen to evidence, and press G to increase the probability of a guilty verdict, and N to increase the probability of a not-guilty verdict.

More concretely, the system works as follows. There are  $J$  jurors, labeled  $\{1, \dots, J\}$ ; let  $\text{pressed}(j, B, t)$  be a variable that is equal to one if juror  $j \in$  is pressing button B button at time  $t$ , and zero otherwise. The “belief state” of this automated system is a single number  $g \in [0, 1]$ , representing the probability of a guilty verdict. When a single juror presses G,  $g$  approaches 1 exponentially, and if they instead press N,  $g$  decays to zero. In the first case (G is pressed) the system evolves according to  $\frac{dg}{dt} = (1 - g)$  while in the second,  $\frac{dg}{dt} = -g$ . The first is the vector field associated with the G button, and the second is the vector field associated with N. The total effect of all buttons is then the sum of that of all buttons across all vector fields, when they are active:

$$\frac{dg}{dt} = \sum_{j=1}^J \text{pressed}(j, G, t)(1 - g) - \text{pressed}(j, N, t)(g),$$

so that  $g$  exponentially approaches 1 when more G buttons are pressed than N buttons, and symmetrically, exponentially approaches 0 when more N buttons than G buttons are pressed. At the end of the trial, the defendant is convicted with probability equal to the final value of  $g$ .

Let  $\phi$  represent a piece of evidence suggesting guilt, presented by the prosecution from time  $t_1$  to time  $t_2$ , and suppose for now that only buttons labeled G are

pressed in this interval. The system measures  $j$ 's confidence in  $\phi$  by

$$w_j := \int_{t_1}^{t_2} G_j(t) dt = \text{total time } j \text{ presses G during } \phi,$$

Note that  $w_j = 0$  if and only if  $j$  does not press any buttons, which (a) indicates that  $j$  does not trust the evidence  $\phi$ , and (b) communicates this fact to the system, by telling it to ignore the evidence. Note that this is an additive representation of confidence, since pressing the button for four seconds, and then three more later, is by definition the same as pressing it for seven. While the maximum possible confidence of  $w_j$  is  $(t_2 - t_1)$ , this system does not allow a juror to express *full* confidence in  $\phi$  because no finite amount of G-pressing will result in a guilty verdict with probability one; it is always possible to increase the value of  $g$  through additional evidence.

Altogether, the system's confidence in  $\phi$  can be measured by as the unique value  $W$  for which

$$\int_{t_1}^{t_2} W(1 - g(t)) dt = g(t_2) - g(t_1),$$

which, so long as only G buttons are pressed, equals  $W := \sum_j w_j$ , so this measure of confidence is additive across jurors as well as across time. This is appropriate, since the jurors are independent and not communicating with each other. As before,  $W = 0$  if and only if no juror presses any buttons between times  $t_1$  and  $t_2$ , indicating zero trust leant to  $\phi$ . In such a case, the system ignores  $\phi$  in updating its beliefs. And just as no individual juror can send a full-confidence update to the system, the system cannot receive a full-confidence from the jurors as a whole.

The picture gets significantly more complicated if we consider the possibility that jurors might press the N button. For example, if  $\phi$ , which was intended as

evidence of guilt, has the effect of getting jurors to press N, there is a sense in which they have *negative* confidence in  $\phi$ , since the belief update happened in the opposite direction of what  $\phi$  represents; rather than *no* trust, this is represents *distrust*. Small negative updates are always possible except at the boundary of belief space, but in this chapter, we focus almost entirely on positive confidence updates.

The introduction of the second button also uncovers a significant source of complexity: unlike Examples 11.1 to 11.3, the order that evidence is presented matters, when there is more than one possible response to it. Evidence presented later has a larger effect, meaning that this system exhibits a recency bias.

Now consider a variant of this system that does not trust all jurors equally; rather, it trusts each juror  $j$  to a degree  $\beta_j \in [0, \infty]$ , and now  $g$  evolves according to

$$\frac{dg}{dt} = \sum_{j=1}^J \beta_j (G_j(t)(1-g) - gN_j(t)).$$

In this case, the system can be said to have trust  $\beta_j$  in juror  $j$ , since  $j$ 's buttons are ignored when  $\beta_j = 0$ . When  $\beta_j = \infty$  (an expression of full confidence in  $j$ ),  $g$  immediately jumps to 0 when  $j$  presses N, or to 1 if  $j$  presses G (unless canceled by another full-confidence juror pressing the opposite button). If all jurors have full confidence, then the verdict of this system is a majority vote at the last moment a button was pressed. Thus, the weights attached to weighted combinations are (additive) expressions of confidence as well.  $\triangle$

[Example 11.10](#) illustrates how a (sufficiently nice) vector field, which is simpler than a smooth path for every starting point, is enough to define an additive notion of confidence, via its integral curves. It may seem strange to define confidence via a vector field, which does not mention confidence at all—but in a sense, it

works because a vector field captures precisely everything about the update *except* for the confidence. We do this formally in [Section 11.3.1](#).

**Example 11.11** (The General Kalman Filters). Suppose we are interested in modeling a dynamical system whose state is a vector  $\mathbf{x} \in \mathbb{R}^n$ , and we receive observations  $\mathbf{z}$  that are assumed to be a linear function of  $\mathbf{x}$ , plus Gaussian noise. In many engineering disciplines, the standard way to track this information is the [Kalman](#) filter [1960]. It prescribes belief state  $(\hat{\mathbf{x}}, P)$ , where  $\hat{\mathbf{x}} \in \mathbb{R}^n$  is our current estimate of  $\mathbf{x}$ , and  $P \in \mathbb{R}^{n \times n}$  is a covariance matrix encoding our certainty  $\hat{\mathbf{x}}$ . (Intuitively, this amounts to a belief that  $\mathbf{x} \sim \mathcal{N}(\hat{\mathbf{x}}, P)$  is normally distributed with mean  $\hat{\mathbf{x}}$  and variance  $P$ .)

Suppose we now receive an observation  $\mathbf{z} = H\mathbf{x} + \xi \in \mathbb{R}^m$ , from a particular sensor, where  $H \in \mathbb{R}^{m \times n}$  is called the sensor's *observation matrix*, and the noise  $\xi \sim \mathcal{N}(0, R)$  has known variance  $R$ . For example, perhaps  $\mathbf{x} = (x_1, x_2)$  is the location of an aircraft, and we have a sensor that measures its first coordinate (plus noise  $\xi$ ). This sensor's observation matrix  $H$  then represents the map  $(x_1, x_2) \mapsto x_1$ , and we observe  $\mathbf{z} = x_1 + \xi$ . How should we update our beliefs in response?

Once again, the right answer depends on how much trust we have in the sensor, and ranges from ignoring  $\mathbf{z}$  to deferring entirely to it. In our example, the latter means replacing  $x_1$  with  $z$ , and altering  $P$  so that  $x_1$  has the same variance as the sensor, and is uncorrelated with  $x_2$ . In general, the Kalman filter measures confidence in the observation with a matrix  $K \in \mathbb{R}^{n \times m}$  called *Kalman gain*, which

is used to compute posterior beliefs  $(\hat{\mathbf{x}}', P')$  according to:

$$\begin{aligned}\hat{\mathbf{x}}' &= \hat{\mathbf{x}} + K(\mathbf{z} - H\hat{\mathbf{x}}) \\ P' &= (I - KH)^T P(I - KH) + KRK^T.\end{aligned}$$

Often introduced as a “blending factor”,  $K$  is similar to  $\alpha$  in [Example 11.1](#), especially for a sensor that directly measures a one-dimensional quantity  $x$  (i.e., with  $H = 1$ ). In this case, our belief state is a pair  $(x, \sigma^2) \in \mathbb{R} \times [0, \infty)$ , and our posterior after observing  $z$  simplifies to:

$$\begin{aligned}x' &= (1 - K)x + (K)z \\ (\sigma^2)' &= (1 - K)^2 \sigma^2 + (K)^2 R.\end{aligned}$$

Observe how  $K$  linearly interpolates between our prior estimate of the mean and the new observation, and in a sense “quadratically” between our prior variance estimate  $\sigma^2$  and the variance  $R$  of the noise  $\xi$ .

More directly than in previous examples, we can also say something prescriptive about how best to select a degree of confidence. This is made possible by two assumptions:

- We know how observations  $\mathbf{z}$  are generated, and, in particular, can objectively quantify their reliability, with the variance  $R$  of the added noise  $\xi$ .
- Our objective is to select  $K$  so as to minimize (the sum of the eigenvalues of) our resulting variance  $P' = \mathbb{E}_{x \sim \mathcal{N}(\hat{\mathbf{x}}, P)}[(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^T]$ , which happens to be the multivariate analogue of mean-square error.

Under these assumptions, the optimal Kalman gain is

$$K = PH^T(HPH^T + R)^{-1} \tag{11.3}$$

(Brown and Hwang 1997, p. 146), and so  $K$  is typically chosen this way in practice (Becker 2003).

With this in mind, let's revisit the state update equations in the of extreme values of confidence. If  $K = 0$ , which is prescribed by (11.3) iff the noise  $\xi$  has unbounded variance (in every coordinate), then the update leaves  $(\hat{x}, P)$  unchanged. Intuitively, so much noise is added to observations  $z$  that we have no trust that incorporating them will improve our estimate of  $x$ . At the other extreme, if there is no noise ( $R = 0$ ), then the optimal Kalman gain  $K = H^+$  is the pseudo-inverse of  $H$ , and we end up with posterior beliefs  $(\hat{x}', P')$  such that  $H\hat{x}' = z$  and  $HP'H^\top = \mathbf{0}$  and so fully trusts the new observation  $z$  (but is otherwise as close to  $(\hat{x}, P)$  as possible).  $\triangle$

Just as mentioned in the discussion after Example 11.4, the general Kalman filter ?? features three distinct kinds of uncertainty:

1.  $R$ , the objective (un)reliability of the observation  $z$  (as measured by variance), a feature of the environment;
2.  $P$ , the subjective “internal confidence” in the current estimate  $x$ , a feature of the current belief state.
3.  $K$ , the subjective confidence in the observation  $z$ , a feature of what one knows about where  $z$  comes from and how it relates to one’s confidence.

## 11.C Extra Properties of Update Rules

### Invertable Update Rules

[L7] For all  $\phi \in \Phi$ , and  $\beta \in \mathbb{R}$ , the update  $F_\phi^\beta : \Theta \rightarrow \Theta$  is invertable. (**Invertability**)

**Proposition 11.13.** *If  $F$  is a differentiable and invertable update rule (i.e., satisfies L1 and L7 and ???), then for all  $\beta \in \mathbb{R}$ ,  $\phi \in \Phi$ , the function  $F_\phi^\beta : \Theta \rightarrow \Theta$  is a diffeomorphism, and its inverse is given by  $F_\phi^{-\beta}$ , in the sense that*

$$F_\phi^{-\beta}(F_\phi^\beta(\mu)) = \mu = F_\phi^\beta(F_\phi^{-\beta}(\mu)).$$

**Corollary 11.13.1.** *If for any  $\beta < \infty$  there exist  $\mu, \phi, A$  such that  $\mu(A) > 0$  but  $F_\phi^\beta(\mu)(A) = 0$ , then  $F$  is not invertable.*

## 11.D Proofs

**Proposition 11.14.** For every learner  $Lrn$ , there exists a believer  $Bel$  such that the pair  $(Lrn, Bel)$  satisfy LB1–3

[ link to proof ]

*Proof.* Given  $Lrn : \Phi \times [\perp, \top] \times \Theta \rightarrow \Theta$ , define

$$Bel(\theta, \phi) := \begin{cases} \top & \text{if } \exists \chi. Lrn(\phi, \chi, \theta) = \theta \\ \perp & \text{otherwise} \end{cases}$$

□

**Theorem 11.3.** If  $Lrn$  satisfies L1, L3 and L4 and ?? (but possibly not L2) then there exist  ${}^+Lrn$  and a continuous function  $g : \Phi \times [\perp, \top] \times \Theta \rightarrow [0, \infty]$  such that

$$\forall \theta, \phi, \chi. \quad Lrn(\phi, \chi, \theta) = {}^+Lrn(\phi, g(\phi, \chi, \theta), \theta) \quad \text{and} \quad {}^+Bel(\phi, \theta) = g(\phi, Bel(\phi, \theta), \theta).$$

Furthermore,  $({}^+F, g)$  is unique up to a multiplicative factor in the output of  $g$ .

*Proof.*

□

**Proposition 11.9.** Boltzmann learners are additive, zero, differentiable, invertible, and commutative.

*Proof.* **Commutativity.** For some normalization factors  $Z, Z', Z''$ , we have:

$$\begin{aligned} F_\phi^\beta(F_{\phi'}^{\beta'}(\mu)) &= F_\phi^\beta\left(\frac{1}{Z} \mu \exp(-\beta' c_{\phi'})\right) \\ &= \frac{1}{Z'} \frac{1}{Z} \mu \exp(-\beta' c_{\phi'}) \exp(-\beta c_\phi) \\ &= \frac{1}{Z''} \mu \exp(-\beta' c_{\phi'} - \beta c_\phi) \end{aligned}$$

which is the same expression when we exchange  $(\phi, \beta)$  and  $(\phi', \beta')$ .

□

**Proposition 11.10.** *The associated vector field is given by  $\text{Boltz}[U]_\phi' \mu = \mu \odot (\mathbb{E}_\mu[U_\phi] - U_\phi)$ .*

*Proof.* Let  $f(X) := \exp(-\beta U(X, \varphi))$ , and  $g(X) := U(X, \varphi)$ .

$$\text{Boltz}'_\varphi \theta = \frac{\partial}{\partial \beta} \text{Boltz}_\varphi^\beta(p) \Big|_{\beta=0}$$

⟨ TODO: finish typesetting algebra ⟩

$$\begin{aligned} &= x \mapsto p(x) \frac{f(x)}{\mathbb{E}_p[f]} \left( \mathbb{E}_p \left[ \frac{f}{\mathbb{E}_p[f]} g \right] - g(x) \right) \Big|_{\beta=0} \\ &= \frac{pf}{\mathbb{E}_p[f]^2} (\mathbb{E}_p [fg] - g \mathbb{E}_p[f]) \Big|_{\beta=0} \\ &= x \mapsto p(x)(\mathbb{E}_p[g] - g(x)) \quad \text{since } f(X) = 1 \text{ when } \beta = 0 \end{aligned}$$

As a sanity check, note that the sum over all components is

$$\sum_{x \in X} ((\text{Boltz } U)'_\varphi \theta)_x = \sum_{x \in X} p(x)(\mathbb{E}_p[g] - g(x)) = \mathbb{E}_p[\mathbb{E}_p[g]] - \mathbb{E}_p[g] = 0,$$

so indeed it lies within the tangent space.  $\square$

On its own, so long as we have the freedom to choose  $[\perp, \top]$ , L2 has no teeth.

**Proposition 11.15.** *If  $F : [\perp, \top] \rightarrow (\Phi \rightarrow (\Theta \rightarrow \Theta))$  satisfies L1 and L6, then we can construct a new update function for  $\Theta$  on  $\Phi$ , that behaves in exactly the same way, but accepts confidences in a different confidence domain  $[\perp, \top]',$  and satisfies ??.*

*Proof.* Consider the new confidence domain

$$[\perp, \top]' := \left\{ \text{finite lists } [c_1, \dots, c_n] \text{ with each } c_i \in [\perp, \top], \quad ::, \quad [], \quad [\top] \right\},$$

whose group operation “ $::$ ” is list concatenation, except that it collapses instances of  $\top$ , i.e.,

$$[c_1, \dots, c_n] :: [d_1, \dots, d_m] := \begin{cases} [\top] & \text{if } \top \in \{c_1, \dots, c_n, d_1, \dots, d_m\} \\ [c_1, \dots, c_n, d_1, \dots, d_m] & \text{otherwise.} \end{cases}$$

Concatenating the empty list  $[]$  on either side has no effect, by construction, for all  $L \in [\perp, \top]',$  we have  $[\top] :: L = [\top] = L :: [\top],$  and  $::$  is clearly associative, so  $[\perp, \top]'$  is also a confidence domain.

The new update rule for this confidence is given by:

$$AF_{\phi}^{[c_1, \dots, c_n]}(\theta) := (F_{\phi}^{c_n} \circ \dots \circ F_{\phi}^{c_1})(\theta).$$

$AF$  has the same behavior as  $F$  on the elements that correspond to the original confidence domain, since  $AF_{\phi}^{[c]}(\theta) = F_{\phi}^c(\theta),$  and it is additive by construction, since

$$\begin{aligned} AF_{\phi}^{[c_1, \dots, c_n]}(AF_{\phi}^{[d_1, \dots, d_m]}(\theta)) &:= F_{\phi}^{d_m} \circ \dots \circ F_{\phi}^{d_1}(F_{\phi}^{c_n} \circ \dots \circ F_{\phi}^{c_1}(\theta)) \\ &= (F_{\phi}^{d_m} \circ \dots \circ F_{\phi}^{d_1} \circ F_{\phi}^{c_n} \circ \dots \circ F_{\phi}^{c_1})(\theta) \\ &= AF_{\phi}^{[c_1, \dots, c_n, d_1, \dots, d_m]}(\theta) \\ &= AF_{\phi}^{[c_1, \dots, c_n] :: [d_1, \dots, d_m]}(\theta). \end{aligned}$$

□

## CHAPTER 12

### RELATIVE ENTROPY SOUP

⟨ INCOMPLETE ⟩

Based on the results of <https://www.cs.cornell.edu/~oli/files/papers/lafi.pdf>; will show that PDGs also arise from using relative entropy as the loss function in the previous chapter, in the update setting where  $\Theta = \Delta\mathcal{V}\mathcal{X}$ .

## CHAPTER 13

### THE CATEGORY THEORY OF PDGS

In this chapter, we investigate a more abstract perspective on PDGs, that was part of the original inspiration for PDGs in the first place. At first Bayesian networks appear to be almost diagrams in the categorical sense—if the BN is a

#### 13.1 A Primer on Category Theory

Category theory is a mathematical interlingua that captures the essential form of many arguments across mathematics. Sometimes (lovingly) called “abstract nonsense”, category theory is often seen as extremely abstract meta-mathematics. Nevertheless, the basics more concrete and simpler than one might imagine. At its core, it’s essentially just the mathematic underpinnings of typed composition.

**Definition 13.1** (category). A *category*  $\mathcal{C}$  consists of four pieces of data:

- a collection of *objects*  $\text{ob}_{\mathcal{C}}$ ;
- a collection of *morphisms*  $\text{Hom}_{\mathcal{C}}(X, Y)$ , also written  $\mathcal{C}(X, Y)$ , for each pair of objects  $(X, Y) \in \text{ob}_{\mathcal{C}}^2$ ;
- a *composition operator*  $\circ_{X,Y,Z} : \mathcal{C}(Y, Z) \times \mathcal{C}(X, Y) \rightarrow \mathcal{C}(X, Z)$  for each triple  $(X, Y, Z) \in \text{ob}_{\mathcal{C}}^3$ , that is written inline (i.e.,  $f \circ g$  instead of  $\circ(f, g)$ ), and is associative, i.e.,  $(f \circ g) \circ h = f \circ (g \circ h)$ ;
- a special *identity element*  $\text{id}_X \in \mathcal{C}(X, X)$  for each object  $X \in \mathcal{C}$ , satisfying  $\text{id}_X \circ f = f$  and  $g \circ \text{id}_X = g$  for any morphism  $f \in \mathcal{C}(Y, X)$  or  $g \in \mathcal{C}(X, Y)$  for some  $Y \in \text{ob}_{\mathcal{C}}$ . □

Common examples of categories include those in [Table 13.1](#).

Category $\mathcal{C}$	Objects $\text{ob } \mathcal{C}$	Morphisms $\text{Hom}_{\mathcal{C}}$
$\mathbb{S}\mathbf{et}$	sets	functions
$\mathbb{R}\mathbf{el}$	sets	relations
$\mathbb{T}\mathbf{op}$	topological spaces	continuous maps
$\mathbb{D}\mathbf{iff}$	smooth manifolds (with boundary or corners)	smooth maps

Table 13.1: A few of the most recognizable categories

Each of these categories can also be much more combinatorial in nature. We will be much more interested in dinkier categories. Here are some more extreme kinds of categories:

- A category with one object is just a monoid—observe that  $\circ$  is associative and has an identity.
- At the opposite extreme, a category with only identity morphisms is just a collection of objects.
- A category with at most one morphism between any two objects is a preorder—in this case we write  $a \leq b$  iff there is a morphism from object  $a$  to object  $b$ ; the relation is reflexive because of the identity, and transitive because of composition.

The kinds of categories we are most interested in, however, are the ones generated by directed graphs.

**Definition 13.2** (free category generated by a graph). If  $G = (N, A)$  is a directed (multi) graph with nodes  $N$  and arrows  $A$ , the *free category generated by G* is the category  $G^*$ , whose objects are the elements of  $N$ , and whose set of morphisms

from  $x$  to  $y$ , for  $x, y \in N$ , is the collection of paths from  $x$  to  $y$ . That is,

$$\text{ob}_{G^*} = N,$$

$$G^*(x, y) = \left\{ \text{sequences } \langle a_1, \dots, a_n \rangle \mid \begin{array}{l} n \in \mathbb{N}, \quad n > 0 \Rightarrow (S_{a_1} = x \wedge T_{a_n} = y), \\ \forall i \in \{1, \dots, n-1\}. T_{a_i} = S_{a_{i+1}} \end{array} \right\},$$

with composition given by sequence concatenation, and the identity being the empty sequence.  $\square$

The superscript-star notation has some standard meanings throughout mathematics, and this construction in [Definition 13.2](#) reduces to several of them in the appropriate contexts.

**(directed multi)**

- A (multi) graph  $G = (\{\ast\}, A)$  with one vertex can be identified with its arc set  $A$ . Every arrow has the same type  $(\ast \rightarrow \ast)$ , and so a path is a sequence  $\langle a_1, a_2, \dots, a_n \rangle$  where each  $a_i \in A$ . So in this case,  $G^*$  (as given by [Definition 13.2](#)) coincides with the familiar set of strings  $A^*$  over the alphabet  $A$ .
- Let  $R \subseteq V \times V$  be a binary relation on  $V$ . Then the transitive closure of  $R$ , often denoted  $R^*$ , is the reachability relation generated by  $R$ . That is,  $(u, v) \in R^*$  if and only if there is a path  $\langle u=u_1, \dots, u_n=v \rangle$  with each  $(u_i, u_{i+1}) \in R$ . Equivalently, we can view  $R$  as a graph  $G = (V, R)$  by regarding each  $(i, j) \in R$  as an arrow  $i \rightarrow j$ . The free category  $G^*$  generated by these arrows (per [Definition 13.2](#)) has an arrow from  $u$  to  $v$  (i.e.,  $G^*(u, v) \neq \emptyset$ ) iff  $(u, v) \in R^*$ .
- A (directed) (multi) graph  $G$  on  $n$  vertices also has an adjacency matrix  $A := \mathbb{A}_G \in \mathbb{N}^{n \times n}$ . Square matrices over a semiring also have notion of a star,

given by:

$$A^* = \sum_{n=0}^{\infty} A^n \in \overline{\mathbb{N}}^{n \times n}, \quad \text{where } \overline{\mathbb{N}} := \mathbb{N} \cup \{\infty\}.$$

And, yet again, we have  $\#G^*(i, j) = (\mathbb{A}_G^*)_{i,j}$

■

The second category of interest to us is that of sets and probabilistic functions between them. We can use the theory in Sections 2.3.1 and 2.3.2 to make this precise.

**Definition 13.3** (category of measurable spaces and Markov kernels). Let  $\mathbb{Stoch}$  be the category whose objects are measurable topological spaces with a base measure, and whose morphisms are Marov kernels that are absolutely continuous with respect to the base measure. Concretely, the objects of  $\mathbb{Stoch}$  are pairs  $(\mathcal{X}, \lambda)$ , where  $\mathcal{X}$  is a measurable topological space, and  $\lambda$  is a strictly positive and  $\sigma$ -finite measure on  $\mathcal{X}$ . The collection of morphisms from  $(X, \mathcal{F}_X, \lambda_X)$  to  $(Y, \mathcal{F}_Y, \lambda_Y)$  is the set of Markov Kernels  $\kappa : \mathcal{X} \rightarrow \mathcal{Y}$  such that  $\kappa(x, -) \ll \lambda_Y$  for all  $x$ . The reason we require this is so that the Radon-Nikodym derivative  $\frac{d\kappa(x)}{d\lambda}$ , i.e., the unique  $\mathcal{F}_Y$ -measurable function satisfying

$$\forall x. \forall A \in \mathcal{F}_Y. \quad \kappa(x, A) = \int_A \frac{d\kappa(x)}{d\lambda} d\lambda, \quad \text{exists.}$$

Composition in  $\mathbb{Stoch}$  is given by Lebesgue Integration: for Markov Kernels  $p(Y|X) : \mathcal{X} \rightarrow \mathcal{Y}$  and  $q(Z|Y) : \mathcal{Y} \rightarrow \mathcal{Z}$ , define  $(p \circ q) : \mathcal{X} \rightarrow \mathcal{Z}$  (i.e.,  $p \circ q : X \times \mathcal{F}_Z \rightarrow [0, 1]$ ) by:

$$(p \circ q)(x, U) := \int_{\mathcal{Y}} q(-, U) dp(x, -).$$

This typechecks because  $q(-, U)$  is a  $\mathcal{Y}$ -measurable function, and  $p(x, -)$  is a measure on  $\mathcal{Y}$ . We must also prove that the result is a Markov Kernel, which we do below. The identities are given by

$$\text{id}_{\mathcal{X}}(x, U) = \begin{cases} 1 & \text{if } x \in U \\ 0 & \text{otherwise} \end{cases}.$$

□

## 13.2 A Categorical Picture of PDGs

Note that  $\mathcal{V}$  is implicit in  $\mathbb{P}$ . The two can be expressed as a functor, which is arguably the most compact definition of an (unweighted) PDG. An *unweighted PDG*  $\langle \mathcal{V}, \mathcal{P} \rangle$  over a structure  $(\mathcal{N}, \mathcal{A})$  is just a functor

$$\mathbb{P} : \mathcal{A}^* \rightarrow \text{Stoch}$$

whose action on objects  $\mathcal{N}$  is  $X \mapsto \mathcal{V}X$ , and whose action on the generating morphisms  $X \xrightarrow{a} Y \in \mathcal{A}$  is written  $\mathbb{P}_a(Y|X)$ . We ~~drop the the symbol  $\mathcal{V}$  in this context, using only the symbol  $\mathbb{P}$ , because  $\mathcal{V}$  can be recovered by the action on the identity morphisms of  $\mathcal{A}^*$~~ . Given small category  $J$  (such as the free category generated by a graph), a functor  $F : J \rightarrow \mathcal{C}$  is often called a *diagram of  $\mathcal{C}$*  (of shape  $J$ ). Therefore, an unweighted PDG is a diagram of the Stoch, of shape generated by its underlying *hyprgraph*. In addition to probabilities, a PDG also contains confidences  $\beta = \{\beta_a\}_{a \in \mathcal{A}}$  about the reliability of those probabilities.

add (in  
parens)  
"weighted"

We now pursue a clean categorical picture of quantitative PDGs. At a quantitative level, positive structural weights can be captured by negative observational weights. This is because the gradient of  $-\hat{\nabla}_\mu H_\mu(Y|X)$ , the gradient of the structural loss corresponding to a hyperarc  $X \xrightarrow{a} Y$ , is the same as

$+\hat{\nabla}D(\mu(X, Y) \parallel \mu(X)\lambda_Y(Y))$ , the gradient of the observational loss corresponding to a uniform distribution. Furthermore, the weight  $\beta_a$  may be absorbed into the cpd  $\mathbb{P}_a$  by dropping the requirement that measures be normalized. This is because the pair  $(p(Y|X), \beta)$  as a can be encoded<sup>1</sup> as a single conditional measure  $(1 - e^{-\beta})p(Y|X)$  losslessly, because  $p(Y|X)$  can be reconstituted by renormalizing, and  $\beta = -\log(1 - k)$  can be recovered from the normalization constant  $k$ . The only exception is when  $\beta = 0$ , but in this case the cpd does not matter semantically, and so if anything it is a bonus that this representation identifies all cpds supplied with confidence  $\beta = 0$ .

Furthermore, with this representation, the effect of composition is very compelling. Suppose we compose  $p(Y|X)$  with confidence  $\beta_1$  with  $q(Z|Y)$  with confidence  $\beta_2$ , where both  $\beta_1, \beta_2 \in [0, \infty]$ . Then the composite

$$\begin{aligned} r(Z|X) &= \int_Y (1 - e^{-\beta_2})p(Z|Y)(1 - e^{-\beta_1})dp(Y|X) \\ &= (1 - e^{-\beta_1} - e^{-\beta_2} + e^{-\beta_1-\beta_2}) (q \circ p)(Z|X) \\ &\approx (1 - \exp(-\min\{\beta_1, \beta_2\})) (q \circ p)(Z|X). \end{aligned}$$

In particular, the composite will be fully trusted iff both components are  $\beta_1 = \beta_2 = \infty$ , and if either has confidence zero, then the composite will also.<sup>2</sup> As a result, all data in a PDG  $(\mathcal{A}, \mathcal{N}, \alpha, \mathcal{V}, \mathbb{P}, \beta)$  may be specified together with a single functor

$$m : \mathcal{A}^* \rightarrow \mathbb{M}\text{eas}_\Delta. \quad (13.1)$$

---

<sup>1</sup>However, there is no way to combine  $\beta$  with  $p$  that results in a quantity  $q$  (independent of  $\mu$ ) that can be plugged directly into the ordinary expression for KL divergence:

$$\beta \log \frac{\mu}{p} = \log \frac{\mu}{q} \implies q = \mu^{1-\beta} p^\beta.$$

<sup>2</sup>Keep in mind that, even if  $p(Y|X)$  and  $q(Z|Y)$  are both marginals of a shared distribution  $\mu(X, Y, Z)$ , and this is known with extreme confidence, their composite will only be correct if the information is somehow “independent”. This is where I think  $\alpha$  should enter the picture, ideally.

In other words, a PDG is a *diagram*, in the usual categorical sense, of conditional (sub)distributions between measurable spaces.

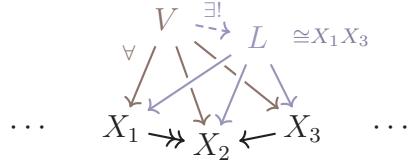
This connection induces a number of category-theory flavored questions about PDGs:

1. A PDG  $m : \mathcal{A}^* \rightarrow \text{Meas}_\Delta$  is a diagram in the category of subprobability distributions. When does it have a limit? What about a colimit? What do limits and colimits of PDGs mean?
2. If PDGs are functors, what are the natural transformations between them? What do they mean?
3. How does inconsistency arise in this categorical picture?
4. Can we study qualitative PDGs separately in this picture? Why are  $\alpha$  combined with  $\beta$  if the former are purely qualitative?
5. PDGs can be given semantics in more than one way, in principle — relative entropy is a natural choice, ~~but, even then, it can be used in either direction.~~ Yet this functorial definition of a PDG does not contain this information. So is there any way it can possibly interact with relative entropy that defines the semantics? If so, what is the categorical picture of the role of relative entropy?
  - For general loss functions (e.g., reverse KL), how does this picture interact with confidence functions?

### 13.2.1 Limits

A *cone*  $(X, \phi)$  over a diagram  $D : \mathcal{J} \rightarrow \mathcal{C}$  is an object  $X \in \text{ob } \mathcal{C}$ , together with an indexed family of morphisms  $\phi_J : X \rightarrow DJ$  for each object  $J \in \text{ob } \mathcal{J}$ , such that, for all morphisms  $f \in \mathcal{J}(A, B)$ ,

... ⟨ INCOMPLETE ⟩



**Definition 13.4.** Let  $\mathbf{m}$  be a PDG, with variables  $\mathcal{X}$ . The *local marginal polytope*

$$\mathbb{L}(\mathbf{m}) := \left\{ \boldsymbol{\mu} = \{\mu_X \in \Delta \mathcal{V}X\}_{X \in \mathcal{X}} \mid \forall S \xrightarrow{a} T \in \mathcal{A}. \mu_T = \mathbb{P}_a \circ \mu_S \right\}. \quad (13.2)$$

consists of all marginals over the variables  $\mathcal{N}$  that are locally consistent with all arcs.  $\square$

**Example 13.1.** 1. Suppose  $\mathbf{m}$  is the PDG representation of a joint distribution

$\mu$ , then  $\lim \mathbf{m} = \mu$

2. if  $\mathbf{m}$  has a single arc  $\mathcal{A} = X \rightarrow X$  associated with a cpd  $q(X|X)$ , then  $\lim \mathbf{m} = \text{Ext}\{\mu : \mu = q \circ \mu\}$  is the set of extreme points over the stationary distributions of  $q$ .

$\triangle$

Our next results are sensitive to the particulars of the PDG encoding as a functor. Let  $\mathbf{m}^{+\mathcal{X}} := \mathbf{m} \cup \{\mathbf{X} \rightarrow \mathbf{Y}\}_{\mathbf{Y} \subset \mathbf{Y} \subseteq \mathcal{X}}$  be the PDG  $\mathbf{m}$  augmented with

additional structure describing the relationships between all subsets of variables. That is,  $\mathcal{A}^*$ , as the free hypergraph over nodes  $2^{\mathcal{N}}$ , complete with structural coherence maps.<sup>3</sup> Let  $\mathbf{m}^{hold}$  be the PDG where  $p(\mathbf{Y}|\mathbf{X})$  is really attached to a hyperarc  $\mathbf{X} \rightarrow \mathbf{X} \cup \mathbf{Y}$ , implicitly the identity along  $\mathbf{X} \setminus \mathbf{Y}$ . Write  $\mathbf{m}^{+\mathcal{X},hold}$  for the PDG with both alterations.

**Theorem 13.1.** *Suppose  $\mathbf{m}$  is a PDG in which every arc has full confidence. Then,*

1.  $\text{Cones}(\mathbf{m}, 1) \cong \mathbb{L}(\mathbf{m});$
2.  $\text{Cones}(\mathbf{m}^{+\mathcal{X},hold}, 1) \cong \{\mathbf{m}\}.$
3.  $\lim \mathbf{m}^{+\mathcal{X},hold} = \text{Ext}\{\mathbf{m}\}$ , where  $\text{Ext}(S)$  is the set of extreme points of a set  $S$  (e.g., the vertices of  $S$ , when  $S$  is a polytope).
4.  $\lim \mathbf{m} = \text{Ext } \mathbb{L}(\mathbf{m}).$

*Proof.* Part 1 is immediate; it just points out that the local marginal polytope, defined in the graphical models literature, is the limit in this context.

Now for part 2. A cone over  $\mathbf{m}^{+\mathcal{X},hold}$  with vertex 1 is a collection of distributions  $\{\mu_X(X)\}_{X \in \mathcal{X}}$  such that, for all  $S \xrightarrow{a} T \in \mathcal{A}$ ,  $\mu_T(S, T) = \mathbb{P}_a(T|S)\mu_S(S)$ . (This familiar notation is not problematic if  $S \cap T = \emptyset$ , but otherwise we mean  $\mu_T(S, T) = \int_S \mathbb{P}_a(S, T|s')\mu_S(s') ds'$ , properly overwriting variables in  $S$  according to  $p$ ). In particular,  $\mathbf{m}^{+\mathcal{X},hold}$  has downprojections, so the cone data must satisfy  $\mu_{\mathbf{Y}}(\mathbf{Y}) = \mu_{\mathbf{X}}(\mathbf{Y})$  whenever  $\mathbf{Y} \subseteq \mathbf{X} \subseteq \mathcal{X}$ . In particular, this means all variables are determined by the particular marginal  $\mu_{\mathcal{X}}$ , pointing to the joint variable  $\mathcal{X}$ , which is present in  $\mathbf{m}^{+\mathcal{X}}$  and  $\mathbf{m}^{+\mathcal{X},hold}$ . Such a distribution (and its induced

---

<sup>3</sup>Note that these coherence maps do not include hyperarcs on these variables to put them back together, e.g.,  $\{\{X\}, \{Y\}\} \rightarrow \{X, Y\}$ . Including such a map is appropriate only if one believes  $X$  and  $Y$  are independent.

marginals) creates a cone over 1 only if it matches the appropriate conditional probability distributions for these other arcs. When  $S_a \cap T_a = \emptyset$  for all  $a$ , that corresponds precisely to the requirement that  $\mu$  matches all of the conditional marginals of  $\mathbb{P}$  (i.e.,  $\mu \in \{\mathcal{M}\}$ ). On the other hand, if  $S_a \cap T_a \neq \emptyset$  for some  $a$ , e.g., for a self loop  $p(X|X)$ ,

□

### 13.2.2 Natural Transformations

Suppose  $\mathcal{m}_1, \mathcal{m}_2 : \mathcal{A}^* \rightarrow \text{Meas}_{\Delta}$  are two PDGs generated by the same (hyper)graph  $\mathcal{A}$ . What is a natural transformation  $\eta : \mathcal{m}_1 \Rightarrow \mathcal{m}_2$ ?

By definition, it is a collection of stochastic maps  $\{\eta_X : \mathcal{m}_1(X) \rightarrow \mathcal{m}_2(X)\}_{X \in \mathcal{N}}$ <sup>4</sup> satisfying the property that, for all  $a : S \rightarrow T \in \mathcal{A}$ ,<sup>5</sup> the diagram

$$\begin{array}{ccc} \mathcal{m}_1(X) & \xrightarrow{m_1(a)} & \mathcal{m}_1(Y) \\ \downarrow \eta_X & & \downarrow \eta_Y \\ \mathcal{m}_2(X) & \xrightarrow{m_2(a)} & \mathcal{m}_2(Y) \end{array} \quad \left( \begin{array}{l} \text{or, in the} \\ \text{original no-} \\ \text{tation,} \end{array} \quad \begin{array}{ccc} \mathcal{V}_1 X & \xrightarrow{\mathbb{P}_a} & \mathcal{V}_1 Y \\ \eta_X \downarrow & & \downarrow \eta_Y \\ \mathcal{V}_2 X & \xrightarrow{\mathbb{P}_a^2} & \mathcal{V}_2 Y \end{array} \right)$$

commutes. This is a diagram in the category  $\text{Meas}_{\Delta}$ .

■

**The relationship between relations and probabilities.** There is a map  $\text{Supp}_X : \Delta X \rightarrow 2^X$  that takes a probability measure to its support set. In fact, it is a natural transformation

---

<sup>4</sup>Normally, we have been using the notation  $\mathcal{V}_1(X)$  and  $\mathcal{V}_2(X)$  for this concept, but for now we'll try this more traditional notation, and see if that works better.

<sup>5</sup>We have to verify this for all  $a \in \mathcal{A}^*$ , technically, but because  $\mathcal{A}^*$  is a free category, it suffices to check it only for the generating arcs  $a \in \mathcal{A}$ .

$$\begin{array}{ccc}
 & \Delta & \\
 \text{FinSet} & \begin{array}{c} \Downarrow \text{Supp} \\ \Downarrow \end{array} & \text{Set} , \\
 & \text{2}^{(-)} &
 \end{array}
 \quad \text{since the diagram} \quad
 \begin{array}{ccc}
 \Delta X & \xrightarrow{\delta f} & \Delta Y \\
 \text{Supp} \downarrow & & \downarrow \text{Supp} \\
 2^X & \xrightarrow{\bar{f}} & 2^Y
 \end{array}$$

commutes for all  $f : X \rightarrow Y$ ,<sup>6</sup> where  $\bar{f}(S) = \{f(x) : x \in S\}$ , often simply written as just  $\bar{f}$  to indicate the obvious extension of  $f$  itself to subsets of  $X$ , is the application of the functor  $2^{(-)}$  on  $f$ .

(( Can we use this to say something about how this interacts with IDef? What about  $H_\mu(Y|X)$  vs  $H_{\text{Supp } \mu}(Y|X)$ ? ))

---

<sup>6</sup>Proof:  $y \in \text{Supp}(\delta f(\mu)) \iff y \in f^{-1}(\text{Supp}(\mu)) \iff y \in \bar{f}(\text{Supp}(\mu))$ .

## **Part V**

## CHAPTER 14

# CONCLUSIONS

### 14.1 Summary

We have now seen a far-reaching unified account of probabilistic modeling, with implications far beyond the usual limits of probabilistically consistent reasoning. This theory is based on probabilistic dependency graphs (PDGs), an extremely expressive probabilistic modeling language. As we saw in [Part I](#), PDGs can capture essentially every standard fragment of epistemic information, ranging from traditional probabilistic graphical models such as Bayesian Networks and factor graphs ([Chapter 3](#)) to standard representations of uncertainty ([Chapter 4](#)). A PDG has two kinds of information: structural, and observational. Understanding the meaning of just the structural information in a PDG has been fruitful, leading to a generalized notion of independence that seems to be the beginning of a bridge between cyclic causal models and multivariate information theory ([Chapter 5](#)). As evidenced by many of the results in [Part II](#), the observational information is perhaps even more useful. With it, PDGs can capture systems composed of neural networks; moreover, the degree of observational inconsistency of the resulting PDG always seems to be the standard loss function used to train the model in question ([Chapter 6](#)). Furthermore, a wide variety of standard algorithms can be viewed as instances of a simple approach to resolving inconsistencies ([Chapter 7](#)). Indeed, learning, inference, and much else can be viewed as different aspects of a single objective: the pursuit of consistency.

In [Part III](#), we saw how these two views of PDGs—as a universal probabilistic model, and as a universal truth-based objective function (the foci of [Parts I](#) and [II](#),

respectively) are in fact two faces of the same concept, both semantically and algorithmically ([Chapter 9](#)). We also gave polynomial algorithm for inference in PDGs that have bounded treewidth ([Chapter 8](#)), finally solving both problems. We also **develop** useful principles for manipulating PDGs, allowing them to function as a visual proof language [Chapter 10](#).

**Sewing Things Back Together, with Self-Consistency.** We conclude our summary by reframing a handful of our simpler results in slightly different form. A PDG's degree observational inconsistency is a number in  $[0, \infty]$ ; it is also additive, in the sense that  $\langle\!\langle m_1 + m_2 \rangle\!\rangle = \langle\!\langle m_1 \rangle\!\rangle + \langle\!\langle m_2 \rangle\!\rangle$  when  $m_1, m_2$  are PDGs over disjoint sets of variables. As shown in [Chapter 11](#), additive measures in the range  $[0, \infty]$  can be equivalently represented as multiplicative measures in  $[0, 1]$ . So, for readers who are more comfortable with the latter range, we quickly summarize a few of our results.

Let the (*observational*) *self-consistency* of a PDG  $m$  be quantity

$$\Xi[m] := \exp(-\langle\!\langle m \rangle\!\rangle_0) \in [0, 1].$$

The following facts are either immediate or follow immediately from a result in [Chapter 4](#) or [6](#).

- A flat out logical contradiction has  $\Xi[\text{False}] = 0$  because it is completely inconsistent, while a collection of fully consistent probabilistic information  $m$  has  $\Xi[m] = 1$ .
- If  $U$  is an event, then  $\Xi[U] = \begin{cases} 0 & \text{if } U = \emptyset \\ 1 & \text{if } U \neq \emptyset \end{cases}$
- If  $U_1$  and  $U_2$  are both events over the same sample space, then  $\Xi[U_1, U_2] = \Xi[U_1 \cap U_2]$ .

- More generally, if  $R_1(\mathbf{X})$  and  $R_2(\mathbf{Y})$  are relations, then  $\Xi[R_1, R_2] = \Xi[R_1 \bowtie R_2]$ , where  $R_1 \bowtie R_2$  is the natural joint of  $R_1$  and  $R_2$  (??).
- If  $\mu$  is a probability distribution, and  $U$  is an event, then  $\Xi[\mu, U] = \mu(U)$ .
- If  $Bel$  is a Dempster-Shafer Belief function and  $Plaus$  is the corresponding plausibility function, and  $U$  is an event, then  $\Xi[Plaus, U] = \Xi[Bel, U] = Plaus(U)$ .
- If  $\mathcal{P}$  is a convex set of probability distributions and  $U$  is an event, then  $\Xi[m_{\mathcal{P}}] = \sup_{P \in \mathcal{P}} P(U)$ .
- If  $D = \{(x_i, y_i)\}$  is a set of training data and  $f : X \rightarrow Y$  is a deterministic classifier, then  $\Xi[\Pr_D, f]$  is the accuracy of  $f$  on  $D$ .
- If  $p, q \in \Delta^{\mathcal{V}X}$  are distributions over a discrete variable  $X$ , then  $\Xi[p(X), q(X)] = \left( \sum_{x \in \mathcal{V}X} \sqrt{p(x)q(x)} \right)^2$  is the square of the Bhattacharya coefficient. If  $p$  and  $q$  are each given confidence  $\frac{1}{2}$ , then  $1 - \Xi[p, q]$  is the squared Hellinger distance between  $p$  and  $q$ .
- If  $p$  and  $q$  are both unit normal distributions over  $X$ , with means  $m_1$  and  $m_2$ , respectively, then  $\Xi[p(X), q(X)] = \exp(-\frac{1}{2}(m_1 - m_2)^2)$  itself resembles the density of a unit normal distribution.

## 14.2 Future Work and Open Questions

The theory presented in this thesis is a significant reframing of the probabilistic modeling process with far-reaching implications. This thesis has presented quite a few of them. But the theory is still in its infancy, and we are still a far away from having fully developing its limits and applications. We conclude with just a few of these avenues for future work.

**Capturing Even More Modeling Formalisms.** There is no reason to think that the material presented here reflects the limits of what PDGs can represent. For instance, PDGs can capture relations (Section 4.2.1), so it makes sense to wonder whether they can capture relational databases (Abiteboul et al. 1995). Preliminary investigations suggest that this is the case, although fully capturing the relationship will require also developing an analogue of a databases’s query language. Another reason to believe this is that data dependencies (Fagin and Vardi 1986) seem to closely match the qualitative semantics of a PDG’s underlying hypergraph. Since databases do not describe probabilities (apart from 0 and 1), encoding them with PDGs should involve only deterministic cpds. This raises another question: what can we encode by allowing probabilistic relationships? Our initial investigations suggest that the result may be quite different from the standard notion of a probabilistic database (Suciu et al. 2011), and perhaps has elements of modern information retrieval systems (Mitra and Craswell 2018). So far we have talked only about relational databases—but there is also a second important class of *graph-based* databases. In many ways, this seems an even closer fit to a PDG, and understanding whether and how they bear relationship to PDG semantics remains a promising open problem.

In Machine Learning, there has been a recent deluge of work introducing new representations and training objectives. At the moment, the two dominant architectures are *diffusion models* (Sohl-Dickstein et al. 2015; Ho et al. 2020) and *transformers* (Vaswani et al. 2017). Because of their roots in VAEs (which PDGs capture well (Section 6.5.1)), it is not hard to show that diffusion models can also be regarded as PDGs. Does this view yield any practical or conceptual benefits, beyond what we have already seen? We do not know. So far, even less said for a relationship between PDGs and transformers. Still, it is a question worth taking

seriously, if one believes in the promise of PDGs as a foundation for modern probabilistic AI systems.

**Further Investigation into Qualitative PDGs.** Section 5.4 explores how one (perhaps surprising) way in which the qualitative information in a PDG is deeply related to the notion of mechanism independence (Definition 5.1). Yet we suspect there is more to the story. First, an analogue of the principle of maximum entropy leads us to a special case of QIM-compatibility in which the witness does not break any unnecessary symmetries. This special variant appears to have an even closer relationship with *SDef*. Furthermore, preliminary experiments suggest that minimizing structural deficiency may suffice to ensure QIM compatibility.

A second avenue of future research into qualitative PDGs involves the relationship between quantifiers and extreme values of the parameter  $\alpha$ . Specifically, it seems that existential quantifiers ( $\exists$ ) can be implemented with qualitative arcs that have  $\alpha = +\infty$ , while universal quantifiers ( $\forall$ ) can be implemented with  $\alpha = -\infty$ . This observation may be useful in developing a query language for PDG-based databases (discussed above). It also raises some deeper questions; can we square this with our understanding of  $\alpha$  as representing independent mechanisms (Chapter 5)?

**Further Investigation into PDG inference** As mentioned in the conclusions to Chapter 8, we have given an  $\tilde{O}(N^{2.0})$  approach to inference in the case of bounded treewidth, but the best lower bound is  $\tilde{\Omega}(N^1)$ . This is still a significant gap, that leaves PDGs trailing far behind other graphical models with analogous inference approaches. Furthermore, our algorithm only provably works for inference in cases where  $\beta \geq \gamma\alpha$  is small. Solving the problem in the case where qualitative

information plays a larger role seems to be a significantly more difficult problem in general, and one that remains open.

+ non-discrete PDGs

There are many different approaches to inference in traditional graphical models; see Chapters 9-14 of [Koller and Friedman \(2009\)](#) for a survey of them. Our approach to inference in PDG is a cousin of the the techniques developed in Ch11. But what of analogues of the other approaches? Particle-based and MCMC methods in in particular (the analogue of Ch12 [ibid]) for PDGs remain almost completely unexplored; can these techniques be adapted for use in PDGs?

A related question is that of a deeper understanding of the local inconsistency resolution (LIR) algorithm (the subject of [Chapter 7](#)). Under what conditions does this procedure converge? And if it converges, when does it produce correct answers? These questions are closely related to similar questions about belief propagation, which have been partially resolved across decades of research ([Yedidia et al. 2000](#); [Wiegerinck and Heskes 2002](#); [Minka 2005](#)).

need more and better citations here.

While treewidth dominates the complexity of (exact) inference in traditional graphical models, it is not obvious that the same must be true for PDGs. PDGs can express many things that other graphical models cannot. Therefore, there may be nontrivial subclasses of PDGs that are completely unrelated to treewidth, which still admit tractable inference. Some inconsistencies, for example, are easier to quantify than others—a fact to which variational inference owes its existence. In some cases, one can easily verify that a PDG has zero inconsistency through its hypergraph alone, without even looking at the probabilities. (That is why it is possible to ensure representations are consistent by construction in the first place.) Neither of these subclasses has anything to do with bounded treewidth. Are there general principles that tell us how difficult it could be to

calculate a PDG’s degree of inconsistency from  $(\alpha, \beta, \mathcal{A})$ ? Might it be possible to automate the search for PDGs that upper and lower bound a PDG’s inconsistency, to get an adaptive general purpose analogue of variational inference?

**Deepening Theoretical Roots.** In [Chapter 13](#), we found that PDGs can be viewed as diagrams in a certain category. We saw that categorical limits of PDGs are important. But what about colimits, and natural transformations between them? The categorical picture opens a vast landscape of abstract research directions. One natural question is whether probability can be replaced by something else—in categorical terms, what is needed to make probability work for a different monad.

Other connections that seem promising connecting to categorical characterizations of relative entropy ([Baez and Fritz 2014](#)) and of information loss ([Leinster 2021](#), Theorem 12.4.9). Doing this properly may also require developing a categorical account of confidence (the subject of [Chapter 11](#)).

Yet the most important future work lies in application rather than theory.

### 14.3 Impact: Implementing and Applying the Theory

The material here has been conceptual, mathematically technical, and sometimes philosophical. But *is it useful?*

Narrowly construed as “can it be used to help us understand AI systems built with probabilistic tools”, I submit that the answer is a resounding yes. For pedagogical reasons alone, there is clear value in unifying so many different

concepts in a way that clarifies the relationships between them. Especially so because it also answers important foundational questions, such as how loss functions are related to one another, and to probability.

Equally narrowly construed as, “can I run it on my laptop?”, the answer is also yes. For the reader interested in playing around with PDGs, I have developed a general-purpose Python library implementing the framework in the discrete case, as well as many of the constructions in this thesis; it can be found at <https://github.com/orichardson/pdg>. Many of the concepts we have seen have proven difficult to manipulate in my head, and the truth has surprised me often. Using and developing this library has helped me explore and understand the material more deeply; I cannot recommend implementing one’s math highly enough. To a reader who wonders why one should bother testing something when you could prove it, I say it’s better to have two ways of understanding things than it is to have just one.

But the deeper question—of whether or not this theory will ultimately help humanity address important real-world problems—remains the most important open question of all. We are already building powerful artificial agents that are quickly reshaping our world; no doubt one of the biggest problems of our time is to get a handle on how we want AI systems to work in the future.

This unified theory of probabilistic modeling and epistemic conflict has the potential to be a key element of that future. For one, it could provide students and researchers with a principled approach to understanding and designing these systems—even the powerful modern ones which are seldom fully consistent. For another, it may help to resolve questions of alignment, as it provides a principled way of pursuing a universal value: self-consistency. Indeed, the entire

theory is based only on truth and trust—analogues of “utility” appear only ephemeral intermediates in this process. There is also a case to be made that universal constructions promote diversity and fairness. Perhaps in this moment, what’s most important is to develop a clearer and more mature understanding of cognition—both for ourselves, and for the generations of powerful AI systems to come.

## BIBLIOGRAPHY

Ternary search. [https://cp-algorithms.com/num\\_methods/ternary\\_search.html](https://cp-algorithms.com/num_methods/ternary_search.html), 2023. [appears to be folklore; accessed online July 2024].

Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Databases*, volume 8. Addison-Wesley Reading, 1995.

Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

S.M. Aji and R.J. McEliece. The generalized distributive law. *IEEE Transactions on Information Theory*, 46(2):325–343, 2000. doi: 10.1109/18.825794.

MOSEK ApS. *MOSEK Optimizer API for Python* 10.0.25, 2022. URL <https://docs.mosek.com/10.0/pythonapi/index.html>.

David P Ausubel and Mohamed Youssef. The effect of spaced repetition on meaningful retention. *The Journal of General Psychology*, 73(1):147–150, 1965.

Riley Badenbroek and Joachim Dahl. An algorithm for nonsymmetric conic optimization inspired by mosek. *Optimization Methods and Software*, pages 1–38, 2021.

John C. Baez and Tobias Fritz. A bayesian characterization of relative entropy, 2014. URL <https://arxiv.org/abs/1402.3067>.

Christel Baier, Clemens Dubslaff, Holger Hermanns, and Nikolai Käfer. On the foundations of cycles in bayesian networks. In *Lecture Notes in Computer Science*, pages 343–363. Springer Nature Switzerland, 2022. doi: 10.1007/978-3-031-22337-2\_17. URL [https://doi.org/10.1007%2F978-3-031-22337-2\\_17](https://doi.org/10.1007%2F978-3-031-22337-2_17).

A. Balke and J. Pearl. Probabilistic evaluation of counterfactual queries. In *Proc. Eleventh National Conference on Artificial Intelligence (AAAI '94)*, pages 200–207, 1994.

Alex Becker. Tutorial on the kalman filter, 2003. URL <https://www.kalmanfilter.net/>.

Sander Beckers, Joseph Y. Halpern, and Christopher Hitchcock. Causal models with constraints, 2023.

Umberto Bertele and Francesco Brioschi. *Nonserial dynamic programming*. Academic Press, Inc., 1972.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Advanced Information Systems Engineering*, pages 387–402. Springer Berlin Heidelberg, 2013. doi: 10.1007/978-3-642-40994-3\_25. URL [https://doi.org/10.1007%2F978-3-642-40994-3\\_25](https://doi.org/10.1007%2F978-3-642-40994-3_25).

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518): 859–877, 2017.

Hans L Bodlaender. A linear time algorithm for finding tree-decompositions of small treewidth. In *Proceedings of the twenty-fifth annual ACM symposium on Theory of computing*, pages 226–234, 1993.

Stephen P Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Robert Grover Brown and Patrick YC Hwang. Introduction to random signals and applied kalman filtering: with matlab exercises and solutions. *Introduction*

*to random signals and applied Kalman filtering: with MATLAB exercises and solutions*, 1997.

Penha Maria Cardoso Dias and Abner Shimony. A critique of jaynes' maximum entropy principle. *Advances in Applied Mathematics*, 2(2):172–211, 1981. ISSN 0196-8858. doi: [https://doi.org/10.1016/0196-8858\(81\)90003-8](https://doi.org/10.1016/0196-8858(81)90003-8). URL <https://www.sciencedirect.com/science/article/pii/0196885881900038>.

Venkat Chandrasekaran, Nathan Srebro, and Prahladh Harsha. Complexity of inference in graphical models. *arXiv preprint arXiv:1206.3240*, 2012.

Nikolai Nikolaevich Chentsov. *Statiscal Decision Rules and Optimal Inference*, volume 53. American Mathematical Society, 1982. ISBN 0-8218-4502-0.

David Maxwell Chickering. A transformational characterization of equivalent bayesian network structures. *CoRR*, abs/1302.4938, 2013. URL <http://arxiv.org/abs/1302.4938>.

Christophe Chipot and Andrew Pohorille. Free energy calculations. *Springer Series in Chemical Physics*, 86:159–184, 2007.

Yoeng-Jin Chu. On the shortest arborescence of a directed graph. *Scientia Sinica*, 14:1396–1400, 1965.

Andrzej Cichocki and Shun-ichi Amari. Families of alpha beta and gamma divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

Bruno Courcelle. The monadic second-order logic of graphs. i. recognizable sets of finite graphs. *Information and Computation*, 85(1):12–75, 1990. ISSN 0890-5401. doi: [https://doi.org/10.1016/0890-5401\(90\)90043-H](https://doi.org/10.1016/0890-5401(90)90043-H). URL <https://www.sciencedirect.com/science/article/pii/089054019090043H>.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

Paul Dagum and Michael Luby. An optimal approximation algorithm for bayesian inference. *Artificial Intelligence*, 93(1):1–27, 1997. ISSN 0004-3702. doi: [https://doi.org/10.1016/S0004-3702\(97\)00013-1](https://doi.org/10.1016/S0004-3702(97)00013-1). URL <https://www.sciencedirect.com/science/article/pii/S0004370297000131>.

Joachim Dahl and Erling D Andersen. A primal-dual interior-point algorithm for nonsymmetric exponential-cone optimization. *Mathematical Programming*, 194(1):341–370, 2022.

A Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

Thomas Dean and Keiji Kanazawa. A model for reasoning about persistence and causation. *Computational Intelligence*, 5(2):142–150, 1989.

René Descartes. The discourse on method, 1637.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Alexander Domahidi, Eric Chu, and Stephen Boyd. Ecos: An socp solver for embedded systems. In *2013 European Control Conference (ECC)*, pages 3071–3076, 2013. doi: 10.23919/ECC.2013.6669541.

Ran Duan, Hongxun Wu, and Renfei Zhou. Faster matrix multiplication via asymmetric hashing. *arXiv preprint*, 2022. doi: 10.48550/ARXIV.2210.10173. URL <https://arxiv.org/abs/2210.10173>.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

DK Fadeev. Zum begriff der entropie einer endlichen wahrscheinlichkeitsschemas. *Arbeiten zur Informationstheorie I*. Deutscher Verlag der Wissenschaften, pages 85–90, 1957.

Ronald Fagin and Moshe Y Vardi. The theory of data dependencies: A survey. In *Mathematics of Information Processing: Proceedings of Symposia in Applied Mathematics*, volume 34, pages 19–71. Amer. Math. Soc., 1986.

Leon Festinger. Cognitive dissonance. *Scientific American*, 207(4):93–106, 1962.

Maurice A Finocchiaro. Fallacies and the evaluation of reasoning. *American Philosophical Quarterly*, 18(1):13–22, 1981.

Roger Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2013.

Brendan J Frey. Extending factor graphs so as to unify directed and undirected graphical models. *arXiv preprint arXiv:1212.2486*, 2012.

Kenneth Friedman and Abner Shimony. Jaynes’s maximum entropy prescription and probability theory. *Journal of Statistical Physics*, 3:381–384, 1971.

Karl Friston. The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, 13(7):293–301, 2009.

Giorgio Gallo, Giustino Longo, Stefano Pallottino, and Sang Nguyen. Directed hypergraphs and applications. *Discrete Applied Mathematics*, 42(2):177–201, 1993. ISSN 0166-218X. doi: [https://doi.org/10.1016/0166-218X\(93\)](https://doi.org/10.1016/0166-218X(93))

90045-P. URL <https://www.sciencedirect.com/science/article/pii/0166218X9390045P>.

Peter Gardenfors. Imaging and conditionalization. *The Journal of Philosophy*, 79(12):747–760, 1982. ISSN 0022362X. URL <http://www.jstor.org/stable/2026039>.

Dan Geiger, Thomas Verma, and Judea Pearl. Identifying independence in bayesian networks. *Networks*, 20(5):507–534, 1990. doi: <https://doi.org/10.1002/net.3230200504>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/net.3230200504>.

Federico Girosi, Michael Jones, and Tomaso Poggio. Regularization theory and neural networks architectures. *Neural computation*, 7(2):219–269, 1995.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2014.

Michael Charles Grant. *Disciplined Convex Programming*. PhD thesis, Stanford University, December 2004. URL [https://web.stanford.edu/~boyd/papers/pdf/mcg\\_thesis.pdf](https://web.stanford.edu/~boyd/papers/pdf/mcg_thesis.pdf).

Adam J. Grove and Joseph Y. Halpern. Probability update: Conditioning vs. cross-entropy. In *Conference on Uncertainty in Artificial Intelligence*, 1997. URL <https://api.semanticscholar.org/CorpusID:14707750>.

Aditya Grover and Stefano Ermon. Lecture notes in deep generative models. [deepgenerativemodels.github.io/notes/](https://deepgenerativemodels.github.io/notes/), 2018.

Peter Hall. On representatives of subsets. *Journal of The London Mathematical Society-second Series*, pages 26–30, 1935. URL <https://api.semanticscholar.org/CorpusID:23252557>.

J. Y. Halpern and S. Leung. Weighted sets of probabilities and minimax weighted expected regret: new approaches for representing uncertainty and making decisions. *Theory and Decision*, 79(3):415–450, 2015.

Joseph Y Halpern. Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research*, 12:317–337, 2000.

Joseph Y Halpern. *Reasoning About Uncertainty*. MIT press, 2017.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition, 2009.

David Heckerman, David Maxwell Chickering, Christopher Meek, Robert Rounthwaite, and Carl Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1(Oct):49–75, 2000.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

Christopher Hitchcock. Causal Models. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2024 edition, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

James E. Matheson Howard, Ronald A. Influence diagrams. *Readings on the Principles and Applications of Decision Analysis*, pages 719–763, 1983.

Shruti Jadon. A survey of loss functions for semantic segmentation. In 2020 *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–7. IEEE, 2020.

Ryan James. (stumbling blocks) on the road to understanding multivariate information theory. Discrete Information Theory package documentation, 2018.  
URL <https://dit.readthedocs.io/en/latest/stumbling.html>.

Ryan G. James and James P. Crutchfield. Multivariate dependence beyond shannon information. *Entropy*, 19(10), 2017. ISSN 1099-4300. doi: 10.3390/e19100531. URL <https://www.mdpi.com/1099-4300/19/10/531>.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

R. C. Jeffrey. Probable knowledge. In I. Lakatos, editor, *International Colloquium in the Philosophy of Science: The Problem of Inductive Logic*, pages 157–185. North-Holland, Amsterdam, 1968.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

Benjamin Lucien Kaminski, Joost-Pieter Katoen, and Christoph Matheja. *Expected Runtime Analysis by Program Verification*, page 185–220. Cambridge University Press, 2020.

Narendra Karmarkar. A new polynomial-time algorithm for linear programming. In *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing*, pages 302–311, 1984.

S Kasangian and RFC Walters. The duality between flow charts and circuits.

*Bulletin of the Australian Mathematical Society*, 42(1):71–79, 1990.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization.

*arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

Varsha Kishore, Xiangyu Chen, Yan Wang, Boyi Li, and Kilian Q Weinberger.

Fixed neural network steganography: Train the images, not the network. In *International Conference on Learning Representations*, 2021.

Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, nov 2021. doi: 10.1109/tpami.2020.2992934. URL <https://doi.org/10.1109%2Ftpami.2020.2992934>.

D. Koller and N. Friedman. *Probabilistic Graphical Models*. MIT Press, Cambridge, MA, 2009.

F. R. Kschischang, B. J. Frey, and H. . Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

Frank R Kschischang, Brendan J Frey, and H-A Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on information theory*, 47(2):498–519, 2001.

S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H.-G. Leimer. Independence properties of directed markov fields. *Networks*, 20(5):491–505, 1990. doi: <https://doi.org/10.1002/net.3230200503>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/net.3230200503>.

Steffen L Lauritzen and David J Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194, 1988.

John M Lee. *Smooth Manifolds*. Springer, 2013.

Tom Leinster. A short characterization of relative entropy, 2017. URL <https://arxiv.org/abs/1712.04903>.

Tom Leinster. *Entropy and Diversity: the Axiomatic Approach*. Cambridge University Press, 2021.

leonbloy. conditioning reduces mutual information. Mathematics Stack Exchange, 2015. URL <https://math.stackexchange.com/q/1219753>. URL: <https://math.stackexchange.com/q/1219753> (version: 2015-04-04).

David Lewis. Probabilities of conditionals and conditional probabilities. In *Ifs*, pages 129–147. Springer, 1976.

Roi Livini. Follow the regularized leader. <https://www.cs.princeton.edu/~rlivni/cos511/lectures/lect21.pdf>, 2017. [Lecture notes; accessed online July 2024].

Jianzhu Ma, Jian Peng, Sheng Wang, and Jinbo Xu. Estimating the partition function of graphical models using langevin importance sampling. In *Artificial Intelligence and Statistics*, pages 433–441. PMLR, 2013.

David J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

Brendan McMahan. Follow-the-regularized-leader. <https://courses.cs.washington.edu/courses/cse599s/14sp/scribes/lecture3/lecture3.pdf>, March 2014. [Lecture Notes; accessed online July 2024].

Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

Tom Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research, Cambridge, U.K., 2005.

Bhaskar Mitra and Nick Craswell. 2018. doi: 10.1561/1500000061.

In Jae Myung. Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, 47(1):90–100, 2003.

Pavel Naumov and Brittany Nicholls. R.e. axiomatization of conditional independence, 2013.

Radford M Neal and Geoffrey E Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. Springer, 1998.

Yu E Nesterov, Michael J Todd, and Yinyu Ye. Infeasible-start primal-dual methods and infeasibility detectors for nonlinear programming problems. Technical report, 1999.

Yurii Nesterov and Arkadii Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM, 1994.

Frank Nielsen. Chernoff information of exponential families. *arXiv preprint arXiv:1102.2684*, 2011.

Otton Nikodym. Sur une généralisation des intégrales de mj radon. *Fundamenta Mathematicae*, 15(1):131–179, 1930.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga,

- Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL <https://arxiv.org/abs/1912.01703>.
- HP Patil. On the structure of k-trees. *Journal of Combinatorics, Information and System Sciences*, 11(2-4):57–64, 1986.
- J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, 1988.
- J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000.
- J Pearl and A Paz. Graphoids: A graph-based logic for reasoning about relevance relations, 1987.
- Judea Pearl. *Causality*. Cambridge university press, 2009.
- Graham Priest, Koji Tanaka, and Zach Weber. Paraconsistent logic. 1996.
- Jason Rennie. On l2-norm regularization and the gaussian prior. 2003.
- Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 547–561. University of California Press, 1961.
- Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.
- Oliver E Richardson. Loss as the inconsistency of a probabilistic dependency graph: Choose your model, not your loss function. *AISTATS '22*, 151, 2022.

- Oliver E Richardson. The local inconsistency resolution algorithm (workshop version), 2023. ICML '23 Workshops: Local Learning Workshop (LLW) and Structured Prediction in Generative Modeling (SPIGM).
- Oliver E Richardson and Jialu Bao. Mixture languages, 2024. Principles of Programming Languages (POPL) Workshop: Languages for Inference (LAFI).
- Oliver E Richardson and Joseph Y Halpern. Probabilistic dependency graphs. *AAAI '21*, 2021.
- Oliver E Richardson, Joseph Y Halpern, and Christopher De Sa. Inference in probabilistic dependency graphs. *UAI '23*, 2023.
- Oliver E Richardson, Spencer Peters, and Joseph Y Halpern. Qualitative mechanism independence. 2024. In Submission: for NeurIPS 2024.
- Dan Roth. On the hardness of approximate reasoning. *Artificial Intelligence*, 82(1): 273–302, 1996. ISSN 0004-3702. doi: [https://doi.org/10.1016/0004-3702\(94\)00092-1](https://doi.org/10.1016/0004-3702(94)00092-1). URL <https://www.sciencedirect.com/science/article/pii/0004370294000921>.
- D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.
- Bertrand Russell. Letter to frege. *From Frege to Gödel*, 6:124–125, 1902.
- San Diego Union. <https://quoteinvestigator.com/2022/07/04/watch/#f+441634+1+1>, 1930. Quote Page 4, Column 1.
- Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkhäuser*, NY, 55(58-63):94, 2015.
- L. J. Savage. *Foundations of Statistics*. Wiley, New York, 1954.

Teddy Seidenfeld. Entropy and uncertainty. *Philosophy of Science*, 53(4):467–491, 1986.

Bart Selman, David G Mitchell, and Hector J Levesque. Generating hard satisfiability problems. *Artificial intelligence*, 81(1-2):17–29, 1996.

Glenn Shafer. *A Mathematical Theory of Evidence*, volume 42. Princeton university press, 1976.

Glenn R Shafer and Prakash P Shenoy. Probability propagation. *Annals of mathematics and Artificial Intelligence*, 2:327–351, 1990.

Michael Sipser. *Introduction to the Theory of Computation* (2nd ed.). Thomson Course Technology., second edition, 2006. ISBN 978-0-534-95097-2.

Anders Skajaa and Yinyu Ye. A homogeneous interior-point algorithm for nonsymmetric convex conic optimization. *Mathematical Programming*, 150(2):391–422, 2015.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL <https://proceedings.mlr.press/v37/sohl-dickstein15.html>.

Dan Suciu, Dan Olteanu, Christopher Ré, and Christoph Koch. *Probabilistic Databases*, volume 3. Morgan & Claypool Publishers, 2011.

Esteban G Tabak and Eric Vanden-Eijnden. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010.

Myron Tribus. Information theory as the basis for thermostatics and thermodynamics. 1961.

Alexandre Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.  
ISBN 9780387790527.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fdb053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fdb053c1c4a845aa-Paper.pdf).

Susan Vineberg. Dutch Book Arguments. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition, 2022.

Martin J Wainwright, Tommi S Jaakkola, and Alan S Willsky. Tree-based reparameterization framework for analysis of sum-product and related algorithms. *IEEE Transactions on information theory*, 49(5):1120–1146, 2003.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, 1991.

Qi Wang, Yue Ma, Kun Zhao, and Yingjie Tian. A comprehensive survey of loss functions in machine learning. *Annals of Data Science*, pages 1–26, 2020.

Wim Wiegerinck and Tom Heskes. Fractional belief propagation. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002. URL [https://proceedings.neurips.cc/paper\\_files/paper/2002/file/35936504a37d53e03abdfbc7318d9ec7-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2002/file/35936504a37d53e03abdfbc7318d9ec7-Paper.pdf).

Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information, 2010.

Peter M Williams. Bayesian regularization and pruning using a laplace prior. *Neural Computation*, 7(1):117–143, 1995.

Jon Williamson. Foundations for bayesian networks. In *Foundations of Bayesianism*, pages 75–115. Springer, 2001.

Jonathan S Yedidia, William Freeman, and Yair Weiss. Generalized belief propagation. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000. URL [https://proceedings.neurips.cc/paper\\_files/paper/2000/file/61b1fb3f59e28c67f3925f3c79be81a1-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2000/file/61b1fb3f59e28c67f3925f3c79be81a1-Paper.pdf).

Alan L Yuille and Anand Rangarajan. The concave-convex procedure. *Neural computation*, 15(4):915–936, 2003.

Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency.

In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL [https://proceedings.neurips.cc/paper\\_files/paper/2003/file/87682805257e619d49b8e0dfdc14affa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2003/file/87682805257e619d49b8e0dfdc14affa-Paper.pdf).

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5982–5991, 2019.