RELATIVE ENTROPY SOUP

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

RELATIVE ENTROPY SOUP

Oliver E. Richardson, Ph.D.

Cornell University 2024

Your abstract goes here. Make sure it sits inside the brackets. If not, your biosketch page may not be roman numeral iii, as required by the graduate school.

BIOGRAPHICAL SKETCH

Your biosketch goes here. Make sure it sits inside the brackets.

This document is dedicated to all Cornell graduate students.

ACKNOWLEDGEMENTS

Your acknowledgements go here. Make sure it sits inside the brackets.

TABLE OF CONTENTS

Chapter 3 Appendices

Chapter 5 Appendices

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1

1.2 Overview of Results

1.3 Themes

1.3.1 Epistemic Humility

I distinguish between

CHAPTER 2

PRELIMINARIES

2.0.1 Basic Notation

If A is a finite set, we write $\#A$ or $|A|$ for its cardinality. We will often be concerned with *variables*, which intuitively correspond to aspects of the world or properties of some object. Mathematically, a variable has two aspects: qualitatively, a variable is just some unique identifier (the variable name). Quantitatively, a variable X is also associated with a set $\mathcal{V}(X)$, or simply $\mathcal{V}X$, of possible values.

2.0.2 Algebra

Definition 2.0.1 (Monoid)**.** A *monoid* is a tuple $(S, *, e)$, where S is a set, $* : S \times S \to S$ is a binary operation, and $e \in S$ is a distinguished identity element, such that:

- (associativity) $\forall a, b, c \in S.\ (a * b) * c = a * (b * c)$;
- (identity) $\forall a \in S.\ a * e = a = e * a$.

A monoid is called *commutative* if it also satisfies

- (commutativity): $\forall a, b \in S.\ a * b = b * a$,

and *idempotent* if it satisfies

- (idempotence): $\forall a \in S.\ a + a = a$.

An idempotent semiring defines partial order by $a \leq b \iff a + b = b$. □

2.0.3 Relations

Let $\mathcal{X} = \{X_1, \dots, X_n\}$ be variables, traditionally called atributes. A *relation* $R(\mathcal{A}) = R(X_1, \dots, X_n) \subseteq \mathcal{V}(X_1) \times \cdots \times \mathcal{V}(X_n)$, or equivalently, $R : \prod_{i=1}^{n} \mathcal{V}X_i \to \{0, 1\}$, is a subset of joint values of attributes. The natural number n is called the *arity* of R.

The *natural join* of two relations $R(A, B)$ and $S(B, C)$ combines them in a particularly obvious way: $(a, b, c) \in R \bowtie S$ iff $(a, b) \in R$ and $(b, c) \in S$. More generally, we have

Definition 2.0.2 (natural join). $R(\mathcal{X}) \bowtie S(\mathcal{Y}) := \Big\{\boldsymbol{\omega} \in \mathcal{V}(\mathcal{X} \cup \mathcal{Y}) \Bigm| \mathcal{X}(\boldsymbol{\omega}) \in R \wedge \mathcal{Y}(\boldsymbol{\omega}) \in S\Big\}$. □

At one extreme, if $\mathcal{X}$ and $\mathcal{Y}$ are disjoint sets of attributes, then $R(\mathcal{X}) \bowtie S(\mathcal{Y})$ coincides with the cartesian product of $R \subseteq \mathcal{V}\mathcal{X}$ and $S \subseteq \mathcal{V}\mathcal{Y}$. At the opposite extreme, if $\mathcal{X} = \mathcal{Y}$ are the same set of variables, then $R(\mathcal{X}) \bowtie S(\mathcal{X})$ coincides with the intersection of the subsets R and S.

Even when $A_1, \dots, A_n \subseteq \mathcal{X}$ are not disjoint, we give a convenient extended syntax by defining the quantity $R(a_1, \dots, a_n)$, where $a_i \in \mathcal{V}(A_i)$. Concretely, define $R(a_1, \dots, a_n) := 0$ if when $\{a_1, \dots, a_n\}$ do not agree on the value of some shared attribute (i.e., if $\exists X \in \mathcal{X}, \exists i, j \in [n].\ X \in A_i \cap A_j \wedge X(a_i) \neq X(a_j)$). When $\{a, b, c\}$ do agree on all values of shared attributes, let $\mathbf{x}$ denote the joint value of $A \cup B \cup C$ obtained from (a, b, c) by removing redundant copies of variable values. In this case, define $R(a, b, c) := R(\mathbf{x})$.

2.0.4 Graph Theory

Definition 2.0.3. A *(directed) (multi) graph* $G = (N, A)$, or simply a *graph*, is a set N of nodes, and a collection A of 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 $S, T\colon A \to N$ often left implicit. □

Definition 2.0.4 (Undirected (Multi) Graph)**.** 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

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

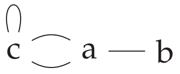
implicit in the definition of G, which we will write $G = (N, E, \iota)$ only when being extra careful. □

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

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

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

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



A (directed) graph has more information than an undirected one. There are natural ways to convert between the two: one forgets the direction of arcs to turn a directed graph into an undirected one, and annotates each arc with arrows in both directions to make an undirected graph directed. These choices are essentially locked in if we want the correspondence with square matrices to hold properly.

$$\begin{tikzcd}
\text{(Directed) Graphs} \arrow[r, bend left, "\mathit{forget}"] & \text{Undirected Graphs} \arrow[l, bend left, "\mathit{annotate}"]
\end{tikzcd}$$

in which $\mathit{forget} \circ \mathit{annotate} = \mathrm{id}_{\text{Undirected Graphs}}$ but $\mathit{anotate} \circ \mathit{forget}$ is not the identity on (directed) graphs. Technically, this makes forget a *retraction*, and $\mathit{annotate}$ a *section*.

Definition 2.0.5. A bipartite graph $G = (L, R, E)$ is a graph $(L \sqcup R, E)$ whose vertices are partiioned into two components $V = L \sqcup R$, and whose edges $E \subset L \times R$ are only between L and R. □

Definition 2.0.6. A directed bipartite graph $G = (L, R, E)$ is a bipartite graph (L, R, E) whose edges $E \subset (L \times R) \cup (R \times L)$ are directed. □

Definition 2.0.7. A *hypergraph* $G = (V, \mathcal{E})$ is a set V of vertices, and a collection $\mathcal{E}$ of edges, which correspond to finite subsets of vertices. □

Thus, a hypergraph is the generalization of an undirected graph in which the codomain of $\iota : \mathcal{E} \to 2^V$ is arbitrary subsets of V, not just those of cardinality 2.

Proposition 2.0.1. *There is a natural bijection between hypergraphs and bipartite graphs:*

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

$$\mathit{hyper}(L, R, E) := (L, \{\{v \in L : (v, r) \in E\} : r \in R\}),$$

$$\mathit{bipart} \circ \mathit{hyper} = \mathrm{id}_{BG} \quad \textit{and} \quad \mathit{hyper} \circ \mathit{bipart} = \mathrm{id}_{HG}.$$

The consequences of this can be unintuitive. It is common to think of bipartite graphs as a strict (particularly nice) special case of ordinary undirected graphs, which themselves are a strict (particularly easy to draw) strict special case of hypergraphs. By transitivity, one might expect bipartite graphs to naturally be an extremely strict special case of hypergraphs—yet in fact they are naturally isomorphic.

Definition 2.0.8. A *directed* hypergraph $(N, \mathcal{A})$ is a set N of nodes, and a collection $\mathcal{A}$ of hyperarcs, each of which has a set $S_a \subset N$ of source variables and a set $T_a \subset N$ of target variables. □

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.

Definition 2.0.9. The *dual* of the hypergraph $G = (V, \mathcal{E})$ is $\check{G} := (\mathcal{E}, \{\{e \in \mathcal{E} : v \in e\} : v \in V\})$. □

Definition 2.0.10. The *dual* of a directed hypergraph $\mathcal{H} = (N, \mathcal{A})$ is $\check{\mathcal{H}} := (\mathcal{A}, N)$, where

$$\check{S}_n = \{a \in \mathcal{A} : n \in T_a\} \quad \text{and} \quad \check{T}_n = \{a \in \mathcal{A} : n \in S_a\}.$$ □

We now verify that $\check{\check{\mathcal{H}}} = \mathcal{H}$. Observe that

$$\begin{aligned}
\check{\check{S}}_a &= \{n \in N : a \in \check{T}_n\} \\
&= \{n \in N : a \in \{a' \in \mathcal{A} : n \in S_a'\}\} \\
&= \{n \in N : n \in S_a\} \\
&= S_a;
\end{aligned}$$

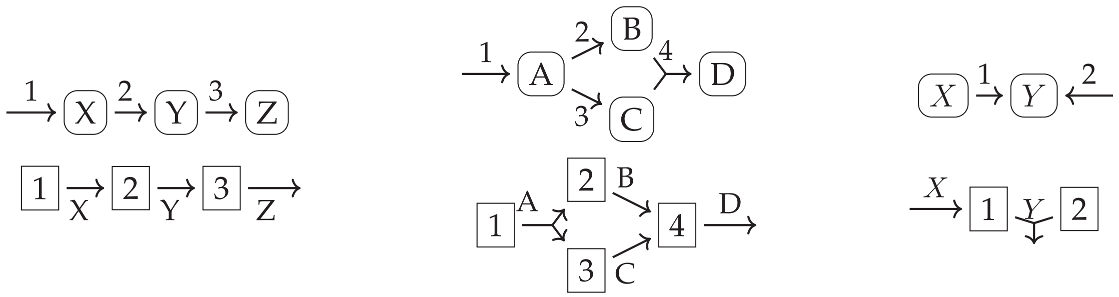


Figure 2.1: Examples of directed hypergraphs (first row) and their duals (second row).

symetrically, $T_a = \check{\check{T}}_a$.

See Figure 2.1 for some visual illustrations. We remark that the left and center diagrams on the top can be viewed as (the hypergraphs corresponding to) qualitative Bayesian Networks, by regarding X,Y,Z and A,B,C,D as variables, and imagining that there is a (randomized) causal determination occuring along each arc. One can also imagine an analogue with cycles—resuling in perhaps a (randomized) causal model of the given shape. But a causal model has one equation corresponding to each variable, and the corresponding hypergraphs thus has exactly one hyperarc leading to it. In the dual hypergraphs, one should view the nodes as processes and the arcs as wires. Such a hypergrah has precisely one hyperarc leading out of every node. When wires branch, one imagines a copy; when two arcs point to the same process (as in process 4, in the middle center), that process takes both of the wires as inputs. In the duals of hypergraphs corresponding to causal models, there are no two-tailed arrows, which might be thought of as a "merge". Yet it is not clear how to merge the values of two variables, when they are not the same, in general—especially if we expect associativity and commutativity, as we do with *copy*.

((What can be done with these objects?))

2.0.5 Categories

Category theory is a mathematical inderlingua 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 2.0.11 (category). A *category* $\mathcal{C}$ consists of four pieces of data:

- a collection of *objects* $\mathrm{ob}_{\mathcal{C}}$;
- a collection of *morphisms* $\mathrm{Hom}_{\mathcal{C}}(X, Y)$, also written $\mathcal{C}(X, Y)$, for each pair of objects $(X, Y) \in \mathrm{ob}_{\mathcal{C}}^2$;
- a *composition* operator $\circ_{X,Y,Z} : \mathcal{C}(Y, Z) \times \mathcal{C}(X, Y) \to \mathcal{C}(X, Z)$ for each triple $(X, Y, Z) \in \mathrm{ob}_{\mathcal{C}}^3$, 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* $\mathrm{id}_X \in \mathcal{C}(X, X)$ for each object $X \in \mathcal{C}$, satisfying $\mathrm{id}_X \circ f = f$ and $g \circ \mathrm{id}_X = g$ for any morphism $f \in \mathcal{C}(Y, X)$ or $g \in \mathcal{C}(X, Y)$ for some $Y \in \mathrm{ob}_{\mathcal{C}}$. □

Common examples of categories include:

- $\mathbb{S}$**et**, the category whose objects are sets and whose morphisms are functions between them,
- $\mathbb{T}$**op**, the category whose objects are topological spaces and whose morphisms are continuous maps,
- $\mathbb{R}$**el**, the category whose objects are sets, and whose morphisms are relations, and

- $\mathbb{D}$**iff**, the category of smooth manifolds (possibly with boundary or corners) and differentiable maps.

All of these are also known as “conrete categories”, because they all build on $\mathbb{S}$et: their objects can be interpereted as sets, and their morphisms interprereted as functions. But categories can also be much more combinitorial 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.

What will be most relevant for our purposes is a construction

Definition 2.0.12 (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,

$$\mathrm{ob}_{G^*} = N, \qquad G^*(x,y) = \left\{ \text{sequences } \langle a_1, \ldots, a_n \rangle \;\middle|\; \begin{array}{c} n \in \mathbb{N}, \;\; n > 0 \Rightarrow (S_{a_1} = x \wedge T_{a_n} = y), \\ \forall i \in \{1, \ldots, n-1\}.\, T_{a_i} = S_{a_{i+1}} \end{array} \right\},$$

with composition given by sequence concatenation, and the identity being the empty sequence. □

The superscript-star notation has some standard meanings throughout mathematics, and this construction in Definition 2.0.12 reduces to several of them in the appropriate contexts.

- A (multi) graph $G = (\{*\}, A)$ with one vertex can be identified with its arc set A. Every arrow has the same type $(* \to *)$, and so a path is a sequence $\langle a_1, a_2, ..., a_n \rangle$ where each $a_i \in A$. So in this case, G^* (as given by Definition 2.0.12) 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 $\langle u{=}u_1, \ldots, u_n{=}v \rangle$ 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 $i \to j$. The free category G^* generated by these arrows (per Definition 2.0.12) 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:

  $$A^* = \sum_{n=0}^{\infty} A^n \in \overline{\mathbb{N}}^{n \times n}, \qquad \text{where } \overline{\mathbb{N}} := \mathbb{N} \cup \{\infty\}.$$

  And, yet again, we have $\#G^*(i, j) = (\mathbb{A}_G^*)_{i,j}$

What about hypergraphs? The free category generated by a directed hypergraph $(N, \mathcal{A})$ is the free category generated by

Definition 2.0.13. □

2.0.6 Measures and Probabilities

Definition 2.0.14 (Measurable Space)**.** A measurable space is a pair $(X, \mathcal{F}_X)$, where X is a set, and $\mathcal{F}_X$ is a sigma-algebra over X, which is to say a set of subsets of X, containing the empty set, and closed under countable union, intersection, and complement with respect to X. The elements of $\mathcal{F}_X$ are referred to as measurable sets. □

Definition 2.0.15 (Measure)**.** A *measure* λ over a measurable space $(X, \mathcal{F})$ is a function $\lambda : \mathcal{F} \to \mathbb{R} \cup \{\infty\}$, with the follwing properties.

- **Null Empty Set.** $\lambda(\emptyset) = 0$.
- **Non-negativity.** $\lambda(U) \geq 0$ for all $U \in \mathcal{F}$.
- **Countable additivity.** For every coutable collection $\{U_i\}_{i=1,2,\ldots}$ of pairwise disjoint measurable sets ($U_i \in \mathcal{F}$), we have $\sum_i \lambda(U_i) = \lambda(\sqcup_i U_i)$.

□

Definition 2.0.16. Consider a measure λ on a mesurable space $\mathcal{X} = (X, \mathcal{F})$.

1. If $\lambda(X) = 1$, then λ is a *probability* measure.

2. If $\mathcal{T}$ is a topology on X, and $\lambda(U) > 0$ for every non-empty open set $U \in \mathcal{F} \cap \mathcal{T}$, then λ is said to be *strictly positive* (wrt $\mathcal{T}$).

3. The measure λ is called *σ-finite* if X can be covered with a countable set of sets with finite measures — that is, if there exist countable sequence $(A_1, A_2, \ldots) \subset \mathcal{F}$ the such that each $\lambda(A_i) < \infty$ is finite, and $X = \cup_{i=1}^{\infty} A_i$. □

Definition 2.0.17 (Measurable Functions)**.** If $\mathcal{X} = (X, \mathcal{F})$, and $\mathcal{Y} = (Y, \mathcal{G})$ are two measurable spaces, a measurable function $f : \mathcal{X} \to \mathcal{Y}$ is a function $f : X \to Y$ on the underlying spaces, such that $f^{-1}(U) \in \mathcal{F}$ for every $U \in \mathcal{G}$. □

It is easy to verify that identity maps are measurable, and that, when f and g are measurable, so is $f \circ g$. It follows that there is a category $\mathbb{M}\text{eas}$ whose objects are measurable spaces, and whose maps are measurable functions.

Definition 2.0.18 (absolute continuity)**.** If μ and ν are measures over a space $(X, \mathcal{F})$, we say that μ is absolutely continuous with respect to ν, denoted $\nu \ll \mu$, if, for every $U \in \mathcal{F}$ such that $\nu(U) = 0$, we also have $\mu(U) = 0$. □

Definition 2.0.19 (Radon-Nikodym Derivative)**.** Suppose μ and ν are both measures over a measurable space $(X, \mathcal{F})$, and $\mu \ll \nu$. The Radon-Nikodym theorem states that there is then a unique $\mathcal{F}$-measurable function f such that for all $A \in \mathcal{F}$,

$$\mu(A) = \int_A f \, \mathrm{d}\nu.$$

The function f is called the Radon-Nikodym derivative, and denoted $\frac{\mathrm{d}\mu}{\mathrm{d}\nu} := f$. □

Definition 2.0.20 (Markov Kernels)**.** If $\mathcal{X} = (X, \mathcal{F})$, and $\mathcal{Y} = (Y, \mathcal{G})$ are two measurable spaces, a Markov Kernel $\kappa : \mathcal{X} \to \mathcal{Y}$, which we sometimes write as "$\kappa(Y|X)$", is a function $\kappa : X \times \mathcal{G} \to \mathbb{R}$, such that

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

□

Definition 2.0.21 (category of measurable spaces and Markov kernels)**.** Let $\mathbb{S}\mathbf{toch}$ be the category whose objects are measureable 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 $\mathbb{S}\mathbf{toch}$ are pairs $(\mathcal{X}, \lambda)$, where $\mathcal{X}$ is a mesurable topological space, and λ is a strictly positive and σ-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 $\kappa : \mathcal{X} \to \mathcal{Y}$ such that $\kappa(x, -) \ll \lambda_Y$ for all x. The reason we require this is so that the Radon-Nikodym derivative $\frac{\mathrm{d}\kappa(x)}{\mathrm{d}\lambda}$, i.e., the unique $\mathcal{F}_Y$-measurable function satisfying

$$\forall x. \forall A \in \mathcal{F}_Y. \qquad \kappa(x, A) = \int_A \frac{\mathrm{d}\kappa(x)}{\mathrm{d}\lambda} \mathrm{d}\lambda\,, \qquad \text{exists.}$$

Composition in $\mathbb{S}\mathbf{toch}$ is given by Lebesgue Integration: for Markov Kernels $p(Y|X) : \mathcal{X} \to \mathcal{Y}$ and $q(Z|Y) : \mathcal{Y} \to \mathcal{Z}$, define $(p \circ q) : \mathcal{X} \to \mathcal{Z}$ (i.e., $p \circ q : X \times \mathcal{F}_Z \to [0, 1]$) by:

$$(p \circ q)(x, U) := \int_{\mathcal{Y}} q(-, U) \mathrm{d}p(x, -).$$

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

$$\mathrm{id}_{\mathcal{X}}(x, U) = \begin{cases} 1 & \text{if } x \in U \\ 0 & \text{otherwise} \end{cases} .$$

These functions are clearly identities, but are they Markov kernels, and can they be made absolutely continuous with respect to our base measure?

More explicitly:

$$\mathrm{id}_{\mathcal{X}}(x, -) \ll \lambda_X \qquad \Longleftrightarrow \qquad \Big(\lambda_X(A) = 0 \;\;\Rightarrow\;\; x \notin A \Big),$$

and so there is a problem if we can find $A \subset \mathcal{V}X$ with $\lambda_X(A) = 0$ but $x \in A$. Or equivalently, a non-empty subset $A \subseteq \mathcal{V}X$ that has measure zero. By strict positivity, this means A cannot be an open set.

In the discrete case, in which every variable can take a finite set of values, and every subset is measurable and clopen, this is not a problem so long as the base measure gives every element positive probability.

□

2.0.7 Independencies

2.0.8 Information Theory

2.0.9 Graphical Models

There are two aspects any graphical model: a "qualitative/structural" aspect, which describes influences between variables, and a "quantitative/observational" aspect, that annotates those influences with data.

A qualitative BN, for example, is a directed graph whose semantics are given in terms of independencies: any variable X is independent of its non-descendents, given the values of its parents, $\mathbf{Pa}\,X$. A quantitative BN, then, includes both that directed graph, and also each variable X to a conditional probability distribution $\Pr_X(X|\mathbf{Pa}\,X)$.

Part I

A Universal Modeling Language

CHAPTER 3

PROBABILISTIC DEPENDENCY GRAPHS (PDGS)

RELATIVE ENTROPY SOUP

Oliver E. Richardson, Ph.D.

Cornell University 2024

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

In this paper we introduce yet another graphical tool for modeling beliefs, *Probabilistic Dependency Graphs* (PDGs). There are already many such models in the literature, including Bayesian networks (BNs) and factor graphs. (For an overview, see Koller and Friedman [19].) 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 union 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 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 probility .1, the law prohibits floomps.

The obvious way to represent this as a BN is to use two random 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.

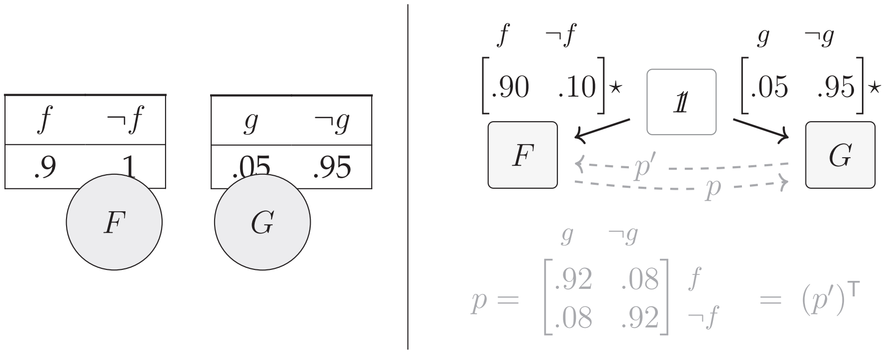


Figure 3.1: A BN (left) and corresponding PDG (right), which can be augmented with additional cpds. The cpds p and/or p' make it inconsistent.

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 *c*onditional *p*robability *d*istribution (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 edges, 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 edges from $\mathbb{1}$ to F and G are associated with the unconditional probabilities of F and G, and the edges between F and G are associated with p and p'.

The original state of knowledge consists of all three nodes and the two solid edges 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 edge from F to G) simply by adding the edge from F to G and the associated cpd p (the new infromation is shown in blue). Doing so does not change the meaning of the original edges. Unlike a Bayesian update, the operation is even reversible: all we need to do recover our original belief state is delete the new edge, making it possible to mull over and then reject an observation. □

The ability of PDGs to model inconsistency, as illustrated in Example 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 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 $\mathit{pdg}(\mathcal{B})$, shown as the solid nodes and edges in Figure 3.2b.

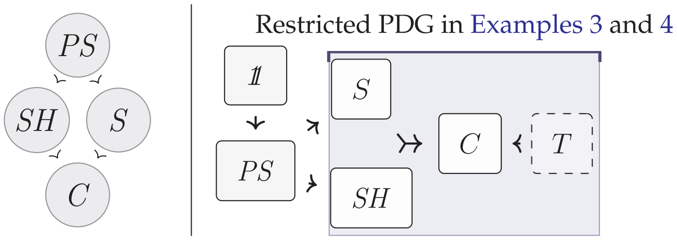


Figure 3.2: (a) The Bayesian Network $\mathcal{B}$ in Example 2 (left), and (b) $\mathit{pdg}(\mathcal{B})$, its corresponding PDG (right). The shaded box indicates a restriction of $\mathit{pdg}(\mathcal{B})$ to only the nodes and edges it contains, and the dashed node T and its arrow to C can be added in the PDG, without taking into account S and SH.

We start with the nodes corresponding to the variables in $\mathcal{B}$, together with the special node $\mathbb{1}$ from Example 1; we add an edge 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 edges $PS \to S$ and $PS \to SH$ in $\mathit{pdg}(\mathcal{B})$. 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 assumptions in the BN.

For (1), we cannot just add the edges $S \to C$ and $SH \to C$ that are present in $\mathcal{B}$. As we saw in Example 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 edge with two tails—a (directed) *hyperedge*—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 independences presumed by $\mathcal{B}$ need not hold for them. Rather than trying to distinguish between them with additional constraints, we develop a a scoring-function semantics for PDGs which is in this case uniquely minimized by the distribution specified by $\mathcal{B}$ (Theorem 3.4.1). 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, $\boldsymbol{\mathit{pdg}}(\mathcal{B})$, on the other hand, has no trouble: We can simply add the node T with an edge to C that is associated with $\mathbb{P}$. 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. □

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

Example 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 [19], 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 $A \to B \to C$, 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 edges 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. □

Being able to form a well-behaved local picture and restrict knowledge is useful, but an even more compelling reason to use PDGs is their ability to aggregate information.

Example 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

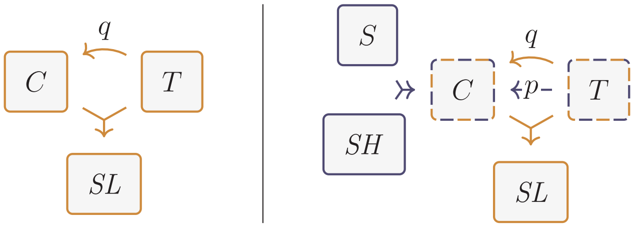


Figure 3.3: Grok's prior (left) and combined (right) knowledge.

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.

Grok is reading about graphical models because she vaguely remembers that the variables in Example 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 $\mathbb{P}$ in Example 2), but before she has time to reflect, we can represent her (conflicted) knowledge state as the union 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 edges from T to C, with inconsistent information. Had we not allowed multigraphs, we would have needed to choose between the two edges, 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. □

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 formal definitions for PDGs. Although it is possible to formalize PDGS with hyperedges directly, we opt for a different approach here, in which PDGs have only regular edges, and hyperedges are captured using a simple construction that involves adding an extra node. [1]

Definition 3.2.1. A *Probabilistic Dependency Graph* is a tuple $\mathcal{m} = (\mathcal{N}, \mathcal{A}, \mathcal{V}, \mathbb{P}, \alpha, \beta)$, where

$\mathcal{N}$ $: \mathbb{S}\mathbf{et}$ is a finite set of nodes, corresponding to variables;

$\mathcal{A}$ $\subseteq \mathcal{N} \times \mathcal{N} \times \mathit{Label}$ is a set of labeled edges $\{X \xrightarrow{a} Y\}$, each with a source S and target T in $\mathcal{N}$;

$\mathcal{V}$ $\mathcal{N} \to \mathbf{Set}$ associates each variable $N \in \mathcal{N}$ with a set $\mathcal{V}(N)$ of values that the variable N can take;

[1] In the factor graph literature, especially with regard to loopy belief propagation [**?**], it is common to call a collection of marginals that are not necessarily all compatible with a distribution *pseudomarginals*, making a PDG in some sense a collection of ‘conditional’ pseudomarginals. This gives an alternate expansion of “PDG” as “Pseudomarginal Dependency Graph”, with nomenclature rooted in the literature.

$\mathbb{P} : \big((A,B,\ell)\colon \mathcal{A}\big) \to \mathcal{V}(A) \to \Delta\mathcal{V}(B)$ associates to each edge $X \xrightarrow{a} Y \in \mathcal{A}$ a distribution $\mathbb{P}_a(x)$ on Y for each $x \in \mathcal{V}(X)$;

$\alpha : \mathcal{A} \to [0,1]$ associates to each edge $X \xrightarrow{a} Y$ a non-negative number α_L which, roughly speaking, is the modeler's confidence in the functional dependence of Y on X implicit in L;

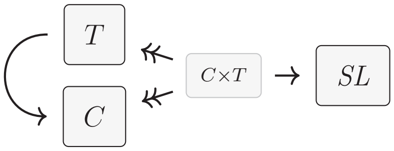
$\beta : \mathcal{A} \to \mathbb{R}^+$ associates to each edge L a positive real number β_L, the modeler's subjective confidence in the reliability of $\mathbb{P}$.

Note that we allow multiple edges in $\mathcal{A}$ with the same source and target; thus $(\mathcal{N}, \mathcal{A})$ is a multigraph. We occasionally write a PDG as $\boldsymbol{m} = (\mathcal{G}, \mathbb{P}, \alpha, \beta)$, where $\mathcal{G} = (\mathcal{N}, \mathcal{A}, \mathcal{V})$, and abuse terminology by referring to $\mathcal{G}$ as a multigraph. We refer to $\boldsymbol{n} = (\mathcal{G}, \mathbb{P})$ as an *unweighted* PDG, and give it semantics as though it were the (weighted) PDG $(\mathcal{G}, \mathbb{P}, \mathbf{1}, \mathbf{1})$, where $\mathbf{1}$ is the constant function (i.e., so that $\alpha_L = \beta_L = 1$ for all L). In this paper, with the exception of Section 3.4.3, we implicitly take $\alpha = \mathbf{1}$ and omit α, writing $\boldsymbol{m} = (\mathcal{G}, \mathbb{P}, \beta)$.[2] □

If $\boldsymbol{m}$ is a PDG, we reserve the names $\mathcal{N}^{\boldsymbol{m}}, \mathcal{A}^{\boldsymbol{m}}, \ldots$, for the components of $\boldsymbol{m}$, so that we may reference one without naming them all explicitly. We write $\mathcal{V}(S)$ for the set of possible joint settings of a set S of variables, and write $\mathcal{V}(\boldsymbol{m})$ for all settings of the variables in $\mathcal{N}^{\boldsymbol{m}}$; we refer to these settings as "worlds". While the definition above is sufficient to represent the class of all legal PDGs, we often use two additional bits of syntax to indicate common constraints: the special variable $\mathbb{1}$ such that $\mathcal{V}(\mathbb{1}) = \{\star\}$ from Examples 1 and 2, and double-headed arrows, $A \twoheadrightarrow B$, which visually indicate that the corresponding cpd is degenerate, effectively representing a deterministic function $f : \mathcal{V}(A) \to \mathcal{V}(B)$.

[2] The appendix gives results for arbitrary α.

Construction 3.2.2. We can now explain how we capture the multi-tailed edges that were used in Examples 2 to 4. That notation can be viewed as shorthand for the graph that results by adding a new node at the junction representing the joint value of the nodes at the tails, with projections going back. For instance, the diagram displaying Grok's prior knowledge in Example 4, on the left of Figure 3.3 is really shorthand for the following PDG, where 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 any joint setting $(c, t) \in \mathcal{V}(C \times T)$ of both variables, the cpd for the edge from $C \times T$ to C gives probability 1 to c; similarly, the cpd for the edge from $C \times T$ to T gives probability 1 to t. □

3.3 Semantics

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 μ, and can be thought of as a *weighted* set of distributions [9]. 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.

Definition 3.3.1. If $\mathcal{M}$ is a PDG (weighted or unweighted) with edges $\mathcal{A}$ and cpds $\mathbb{P}$, let $\{\!\!\{\mathcal{M}\}\!\!\}$ be the *s*et of *d*istributions over the variables in $\mathcal{M}$ whose conditional marginals are exactly those given by $\mathbb{P}$. That is, $\mu \in \{\!\!\{\mathcal{M}\}\!\!\}$ iff, for all edges $a \in \mathcal{A}$ from X to Y, $x \in \mathcal{V}(X)$, and $y \in \mathcal{V}(Y)$, we have that $\mu(Y\!=\!y \mid X\!=\!x) = \mathbb{P}_a(T|s)$. Formally,

$$\{\!\!\{\mathcal{M}\}\!\!\} = \left\{ \mu \in \Delta\mathcal{V}(\mathcal{M}) \;\middle|\; \begin{array}{l} \mu(B\!=\!b \mid A\!=\!a) \geq \mathbb{P}_a(b \mid a) \\ \forall (A, B, \ell) \in \mathcal{A},\, a \in \mathcal{V}(A),\, b \in \mathcal{V}(B) \end{array} \right\}$$

$\mathcal{M}$ is *inconsistent* if $\{\!\!\{\mathcal{M}\}\!\!\} = \emptyset$, and *consistent* otherwise. □

Note that $\{\!\!\{\mathcal{M}\}\!\!\}$ is independent of the weights α and β.

3.3.2 PDGs As Distribution Scoring Functions

We now generalize the previous semantics by viewing a PDG $\mathcal{M}$ as a *scoring function* that, given an arbitrary distribution μ on $\mathcal{V}(\mathcal{M})$, returns a real-valued score indicating how well μ fits $\mathcal{M}$. Distributions with the lowest (best) scores are those that most closely match the cpds in $\mathcal{M}$, and contain the fewest unspecified correlations.

We start with the first component of the score, which assigns higher scores to distributions that require a larger perturbation in order to be consistent with

$\mathcal{m}$. We measure the magnitude of this perturbation with relative entropy. In particular, for an edge $X \xrightarrow{a} Y$ and $x \in \mathcal{V}(X)$, we measure the relative entropy from $\mathbb{P}_a(x)$ to $\mu(Y = \cdot \mid X = x)$, and take the expectation over μ_X (that is, the marginal of μ on X). We then sum over all the edges L in the PDG, weighted by their reliability.

Definition 3.3.2. For a PDG $\mathcal{m}$, the *incompatibility* of a a joint distribution μ over $\mathcal{V}(\mathcal{m})$, is given by

$$Inc_{\mathcal{m}}(\mu) := \sum_{\mathcal{A}LX\ Y \in \mathcal{A}^{\mathcal{m}}} \beta_L^{\mathcal{m}} \mathop{\mathbb{E}}_{x \sim \mu_X} \left[\boldsymbol{D}\Big(\mu(Y \mid X{=}x) \,\Big\|\, \mathbb{P}_a^{\mathcal{m}}(x) \Big) \right],$$

where $\boldsymbol{D}(\mu \parallel \nu) = \sum_{w \in \mathrm{Supp}(\mu)} \mu(w) \log \frac{\mu(w)}{\nu(w)}$ is the relative entropy from ν to μ. The *inconsistency of PDG* $\mathcal{m} = (\mathcal{N}, \mathcal{A}, \mathbb{P}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, denoted $Inc(\mathcal{m})$, is the minimum possible incompatibility of $\mathcal{m}$ with any distribution μ,

$$Inc(\mathcal{m}) = \inf_{\mu \in \Delta[W_{\mathcal{V}}]} Inc_{\mathcal{m}}(\mu).$$

□

The idea behind this definition of inconsistency is that we want to choose a distribution μ that minimizes the total number of bits required to encode all of the relevant conditional marginals. More precisely, fix a distribution μ. For each edge $L = (X, Y, \ell) \in \mathcal{A}$ and $x \in \mathcal{V}(X)$, we are given a code for Y optimized for the distribution $\mathbb{P}_a(x)$, and asked to transmit data from $\mu(Y \mid x)$; we incur a cost for each bit required beyond what we would have used had we used a code optimized for the actual distribution $\mu(Y \mid X = x)$. To obtain the cost for L, we take a weighted average of these costs, where the weight for the value x is the probability $\mu_X(x)$. We do this for every edge $L \in \mathcal{A}$, summing the cost.

For even more intuition, imagine two agents (A and B) with identical beliefs described by a PDG $\mathcal{m}$ about a set of variables that are in fact distributed according to μ. For each edge $L = (X, Y, \ell) \in \mathcal{A}^{\mathcal{m}}$, values $x, y \in \mathcal{V}(X)$ are chosen according to μ_{XY} and x is given to both agents.

At this point, the agents, having the same conditional beliefs, and the same information about Y, agree on the optimal encoding of the possible values of Y as sequences of bits, so that if y were drawn from $\mathbb{P}_a(x)$, the fewest number of bits would be needed to communicate it in expectation. The value of y—which is distributed not according to $\mathbb{P}_a(x)$, but $\mu(Y \mid X = x)$—is now given to agent A. The agents pay a cost equal the number of bits needed to encode y according to the agreed-upon optimal code, but reimbursed the (smaller) cost that would have been paid, had the agents beliefs lined up with the true distribution μ.

Repeating for each edge and summing the expectations of these costs, we can view $\mathit{Inc}_{\mathcal{m}}(\mu)$ as the total number of *additional* expected bits required to communicate y with a code optimized for $\mathbb{P}_a(x)$ instead of the true conditional distribution $\mu(Y \mid X = x)$.

If $\mathcal{m}$ is inconsistent, then there will be a cost no matter what distribution μ is the true distribution. Conversely, if $\mathcal{m}$ is consistent, then any distribution $\mu \in \{\!\{\mathcal{m}\}\!\}$ will have $\mathit{Inc}_{\mathcal{m}}(\mu) = 0$.

Example ?? (continuing from p. **??**)**.** Recall our simplest example, which directly encodes an entire distribution p over the set W. In this case, there is only one edge, the expectation is over a single element, and the marginal on W is the entire distribution. Therefore, $\mathit{Inc}(\mathcal{m}; \mu) = \boldsymbol{D}(\mu \parallel \mu)$, so the inconsistency is just the information μ and p, so is minimized uniquely when μ is p □

$\{\!\{\mathcal{m}\}\!\}$ and $\mathit{Inc}_{\mathcal{m}}$ distinguish between distributions based on their compatibility with $\mathcal{m}$, but even among distributions that match the marginals, some more closely match the qualitative structure of the graph than others. We think of each edge $X\xrightarrow{a}Y$ as representing a qualitative claim (with confidence α_L) that the value of Y can be computed from X alone. To formalize this, we require only the multigraph $\mathcal{G}^{\mathcal{m}}$.

Given a multigraph G and distribution μ on its variables, contrast the amount of information required to

(a) directly describe a joint outcome $\mathbf{w}$ drawn from μ, and

(b) separately specify, for each edge $X\xrightarrow{a}Y$, the value $\mathbf{w}_Y$ (of Y in world $\mathbf{w}$) given the value $\mathbf{w}_X$, in expectation.

If (a) = (b), a specification of (b) has exactly the same length as a full desciption of the world. If (b) > (a), then there are correlations in μ that allow for a more compact representation than G provides. The larger the difference, the more information is needed to determine targets Y beyond the conditional probabilities associated with the edges $X \to Y$ leading to Y (which according to G should be sufficient to compute them), and the poorer the qualitative fit of μ to G. Finally, if (a) > (b), then μ requires additional information to specify, beyond what is necessary to determine outcomes of the marginals selected by G.

Definition 3.3.3. For a multigraph $G = (\mathcal{N}, \mathcal{A}, \mathcal{V})$ over a set $\mathcal{N}$ of variables, define the *G-information deficiency* of distribution μ, denoted $\mathit{IDef}_G(\mu)$, by considering the difference between (a) and (b), where we measure the amount of information needed for a description using entropy:

$$\mathit{IDef}_G(\mu) := \sum_{(X,Y)\in\mathcal{A}} \mathrm{H}_\mu(Y \mid X) - \mathrm{H}(\mu). \tag{3.1}$$

(Recall that $H_\mu(Y \mid X)$, the (μ-)*conditional entropy of* Y *given* X, is defined as $-\sum_{x,y \in \mathcal{V}(X,Y)} \mu(x,y) \log \mu(y \mid x)$.) For a PDG $\mathcal{m}$, we take $\mathit{IDef}_{\mathcal{m}} = \mathit{IDef}_{\mathcal{G}^{\mathcal{m}}}$. $\square$

We illustrate $\mathit{IDef}_{\mathcal{m}}$ with some simple examples. Suppose that $\mathcal{m}$ has two nodes, X and Y. If $\mathcal{m}$ has no edges, the $\mathit{IDef}_{\mathcal{m}}(\mu) = -\operatorname{H}(\mu)$. There is no information required to specify, for each edge in $\mathcal{m}$ from X to Y, the value w_Y given w_X, since there are no edges. Since we view smaller numbers as representing a better fit, $\mathit{IDef}_{\mathcal{m}}$ in this case will prefer the distribution that maximizes entropy. If $\mathcal{m}$ has one edge from X to Y, then since $\operatorname{H}(\mu) = \operatorname{H}_\mu(Y \mid X) + \operatorname{H}_\mu(X)$ by the well known *entropy chain rule* [24], $\mathit{IDef}_{\mathcal{m}}(\mu) = -\operatorname{H}_\mu(X)$. Intuitively, while knowing the conditional probability $\mu(Y \mid X)$ is helpful, to completely specify μ we also need $\mu(X)$. Thus, in this case, $\mathit{IDef}_{\mathcal{m}}$ prefers distributions that maximize the entropy of the marginal on X. If $\mathcal{m}$ has sufficiently many parallel edges from X to Y and $\operatorname{H}_\mu(Y \mid X) > 0$ (so that Y is not totally determined by X) then we have $\mathit{IDef}_{\mathcal{m}}(\mu) > 0$, because the redundant edges add no information, but there is still a cost to specifying them. In this case, $\mathit{IDef}_{\mathcal{m}}$ prefers distributions that make Y a deterministic function of X will maximizing the entropy of the marginal on X. Finally, if $\mathcal{m}$ has an edge from X to Y and another from Y to X, then a distribution μ minimizes $\mathit{IDef}_{\mathcal{m}}$ when X and Y vary together (so that $\operatorname{H}_\mu(Y \mid X) = \operatorname{H}_\mu(X \mid Y) = 0$) while maximizing $\operatorname{H}(\mu)$, for example, by taking $\mu(0,0) = \mu(1,1) = 1/2$.

$\mathit{Inc}_{\mathcal{m}}(\mu)$ and $\mathit{IDef}_{\mathcal{m}}(\mu)$ give us two measures of compatibility between $\mathcal{m}$ and a distribution μ. We take the score of interest to be their sum, with the tradeoff controlled by a parameter $\gamma \geq 0$:

$$[\![\mathcal{m}]\!]_\gamma(\mu) := \mathit{Inc}_{\mathcal{m}}(\mu) + \gamma \mathit{IDef}_{\mathcal{m}}(\mu) \tag{3.2}$$

The following just makes precise that the scoring semantics generalizes the first semantics.

[link to proof]

Proposition 3.3.1. $\{\!\!\{ \mathcal{m} \}\!\!\} = \{\mu : \llbracket \mathcal{m} \rrbracket_0(\mu) = 0\}$ *for all* $\mathcal{m}$*.*

While we focus on this particular scoring function in the paper, in part because it has deep connections to the free energy of a factor graph [19], other scoring functions may well end up being of interest.

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

$$\llbracket \mathcal{m} \rrbracket^*_\gamma = \arg\min_{\mu \in \Delta \mathcal{V}(\mathcal{m})} \llbracket \mathcal{m} \rrbracket_\gamma(\mu). \tag{3.3}$$

In general, $\llbracket \mathcal{m} \rrbracket^*_\gamma$ does not give a unique distribution. But if γ is sufficiently small, then it does:

[link to proof]

Proposition 3.3.2. *If* $\mathcal{m}$ *is a PDG and* $0 < \gamma \leq \min_L \beta^{\mathcal{m}}_L$*, then* $\llbracket \mathcal{m} \rrbracket^*_\gamma$ *is a singleton.*

In this paper, we are interested in the case where γ is small; this amounts to emphasizing the accuracy of the probability distribution as a description of probabilistic information, rather than the graphical structure of the PDG. This motivates us to consider what happens as γ goes to 0. If S_γ is a set of probability distributions for all $\gamma \in [0,1]$, we define $\lim_{\gamma\to 0} S_\gamma$ to consist of all distributions μ such that there is a sequence $(\gamma_i, \mu_i)_{i\in\mathbb{N}}$ with $\gamma_i \to 0$ and $\mu_i \to \mu$ such that $\mu_i \in S_{\gamma_i}$ for all i. It can be further shown that

link to proof

Proposition 3.3.3. *For all $\mathcal{M}$, $\lim_{\gamma\to 0}[\![\mathcal{M}]\!]^*_\gamma$ is a singleton.*

Let $[\![\mathcal{M}]\!]^*$ be the unique element of $\lim\limits_{\gamma\to 0}[\![\mathcal{M}]\!]^*_\gamma$. The semantics has an important property:

link to proof

Proposition 3.3.4. *$[\![\mathcal{M}]\!]^* \in [\![\mathcal{M}]\!]^*_0$, so if $\mathcal{M}$ is consistent, then $[\![\mathcal{M}]\!]^* \in \{\!\{\mathcal{M}\}\!\}$.*

3.4 Relationships to Other Graphical Models

We start by relating PDGs to two of the most popular graphical models: BNs and factor graphs. PDGs are strictly more general than BNs, and can emulate factor graphs for a particular value of γ.

3.4.1 Bayesian Networks

Construction 3.2.2 can be generalized to convert arbitrary Bayesian Networks into PDGs. Given a BN $\mathcal{B}$ and a positive confidence β_X for the cpd of each variable X of $\mathcal{B}$, let $\boldsymbol{\mathit{pdg}}(\mathcal{B}, \beta)$ be the PDG comprising the cpds of $\mathcal{B}$ in this way; we defer the straightforward formal details to the appendix.

link to proof

Theorem 3.4.1. *If $\mathcal{B}$ is a Bayesian network and $\Pr_{\mathcal{B}}$ is the distribution it specifies, then for all $\gamma > 0$ and all vectors β such that $\beta_L > 0$ for all edges L, $[\![\boldsymbol{pdg}(\mathcal{B}, \beta)]\!]^*_\gamma = \{\Pr_{\mathcal{B}}\}$, and thus $[\![\boldsymbol{pdg}(\mathcal{B}, \beta)]\!]^* = \Pr_{\mathcal{B}}$.*

Theorem 3.4.1 is quite robust to parameter choices: it holds for every weight vector β and all $\gamma > 0$. However, it does lean heavily on our assumption that $\alpha = \mathbf{1}$, making it our only result that does not have a natural analog for general α.

3.4.2 Factor Graphs

Factor graphs [21], like PDGs, generalize BNs. In this section, we consider the relationship between factor graphs (FGs) and PDGs.

Definition 3.4.1. A *factor graph* Φ is a set of random variables $\mathcal{X} = \{X_i\}$ and *factors* $\{\phi_J \colon \mathcal{V}(X_J) \to \mathbb{R}_{\geq 0}\}_{J \in \mathcal{J}}$, where $X_J \subseteq \mathcal{X}$. More precisely, each factor ϕ_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. Φ specifies a distribution

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

where $\vec{x}$ is a joint setting of all of the variables, $\vec{x}_J$ is the restriction of $\vec{x}$ to only the variables X_J, and Z_Φ is the constant required to normalize the distribution. □

The cpds of a PDG naturally constitute a collection of factors, so it natural to wonder how the semantics of a PDG compares to simply treating the cpds as factors in a factor graph. To answer this, we start by making the translation precise.

Definition 3.4.2 (unweighted PDG to factor graph)**.** If $\mathcal{M} = (\mathcal{G}, \mathbb{P})$ is an unweighted PDG, define the associated FG $\Phi_{\mathcal{M}}$ on the variables $(\mathcal{N}, \mathcal{V})$ by taking $\mathcal{J}$ to be the set of edges, and for an edge L from Z to Y, taking $X_L = \{Z, Y\}$, and $\phi_L(z, y)$ to be $\mathbb{P}_a^{\mathcal{M}}(y \mid z)$ (i.e., $(\mathbb{P}_a^{\mathcal{M}}(z))(y)$). □

It turns out we can also do the reverse. Using essentially the same idea as in Construction 3.2.2, we can encode a factor graph as an assertion about the unconditional probability distribution over the variables associated to each factor.

Definition 3.4.3 (factor graph to unweighted PDG)**.** For a FG Φ, let $\boldsymbol{\mathit{updg}}(\Phi)$ be the unweighted PDG consisting of

- the variables in Φ together with $\mathbb{1}$ and a variable $X_J := \prod_{j \in J} X_j$ for every factor $J \in \mathcal{J}$, and
- edges $\mathbb{1} \!\to\! X_J$ for each J and $X_J \!\twoheadrightarrow\! X_j$ for each $X_j \in \mathbf{X}_J$,

where the edges $X_J \twoheadrightarrow X_j$ are associated with the appropriate projections, and each $\mathbb{1} \to X_J$ is associated with the unconditional joint distribution on X_J obtained by normalizing ϕ_J. The process is illustrated in Figure 3.4. □

PDGs are directed graphs, while factors graphs are undirected. The map from PDGs to factor graphs thus loses some important structure. As shown in Figure 3.4, this mapping can change the graphical structure significantly. Nevertheless,

link to proof

Theorem 3.4.2. $\Pr_\Phi = [\![\boldsymbol{\mathit{updg}}(\Phi)]\!]^*_1$ *for all factor graphs* Φ.[3]

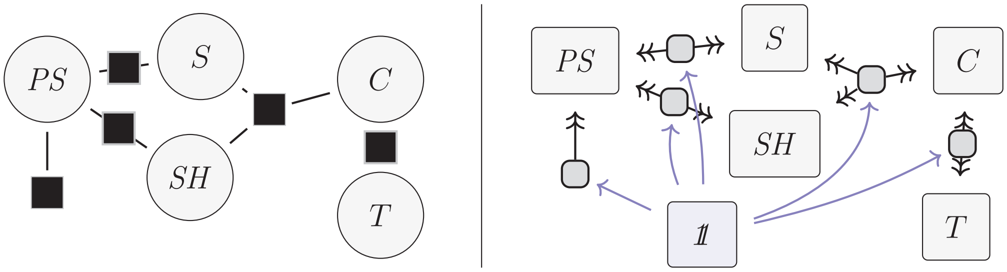


Figure 3.4: Conversion of the PDG in Example 2 to a factor graph according to Definition 3.4.2 (left), and from that factor graph back to a PDG by Definition 3.4.3 (right). In the latter, for each J we introduce a new variable X_J (displayed as a smaller darker rectangle), whose values are joint settings of the variables connected it, and also an edge $1 \to X_J$ (shown in blue), to which we associate the unconditional distribution given by normalizing ϕ_J.

(unproven!)

Theorem 3.4.3. $[\![\mathcal{n}]\!]^*_1 = \Pr_{\Phi_{\mathcal{n}}}$ *for all unweighted PDGs* $\mathcal{n}$*.*

The correspondence hinges on the fact that we take $\gamma = 1$, so that Inc and IDef are weighted equally. Because the user of a PDG gets to choose γ, the fact that the translation from factor graphs to PDGs preserves semantics only for $\gamma = 1$ poses no problem. Conversely, the fact that the reverse correspondence requires $\gamma = 1$ suggests that factor graphs are less flexible than PDGs.

What about weighted PDGs $(\mathcal{G}, \mathbb{P}, \beta)$ where $\beta \neq \mathbf{1}$? There is also a standard notion of weighted factor graph, but as long as we stick with our convention of taking $\alpha = \mathbf{1}$, we cannot relate them to weighted PDGs. As we are about to see, once we drop this convention, we can do much more.

[3] Recall that we identify the unweighted PDG $(\mathcal{G}, \mathbf{p})$ with the weighted PDG $(\mathcal{G}, \mathbb{P}, \mathbf{1}, \mathbf{1})$.

3.4.3 Factored Exponential Families

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

$$\mathit{GFE}_\Psi(\mu) := \mathop{\mathbb{E}}_{\vec{x} \sim \mu} \left[\sum_{J \in \mathcal{J}} \theta_J \log \frac{1}{\phi_J(\vec{x}_J)} \right] - \mathrm{H}(\mu), \tag{3.4}$$

called the *variational Gibbs free energy* [25]. GFE_Ψ 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 Φ's *exponential family* and is a central tool in the analysis and development of many algorithms for graphical models [36].

PDGs can in fact capture the full exponential family of a factor graph, but only by allowing values of α other than $\mathbf{1}$. In this case, the only definition that requires alteration is IDef, which now depends on the *weighted multigraph* $(\mathcal{G}^m, \alpha^m)$, and is given by

$$\mathit{IDef}_{m}(\mu) := \sum_{X \xrightarrow{a} Y \in \mathcal{A}} \alpha_L \, \mathrm{H}_\mu(Y \mid X) - \mathrm{H}(\mu). \tag{3.5}$$

Thus, the conditional entropy $\mathrm{H}_\mu(Y \mid X)$ associated with the edge $X \xrightarrow{a} Y$ is multiplied by the weight α_L of that edge.

One key benefit of using α is that we can capture arbitrary WFGs, not just ones with a constant weight vector. All we have to do is to ensure that in our translation from factor graphs to PDGs, the ratio α_L / β_L is a constant. (Of course, if we allow arbitrary weights, we cannot hope to do this if $\alpha_L = 1$ for all edges L.) We therefore define a family of translations, parameterized by the ratio of α_L to β_L.

Definition 3.4.4 (WFG to PDG)**.** Given a WFG $\Psi = (\Phi, \theta)$, and postive number k, we define the corresponding PDG $\boldsymbol{\mathit{pdg}}(\Psi, k) = (\boldsymbol{\mathit{updg}}(\Phi), \alpha_\theta, \beta_\theta)$ by taking $\beta_J = k\theta_J$ and $\alpha_J = \theta_J$ for the edge $\mathbb{1} \to X_J$, and taking $\beta_L = k$ and $\alpha_L = 1$ for the projections $X_J \twoheadrightarrow X_j$. □

We now extend Definitions 3.4.2 and 3.4.3 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 have one. Here we throw out α and keep β, though in its role here as a left inverse of Definition 3.4.4, either choice would suffice.

Definition 3.4.5 (PDG to WFG)**.** Given a (weighted) PDG $\mathit{m} = (\mathit{n}, \beta)$, we take its corresponding WFG to be $\Psi_{\mathit{m}} := (\Phi_{\mathit{n}}, \beta)$; that is, $\theta_L := \beta_L$ for all edges L. □

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

(unproven!)

Theorem 3.4.4. *For all WFGs $\Psi = (\Phi, \theta)$ and all $\gamma > 0$, we have that $GFE_\Psi = {}^1\!/_{\gamma}[\![\mathit{m}_{\Psi,\gamma}]\!]_\gamma + C$ for some constant C, so $\Pr_\Psi$ is the unique element of $[\![\mathit{m}_{\Psi,\gamma}]\!]^*_\gamma$.*

In particular, for $k{=}1$, so that θ is used for both the functions α and β of the resulting PDG, Theorem 3.4.4 strictly generalizes Theorem 3.4.2.

Corollary 3.4.4.1. *For all weighted factor graphs (Φ, θ), we have that $\Pr_{(\Phi,\theta)} = [\![(\boldsymbol{\mathit{updg}}(\Phi), \theta, \theta)]\!]^*_1$*

Conversely, as long as the ratio of α_L to β_L is constant, the reverse translation also preserves semantics.

[link to proof]

Theorem 3.4.5. *For all unweighted PDGs $\mathcal{n}$ and non-negative vectors $\mathbf{v}$ over $\mathcal{A}^{\mathcal{n}}$, and all $\gamma > 0$, we have that $[\![(\mathcal{n}, \mathbf{v}, \gamma\mathbf{v})]\!]_\gamma = \gamma\, \mathit{GFE}_{(\Phi_{\mathcal{n}}, \mathbf{v})}$; consequently, $[\![(\mathcal{n}, \mathbf{v}, \gamma\mathbf{v})]\!]^*_\gamma = \{\Pr_{(\Phi_{\mathcal{n}}, \mathbf{v})}\}$.*

The key step in proving Theorems 3.4.4 and 3.4.5 (and in the proofs of a number of other results) involves rewriting $[\![\mathcal{m}]\!]_\gamma$ as follows:

link to proof

Proposition 3.4.6. *Letting $x^{\mathbf{w}}$ and $y^{\mathbf{w}}$ denote the values of X and Y, respectively, in $\mathbf{w} \in \mathcal{V}(\mathcal{m})$, we have*

$$[\![\mathcal{m}]\!](\mu) = \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu} \Bigg\{ \sum_{X \xrightarrow{a} Y} \Bigg[\overbrace{\beta_L \log \frac{1}{\mathbb{P}_a(y^{\mathbf{w}}|x^{\mathbf{w}})}}^{\textit{log likelihood / cross entropy}} + \underbrace{(\alpha_L\gamma - \beta_L) \log \frac{1}{\mu(y^{\mathbf{w}}|x^{\mathbf{w}})}}_{\textit{local regularization (if } \beta_L > \alpha_L\gamma\textit{)}} \Bigg] - \underbrace{\gamma \log \frac{1}{\mu(\mathbf{w})}}_{\textit{global regularization}} \Bigg\}. \tag{3.6}$$

For a fixed γ, the first and last terms of (3.6) are equal to a scaled version of the free energy, $\gamma\, \mathit{GFE}_\Phi$, if we set $\phi_J := \mathbb{P}_a$ and $\theta_J := {}^{\beta_L}\!/\!{}_{\gamma}$. If, in addition, $\beta_L = \alpha_L\gamma$ for all edges L, then the local regularization term disappears, giving us the desired correspondence.

Equation (3.6) also makes it clear that taking $\beta_L = \alpha_L\gamma$ for all edges L is essentially necessary to get Theorems 3.4.2 and 3.4.3. Of course, fixed γ precludes taking the limit as γ goes to 0, so Proposition 3.3.4 does apply. This is reflected in some strange behavior in factor graphs trying to capture the same phenomena as PDGs, as the following example shows.

Example 5. Consider the PDG $\mathcal{m}$ containing just X and 1, and two edges $p, q : 1 \to X$. (Recall that 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!) Consider the simplest situation, where p and q are both associated with the same distribution on X; further suppose that the agent is certain about the distribution, so $\beta_p = \beta_q = 1$. For definiteness, suppose that $\mathcal{V}(X) = \{x_1, x_2\}$, and that the distribution associated with both edges is $\mu_{.7}$, which ascribes probability .7 to x_1. Then, as we would hope $[\![\mathcal{m}]\!]^* = \{\mu_{.7}\}$; after all, both sources agree on the information. However, it can be shown that $\Pr_{\Psi_{\mathcal{m}}} = \mu_{.85}$, so $[\![\mathcal{m}]\!]^*_1 = \{\mu_{.85}\}$. □

Although both θ and β are measures of confidence, the way that the Gibbs free energy varies with θ is quite different from the way that the score of a PDG varies with β. The scoring function that we use for PDGs can be viewed as extending $\mathit{GFE}_{\Phi,\theta}$ by including the local regularization term. As γ 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 probabilistic information. They have a number of advantages over other probablisitic graphical models.

- They allow us to capture inconsistency, including conflicting information from multiple sources with varying degrees of reliability.
- They 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 (edges) and features (nodes) without affecting previously-received information.

- They allow for a clean separation between quantitiatve information (the cpds and weights β) and more qualitative information contained by the graph structure (and the weights α); this is captured by the terms Inc and IDef in our scoring function.
- PDGs have (several) natural semantics; one of them allows us to pick out a unique distribution. Using this distrbution, 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.

We have only scratched the surface of what can be done with PDGs here. 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. Due to space limitations, we just briefly give some intuitions and examples here.

Suppose that we want to compute the probability of Y given X in a PDG $\mathcal{M}$. For a cpd $p(Y|X)$, let $\mathcal{M}^{+p}$ be the PDG obtained by associating p with a new edge in $\mathcal{M}$ from X to Y, with $\alpha_p = 0$. We judge the quality of a candidate answer p by the best possible score that $\mathcal{M}^{+p}$ gives to any distribution (which we call the *degree of inconsistency* of $\mathcal{M}^{+p}$). It can be shown that the deegree of inconsistency is minimized by $[\![\mathcal{M}]\!]^*(Y \mid X)$. Since the degree of inconsistency of $\mathcal{M}^{+p}$ is smooth and strongly convex as a function of p, we can compute its optimum values by standard gradient methods. This approach is inefficient as written (since it involves computing the full joint distribution $[\![\mathcal{M}^{+p}]\!]^*$), but we believe that standard approximation techniques will allow us to draw inferences efficiently.

To take another 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 edge from $\mathbb{1}$ to Y in $\mathcal{M}$, annoted with the cpd that gives probability 1 to y, to get the (possibly inconsistent) PDG $\mathcal{M}^{+(Y=y)}$. The distribution $[\![\mathcal{M}^{+(Y=y)}]\!]^*$ turns out to be the result of conditioning $[\![\mathcal{M}]\!]^*$ on $Y = y$. This account of conditioning generalizes without modification to give Jeffrey's Rule [16], a more general approach to belief updating.

Issues of updating and inconsistency also arise in variational inference. A variational autoencoder [17], for instance, is essentially three cpds: a prior $p(Z)$, a decoder $p(X \mid Z)$, and an encoder $q(Z \mid X)$. 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. We hope to report further on the deep connection between inference, updating, and the resolution of inconsistency in PDGs in future work.

Ethics Statement

Because PDGs are a recent theoretical development, there is a lot of guesswork in evaluating the impact. Here are two views of opposite polarity.

Positive Impacts

One can imagine many applications of enabling simple and coherent aggregation of (possibly inconsistent) information. In particular we can imagine using PDGs to build and interpret a communal and global database of statistical models, in a way that may not only enable more accurate predictions, but also highlights conflicts between information.

This could have many benefits. Suppose, for instance, that two researchers train models, but use datasets with different racial makeups. Rather than trying to get an uninterpretable model to "get it right" the first time, we could simply highlight any such clashes and flag them for review.

Rather than trying to ensure fairness by design, which is both tricky and costly, we envision an alternative: simply aggregate (conflicting) statistically optimal results, and allow existing social structure to resolve conflicts, rather than sending researchers to fiddle with loss functions until they look fair.

Negative Impacts

We can also imagine less rosy outcomes. To the extent that PDGs can model and reason with inconsistency, if we adopt the attitude that a PDG need not wait until it is consistent to be used, it is not hard to imagine a world where a PDG gives biased and poorly-thought out conclusions. It is clear that PDGs need a great deal more vetting before they can be used for such important purposes as aggregating the world's statistical knowledge.

PDGs are powerful statistical models, but are by necessity semantically more complicated than many existing methods. This will likely restrict their accessibility. To mitigate this, we commit to making sure our work is widely accessible to researchers of different backgrounds.

APPENDICES FOR CHAPTER 3

3.A Proofs

For brevity, we use the standard notation and write $\mu(x, y)$ instead of $\mu(X = x, Y = y)$, $\mu(x \mid y)$ instead of $\mu(X = x \mid Y = y)$, and so forth.

3.A.1 Properties of Scoring Semantics

In this section, we prove the properties of scoring functions that we mentioned in the main text, Propositions 3.3.1, 3.3.2, and 3.3.4. We repeat the statements for the reader's convenience.

Proposition 3.3.1. *$\{\!\{\mathcal{m}\}\!\} = \{\mu : [\![\mathcal{m}]\!]_0(\mu) = 0\}$ for all $\mathcal{m}$.*

Proof. By taking $\gamma = 0$, the score is just Inc. By definition, a distribution $\mu \in \{\!\{\mathcal{m}\}\!\}$ satisfies all the constraints, so $\mu(Y = \cdot \mid X = x) = \mathbb{P}_a(x)$ for all edges $X \to Y \in \mathcal{A}^{\mathcal{m}}$ and x with $\mu(X = x) > 0$. By Gibbs inequality [24], $\boldsymbol{D}(\mu(Y|x) \parallel \mathbb{P}_a(x)) = 0$. Since this is true for all edges, we must have $\mathit{Inc}_{\mathcal{m}}(\mu) = 0$. Conversely, if $\mu \notin \{\!\{\mathcal{m}\}\!\}$, then it fails to marginalize to the cpd $\mathbb{P}_a$ on some edge L, and so again by Gibbs inequality, $\boldsymbol{D}(\mu(Y|x) \parallel \mathbb{P}_a(x)) > 0$. As relative entropy is non-negative, the sum of these terms over all edges must be positive as well, and so $\mathit{Inc}_{\mathcal{m}}(\mu) \neq 0$. □

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

Lemma 3.A.1. *$\mathit{Inc}_{\boldsymbol{m}}(\mu)$ is a convex function of μ.*

Proof. It is well known that $\boldsymbol{D}$ is convex [5, Theorem 2.7.2], in the sense that

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

Given an edge $\ell \in \mathcal{A}$ from A to B and $a \in \mathcal{V}(A)$, and setting $q_1 = q_2 = \mathbb{P}_\ell(a)$, we get that

$$\boldsymbol{D}(\mathbb{P}_\ell(a) \; || \; \lambda p_1 + (1-\lambda)p_2) \leq \lambda \boldsymbol{D}(\mathbb{P}_\ell(a) \; || \; p_1) + (1-\lambda)\boldsymbol{D}(\mathbb{P}_\ell(a) \; || \; p_2).$$

Since this is true for every a and edge, we can take a weighted sum of these inequalities for each a weighted by $p(A = a)$; thus,

$$\mathop{\mathbb{E}}_{a \sim p_A} \boldsymbol{D}(\mathbb{P}_\ell(a) \parallel \lambda p_1 + (1-\lambda)p_2) \leq \mathop{\mathbb{E}}_{a \sim p_A} \lambda \boldsymbol{D}(\mathbb{P}_\ell(a) \parallel p_1) + (1-\lambda)\boldsymbol{D}(\mathbb{P}_\ell(a) \parallel p_2).$$

Taking a sum over all edges, we get that

$$\sum_{(A,B)\in\mathcal{A}} \mathop{\mathbb{E}}_{a \sim p_A} \boldsymbol{D}(\mathbb{P}_\ell(a) \parallel \lambda p_1 + (1-\lambda)p_2) \leq \sum_{(A,B)\in\mathcal{A}} \mathop{\mathbb{E}}_{a \sim p_A} \lambda \boldsymbol{D}(\mathbb{P}_\ell(a) \parallel p_1) + (1-\lambda)\boldsymbol{D}(\mathbb{P}_\ell(a) \parallel p_2).$$

It follows that

$$\mathit{Inc}_{\boldsymbol{m}}(\lambda p_1) + (1-\lambda)p_2) \leq \lambda \mathit{Inc}_{\boldsymbol{m}}(p_1) + (1-\lambda)\mathit{Inc}_{\boldsymbol{m}}(p_2).$$

Therefore, $\mathit{Inc}_{\boldsymbol{m}}(\mu)$ is a convex function of μ. □

The next proposition gives us a useful representation of $[\![M]\!]_\gamma$.

Proposition 3.4.6. *Letting $x^{\mathbf{w}}$ and $y^{\mathbf{w}}$ denote the values of X and Y, respectively, in $\mathbf{w} \in \mathcal{V}(\boldsymbol{m})$, we have*

$$[\![\boldsymbol{m}]\!](\mu) = \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu} \Bigg\{ \sum_{X \xrightarrow{a} Y} \Bigg[\overbrace{\beta_L \log \frac{1}{\mathbb{P}_a(y^{\mathbf{w}}|x^{\mathbf{w}})}}^{\text{log likelihood / cross entropy}} + \underbrace{(\alpha_L\gamma - \beta_L) \log \frac{1}{\mu(y^{\mathbf{w}}|x^{\mathbf{w}})}}_{\text{local regularization (if } \beta_L > \alpha_L\gamma)} \Bigg] - \underbrace{\gamma \log \frac{1}{\mu(\mathbf{w})}}_{\text{global regularization}} \Bigg\}. \tag{3.6}$$

Proof. We use the more general formulation of *IDef* given in Section 3.4.3, in which each arc a's conditional information is weighted by α_a.

$$
\begin{aligned}
[\![\mathcal{M}]\!]_\gamma(\mu) &:= \mathit{Inc}_{\mathcal{M}}(\mu) + \gamma \mathit{IDef}_{\mathcal{M}}(\mu) \\
&= \left[\sum_{X \xrightarrow{a} Y} \beta_L \mathop{\mathbb{E}}_{x \sim \mu_X} \boldsymbol{D}\Big(\mu(Y|X{=}x) \,\Big\|\, \mathbb{P}_a(x) \Big) \right] + \gamma \left[\sum_{X \xrightarrow{a} Y} \alpha_L \, \mathrm{H}_\mu(Y \mid X) \; - \mathrm{H}(\mu) \right] \\
&= \sum_{X \xrightarrow{a} Y} \mathop{\mathbb{E}}_{x \sim \mu_X} \Big[\beta_L \, \boldsymbol{D}\Big(\mu(Y \mid x) \,\Big\|\, \mathbb{P}_a(Y \mid x) \Big) + \gamma \, \alpha_L \, \mathrm{H}(Y \mid X{=}x) \Big] - \gamma \, \mathrm{H}(\mu) \\
&= \sum_{X \xrightarrow{a} Y} \mathop{\mathbb{E}}_{x \sim \mu_X} \left[\beta_L \left(\sum_{y \in \mathcal{V}(Y)} \mu(y \mid x) \log \frac{\mu(y \mid x)}{\mathbb{P}_a(y \mid x)} \right) + \alpha_L \gamma \left(\sum_{y \in \mathcal{V}(Y)} \mu(y \mid x) \log \frac{1}{\mu(y \mid x)} \right) \right] - \\
&= \sum_{X \xrightarrow{a} Y} \mathop{\mathbb{E}}_{x \sim \mu_X} \left[\sum_{y \in \mathcal{V}(Y)} \mu(y \mid x) \left(\beta_L \; \log \frac{\mu(y \mid x)}{\mathbb{P}_a(y \mid x)} + \alpha_L \gamma \; \log \frac{1}{\mu(y \mid x)} \right) \right] - \gamma \, \mathrm{H}(\mu) \\
&= \sum_{X \xrightarrow{a} Y} \mathop{\mathbb{E}}_{x \sim \mu_X} \left[\mathop{\mathbb{E}}_{y \sim \mu(Y|X=x)} \left(\beta_L \; \log \frac{\mu(y \mid x)}{\mathbb{P}_a(y \mid x)} + \alpha_L \gamma \; \log \frac{1}{\mu(y \mid x)} \right) \right] - \gamma \sum_{\mathbf{w} \in \mathcal{V}(\mathcal{M})} \mu(\mathbf{w}) \log \frac{1}{\mu(\mathbf{w}} \\
&= \sum_{X \xrightarrow{a} Y} \mathop{\mathbb{E}}_{x,y \sim \mu_{XY}} \left[\beta_L \; \log \frac{\mu(y \mid x)}{\mathbb{P}_a(y \mid x)} + \alpha_L \gamma \; \log \frac{1}{\mu(y \mid x)} \right] - \gamma \mathop{\mathbb{E}}_{\mathbf{w} \sim \mu} \left[\log \frac{1}{\mu(\mathbf{w})} \right] \\
&= \mathop{\mathbb{E}}_{\mathbf{w} \sim \mu} \left\{ \sum_{X \xrightarrow{L} Y} \left[\beta_L \log \frac{1}{\mathbb{P}_a(y \mid x)} - \beta_L \log \frac{1}{\mu(y \mid x)} + \alpha_L \gamma \log \frac{1}{\mu(y \mid x)} \right] \right\} - \gamma \mathop{\mathbb{E}}_{\mathbf{w} \sim \mu} \left[\log \frac{1}{\mu(\mathbf{w})} \right. \\
&= \mathop{\mathbb{E}}_{\mathbf{w} \sim \mu} \left\{ \sum_{X \xrightarrow{L} Y} \left[\beta_L \log \frac{1}{\mathbb{P}_a(y \mid x)} + (\alpha_L \gamma - \beta_L) \log \frac{1}{\mu(y \mid x)} \right] - \gamma \log \frac{1}{\mu(\mathbf{w})} \right\}.
\end{aligned}
$$

□

We can now prove Proposition 3.3.2.

Proposition 3.3.2. *If $\mathcal{M}$ is a PDG and $0 < \gamma \leq \min_L \beta_L^{\mathcal{M}}$, then $[\![\mathcal{M}]\!]_\gamma^*$ is a singleton.*

Proof. It suffices to show that $[\![\mathcal{M}]\!]_\gamma$ is a strictly convex function of μ, since every strictly convex function has a unique minimum. Note that

$$\begin{aligned}
[\![M]\!]_\gamma(\mu) &= \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu} \left\{ \sum_{X \xrightarrow{L} Y} \left[\beta_L \log \frac{1}{\mathbb{P}_a(y \mid x)} + (\alpha_L\gamma - \beta_L) \log \frac{1}{\mu(y \mid x)} \right] - \gamma \log \frac{1}{\mu(\mathbf{w})} \right\} \\
&= \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu} \left\{ \sum_{X \xrightarrow{L} Y} \left[\gamma\alpha_L \log \frac{1}{\mathbb{P}_a(y \mid x)} + (\beta_L - \alpha_L\gamma) \log \frac{1}{\mathbb{P}_a(y \mid x)} - (\beta_L - \alpha_L\gamma) \log \frac{1}{\mu(y \mid x)} \right] - \right. \\
&= \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu} \left\{ \sum_{X \xrightarrow{L} Y} \left[\gamma\alpha_L \log \frac{1}{\mathbb{P}_a(y \mid x)} + (\beta_L - \alpha_L\gamma) \log \frac{\mu(y \mid x)}{\mathbb{P}_a(y \mid x)} \right] - \gamma \log \frac{1}{\mu(\mathbf{w})} \right\} \\
&= \sum_{X \xrightarrow{L} Y} \left[\gamma\alpha_L \mathop{\mathbb{E}}_{x,y\sim\mu_{XY}} \left[\log \frac{1}{\mathbb{P}_a(y \mid x)} \right] + (\beta_L - \alpha_L\gamma) \mathop{\mathbb{E}}_{x\sim\mu_X} \boldsymbol{D}\Big(\mu(Y \mid x) \,\Big\|\, \mathbb{P}_a(x) \Big) \right] - \gamma \, \mathrm{H}(\mu).
\end{aligned}$$

The first term, $\mathbb{E}_{x,y\sim\mu_{XY}} [-\log \mathbb{P}_a(y \mid x)]$ is linear in μ, as $\mathbb{P}_a(y \mid x)$ does not depend on μ. As for the second term, it is well-known that KL divergence is convex, in the sense that

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

Therefore, for a distribution on Y, setting $p_1 = p_2 = \mathbb{P}_a(x)$, for all conditional marginals $\mu_1(Y \mid X = x)$ and $\mu_2(Y \mid X = x)$,

$$\boldsymbol{D}(\lambda\mu_1(Y \mid x) + (1-\lambda)\mu_2(Y \mid x) \parallel \mathbb{P}_a(x)) \leq \lambda\boldsymbol{D}(\mu_1(Y \mid x) \parallel \mathbb{P}_a(x)) + (1-\lambda)\boldsymbol{D}(\mu_2(Y \mid x) \parallel \mathbb{P}_a(x)).$$

So $\boldsymbol{D}(\mu(Y \mid x) \parallel \mathbb{P}_a(Y|x))$ is convex. As convex combinations of convex functions are convex, the second term, $\mathbb{E}_{x\sim\mu_X} \boldsymbol{D}(\mu(Y \mid x) \parallel \mathbb{P}_a(x))$, is convex. Finally, negative entropy 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 $\beta_L \geq \gamma$ for all edges $L \in \mathcal{A}^{\mathcal{m}}$, $[\![\mathcal{m}]\!]_\gamma$ is strictly convex. The result follows. □

We next prove Proposition 3.3.3. The first step is provided by the following lemma.

Lemma 3.A.2. $\lim\limits_{\gamma \to 0} [\![\mathcal{M}]\!]^*_\gamma \subseteq [\![\mathcal{M}]\!]^*_0$.

Proof. Since $\mathit{IDef}_{\mathcal{M}}$ is a finite weighted sum of entropies and conditional entropies over the variables $\mathcal{N}^{\mathcal{M}}$, which have finite support, it is bounded. Thus, there exist bounds k and K depending only on $\mathcal{N}^{\mathcal{M}}$ and $\mathcal{V}^{\mathcal{M}}$, such that $k \leq \mathit{IDef}_{\mathcal{M}}(\mu) \leq K$ for all μ. Since $[\![\mathcal{M}]\!]_\gamma = \mathit{Inc}_{\mathcal{M}} + \gamma \mathit{IDef}_{\mathcal{M}}$, it follows that, for all $\mu \in \mathcal{V}(\mathcal{M})$, we have

$$\mathit{Inc}_{\mathcal{M}}(\mu) + \gamma k \leq \; [\![\mathcal{M}]\!]_\gamma(\mu) \leq \; \mathit{Inc}_{\mathcal{M}}(\mu) + \gamma K.$$

For a fixed γ, since this inequality holds for all μ, and both Inc and IDef are bounded below, it must be the case that

$$\min_{\mu \in \Delta\mathcal{V}(\mathcal{M})} \Big[\mathit{Inc}_{\mathcal{M}}(\mu) + \gamma k \Big] \;\leq\; \min_{\mu \in \Delta\mathcal{V}(\mathcal{M})} [\![\mathcal{M}]\!]_\gamma(\mu) \;\leq\; \min_{\mu \in \Delta\mathcal{V}(\mathcal{M})} \Big[\mathit{Inc}_{\mathcal{M}}(\mu) + \gamma K \Big],$$

even though the distributions that minimize each expression will in general be different. Let $\mathit{Inc}(\mathcal{M}) = \min_\mu \mathit{Inc}_{\mathcal{M}}(\mu)$. Since $\Delta\mathcal{V}(\mathcal{M})$ is compact, the minimum of the middle term is achieved. Therefore, for $\mu_\gamma \in [\![\mathcal{M}]\!]^*_\gamma(\mu)$ that minimizes it, we have

$$\mathit{Inc}(\mathcal{M}) + \gamma k \leq [\![\mathcal{M}]\!]_\gamma(\mu_\gamma) \leq \mathit{Inc}(\mathcal{M}) + \gamma K$$

for all $\gamma \geq 0$. Now taking the limit as $\gamma \to 0$ from above, we get that $\mathit{Inc}(\mathcal{M}) = [\![\mathcal{M}]\!]_0(\mu^*)$. Thus, $\mu^* \in [\![\mathcal{M}]\!]^*_0$, as desired. □

We now apply Lemma 3.A.2 to show that the limit as $\gamma \to 0$ is unique, as stated in Proposition 3.3.3.

Proposition 3.3.3. *For all $\mathcal{M}$, $\lim_{\gamma \to 0} [\![\mathcal{M}]\!]^*_\gamma$ is a singleton.*

Proof. First we show that $\lim_{\gamma \to 0} [\![\mathcal{M}]\!]^*_\gamma$ cannot be empty. Let $(\gamma_n) = \gamma_1, \gamma_2, \ldots$ be a sequence of positive reals converging to zero. For all n, choose some

$\mu_n \in [\![\mathcal{m}]\!]^*_{\gamma_n}$. Because $\Delta\mathcal{V}(\mathcal{m})$ is a compact metric space, it is sequentially compact, and so, by the Bolzano–Weierstrass Theorem, the sequence (μ_n) has at least one accumulation point, say ν. By our definition of the limit, $\nu \in \lim_{\gamma\to 0}[\![\mathcal{m}]\!]^*_\gamma$, as witnessed by the sequence $(\gamma_n, \mu_n)_n$. It follows that $\lim_{\gamma\to 0}[\![\mathcal{m}]\!]^*_\gamma \neq \emptyset$.

Now, choose $\nu_1, \nu_2 \in \lim_{\gamma\to 0}[\![\mathcal{m}]\!]^*_\gamma$. Thus, there are subsequences (μ_i) and (μ_j) of (μ_n) converging to ν_1 and ν_2, respectively. By Lemma 3.A.2, $\nu_1, \nu_2 \in [\![\mathcal{m}]\!]^*_0$, so $\mathit{Inc}_{\mathcal{m}}(\nu_1) = \mathit{Inc}_{\mathcal{m}}(\nu_2)$. Because $(\mu_{j_n}) \to \nu_1$, $(\mu_{k_n}) \to \nu_2$, and $\mathit{IDef}_{\mathcal{m}}$ is continuous on $\Delta\mathcal{V}(\mathcal{m})$, we conclude that $(\mathit{IDef}_{\mathcal{m}}(\mu_i)) \to \mathit{IDef}_{\mathcal{m}}(\nu_1)$ and $(\mathit{IDef}_{\mathcal{m}}(\mu_j)) \to \mathit{IDef}_{\mathcal{m}}(\nu_2)$.

Suppose that $\mathit{IDef}_{\mathcal{m}}(\nu_1) \neq \mathit{IDef}_{\mathcal{m}}(\nu_2)$. Without loss of generality, suppose that $\mathit{IDef}_{\mathcal{m}}(\nu_1) > \mathit{IDef}_{\mathcal{m}}(\nu_2)$. Since $(\mathit{IDef}_{\mathcal{m}}(\mu_i)) \to \mathit{IDef}_{\mathcal{m}}(\nu_1)$, there exists some $i^* \in \mathbb{N}$ such that for all $i > i^*$, $\mathit{IDef}_{\mathcal{m}}(\mu_i) > \mathit{IDef}_{\mathcal{m}}(\nu_2)$. But then for all γ and $i > i^*$, we have

$$[\![\mathcal{m}]\!]_\gamma(\mu_i) = \mathit{Inc}(\mu_i) + \gamma\mathit{IDef}_{\mathcal{m}}(\mu_i) > \mathit{Inc}(\nu_2) + \gamma\mathit{IDef}_{\mathcal{m}}(\nu_2) = [\![\mathcal{m}]\!]_\gamma(\nu_2),$$

contradicting the assumption that μ_i minimizes $[\![\mathcal{m}]\!]_{\gamma_i}$. We thus conclude that we cannot have $\mathit{IDef}_{\mathcal{m}}(\nu_1) > \mathit{IDef}_{\mathcal{m}}(\nu_2)$. By the same argument, we also cannot have $\mathit{IDef}_{\mathcal{m}}(\nu_1) < \mathit{IDef}_{\mathcal{m}}(\nu_2)$, so $\mathit{IDef}_{\mathcal{m}}(\nu_1) = \mathit{IDef}_{\mathcal{m}}(\nu_2)$.

Now, suppose that ν_1 and ν_2 distinct. Since $[\![\mathcal{m}]\!]_\gamma$ is strictly convex for $\gamma > 0$, among the possible convex combinations of ν_1 and ν_2, the distribution $\nu_3 = \lambda\nu_1 + (1-\lambda)\nu_2$ that minimizes $[\![\mathcal{m}]\!]_\gamma$ must lie strictly between ν_1 and ν_2. Because Inc itself is convex and $\mathit{Inc}_{\mathcal{m}}(\nu_1) = \mathit{Inc}_{\mathcal{m}}(\nu_2) =: v$, we must have $\mathit{Inc}_{\mathcal{m}}(\nu_3) \leq v$. But since $\nu_1, \nu_2 \in [\![\mathcal{m}]\!]^*_0$ minimize Inc, we must have $\mathit{Inc}_{\mathcal{m}}(\nu_3) \geq v$. Thus, $\mathit{Inc}_{\mathcal{m}}(\nu_3) = v$. Now, because, for all $\gamma > 0$,

$$[\![\mathcal{m}]\!]_\gamma(\nu_3) = v + \gamma\mathit{IDef}_{\mathcal{m}}(\nu_3) < v + \gamma\mathit{IDef}_{\mathcal{m}}(\nu_1) = [\![\mathcal{m}]\!]_\gamma(\nu_1),$$

it must be the case that $\mathit{IDef}_{\mathcal{M}}(\nu_3) < \mathit{IDef}_{\mathcal{M}}(\nu_1)$.

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

$$\llbracket \mathcal{M} \rrbracket_\gamma(\mu_i) = \mathit{Inc}(\mu_i) + \gamma \mathit{IDef}_{\mathcal{M}}(\mu_i) > \mathit{Inc}(\nu_3) + \gamma \mathit{IDef}_{\mathcal{M}}(\nu_3) = \llbracket \mathcal{M} \rrbracket_\gamma(\nu_3),$$

again contradicting the assumption that μ_i minimizes $\llbracket \mathcal{M} \rrbracket_{\gamma_i}$. Thus, our supposition that ν_1 was distinct from ν_2 cannot hold, and so $\lim_{\gamma\to 0} \llbracket \mathcal{M} \rrbracket_\gamma^*$ must be a singleton, as desired. □

Finally, Proposition 3.3.4 is a simple corollary of Lemma 3.A.2 and Proposition 3.3.3, as we now show.

Proposition 3.3.4. *$\llbracket \mathcal{M} \rrbracket^* \in \llbracket \mathcal{M} \rrbracket_0^*$, so if $\mathcal{M}$ is consistent, then $\llbracket \mathcal{M} \rrbracket^* \in \{\!\{ \mathcal{M} \}\!\}$.*

Proof. By Proposition 3.3.3, $\lim_{\gamma\to 0} \llbracket \mathcal{M} \rrbracket_\gamma^*$ is a singleton. As in the body of the paper, we refer to its unique element by $\llbracket \mathcal{M} \rrbracket^*$ Lemma 3.A.2 therefore immediately gives us $\llbracket \mathcal{M} \rrbracket^* \in \llbracket \mathcal{M} \rrbracket_0^*$.

If $\mathcal{M}$ is consistent, then by Proposition 3.3.1, $\mathit{Inc}(\mathcal{M}) = 0$, so $\llbracket \mathcal{M} \rrbracket_0(\llbracket \mathcal{M} \rrbracket^*) = 0$, and thus $\llbracket \mathcal{M} \rrbracket^* \in \{\!\{ \mathcal{M} \}\!\}$. □

3.A.2 PDGs as Bayesian Networks

In this section, we prove Theorem 3.4.1. We start by recounting some standard results and notation, all of which can be found in a standard introduction to information theory (e.g., [24, Chapter 1]).

First, note that just as we introduced new variables to model joint dependence in PDGs, we can view a finite collection $\mathcal{X} = X_1, \ldots, X_n$ of random variables, where each X_i has the same sample space, as itself a random variable, taking the value $(x_1, \ldots, x_n)$ iff each X_i takes the value x_i. Doing so allows us to avoid cumbersome and ultimately irrelevant notation which treats sets of raomd variables differently, and requires lots of unnecessary braces, bold face, and uniqueness issues. Note the notational convention that the joint variable X, Y may be indicated by a comma.

Definition 3.A.1 (Conditional Independence)**.** If X, Y, and Z are random variables, and μ is a distribution over them, then X is *conditionally independent of Z given Y*, (according to μ), denoted '$X \perp\!\!\!\perp_\mu Z \mid Y$, iff for all $x, y, z \in \mathcal{V}(X, Y, Z)$, we have $\mu(x \mid y)\mu(z \mid y) = \mu(x, z \mid y)$. □

Fact 3.A.3 (Entropy Chain Rule)**.** *If X and Y are random variables, then the entropy of the joint variable (X, Y) can be written as $\mathrm{H}_\mu(X, Y) = \mathrm{H}_\mu(Y \mid X) + \mathrm{H}_\mu(X)$. It follows that if μ is a distribution over the n variables $X_1, \ldots, X_n$, then*

$$\mathrm{H}(\mu) = \sum_{i=1}^{n} \mathrm{H}_\mu(X_i \mid X_1, \ldots X_{i-1}).$$

Definition 3.A.2 (Conditional Mutual Information)**.** The *conditional mutual information* between two (sets of) random variables is defined as

$$\mathrm{I}_\mu(X; Y \mid Z) := \sum_{x,y,z \in \mathcal{V}(X,Y,Z)} \mu(x, y, z) \log \frac{\mu(z)\mu(x, y, z)}{\mu(x, z)\mu(y, z)}.$$

□

Fact 3.A.4 (Properties of Conditional Mutual Information)**.** *For random variables X, Y, and Z over a common set of outcomes, distributed according to a distribution μ, the following properties hold:*

1. ***(difference identity)*** $\mathrm{I}_\mu(X; Y \mid Z) = \mathrm{H}_\mu(X \mid Y) - \mathrm{H}_\mu(X \mid Y, Z)$*;*
2. ***(non-negativity)*** $\mathrm{I}_\mu(X; Y \mid Z) \geq 0$*;*
3. ***(relation to independence)*** $\mathrm{I}_\mu(X; Y \mid Z) = 0$ *iff* $X \perp\!\!\!\perp_\mu Z \mid Y$*.*

We now provide the formal details of the transformation of a BN into a PDG.

Definition 3.A.3 (Transformation of a BN to a PDG)**.** Recall that a (quantitative) Bayesian Network (G, f) consists of two parts: its qualitative graphical structure G, described by a dag, and its quantitative data f, an assignment of a cpd $p_i(X_i \mid \mathbf{Pa}(X_i))$ to each variable X_i. If $\mathcal{B}$ is a Bayesian network on random variables $X_1, \ldots, X_n$, we construct the corresponding PDG $\mathit{pdg}(\mathcal{B})$ as follows: we take $\mathcal{N} := \{X_1, \ldots, X_n\} \cup \{\mathbf{Pa}(X_1), \ldots, \mathbf{Pa}(X_n)\}$. That is, the variables of $\mathit{pdg}(\mathcal{B})$ consist of all the variables in $\mathcal{B}$ together with a variable corresponding to the parents of X_i. (This will be used to deal with the hyperedges.) The values $\mathcal{V}(X_i)$ for a random variable X_i are unchanged, (i.e., $\mathcal{V}^{\mathit{pdg}(\mathcal{B})}(\{X_i\}) := \mathcal{V}(X_i)$) and $\mathcal{V}^{\mathit{pdg}(\mathcal{B})}(\mathbf{Pa}(X_i)) := \prod_{Y \in \mathbf{Pa}(X_i)} \mathcal{V}(Y)$ (if $\mathbf{Pa}(X_i) = \emptyset$, so that X_i has no parents, then we then we identify $\mathbf{Pa}(X_i)$ with $\mathbb{1}$ and take $\mathcal{V}(\mathbf{Pa}(X_i)) = \{\star\}$). We take the set of edges $\mathcal{A}^{\mathit{pdg}(\mathcal{B})} := \{(\mathbf{Pa}(X_i), X_i) : i = 1, \ldots, n\} \cup \{(\mathbf{Pa}_i, Y) : Y \in \mathbf{Pa}(X_i)\}$ to be the set of edges to a variable X_i from its parents, together with an edge from from $\mathbf{Pa}(X_i)$ to each of the elements of $\mathbf{Pa}(X_i)$, for $i = 1, \ldots, n$. Finally, we set $\mathbb{P}^{\mathit{pdg}(\mathcal{B})}_{(\mathbf{Pa}(X_i), X_i)}$ to be the cpd associated with X_i in $\mathcal{B}$, and for each node $X_j \in \mathbf{Pa}(X_i)$, we define

$$\mathbb{P}^{\mathit{pdg}(\mathcal{B})}_{(\mathbf{Pa}(X_i), X_j)}(\ldots, x_j, \ldots) = \delta_{x_j};$$

that is, $\mathbb{P}^{\boldsymbol{pdg}(\mathcal{B},\beta)}_{(\mathbf{Pa}(X_i),X_j)}$ is the the cpd on X_j that, given a setting $(\ldots, x_j, \ldots)$ of $\mathbf{Pa}(X_i)$, yields the distribution that puts all mass on x_j. □

Let $\mathcal{X}$ be the variables of some BN $\mathcal{B}$, and $\mathcal{M} = (\mathcal{N}, \mathcal{A}, \mathbb{P}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ be the PDG $\boldsymbol{pdg}(\mathcal{B})$. Because the set $\mathcal{N}$ of variables in $\boldsymbol{pdg}(\mathcal{B},\beta)$ includes variables of the form $\mathbf{Pa}(X_i)$, it is a strict superset of $\mathcal{X} = \{X_1, \ldots, X_n\}$, the set of variables of $\mathcal{B}$. For the purposes of this theorem, we identify a distribution $\mu_{\mathcal{X}}$ over $\mathcal{X}$ with the unique distribution $\Pr_{\mathcal{B}}$ whose marginal on the variables in $\mathcal{X}$ is $\mu_{\mathcal{X}}$ such that if $X_j \in \mathbf{Pa}(X_i)$, then $\mu_{\mathcal{N}}(X_j = x'_j \mid \mathbf{Pa}(X_i) = (\ldots, x_j, \ldots)) = 1$ iff $x_j = x'_j$. In the argument below, we abuse notation, dropping the the subscripts $\mathcal{X}$ and $\mathcal{N}$ on a distribution μ.

Theorem 3.4.1. *If $\mathcal{B}$ is a Bayesian network and $\Pr_{\mathcal{B}}$ is the distribution it specifies, then for all $\gamma > 0$ and all vectors β such that $\beta_L > 0$ for all edges L, $[\![\boldsymbol{pdg}(\mathcal{B},\beta)]\!]^*_\gamma = \{\Pr_{\mathcal{B}}\}$, and thus $[\![\boldsymbol{pdg}(\mathcal{B},\beta)]\!]^* = \Pr_{\mathcal{B}}$.*

Proof. For the cpd $p(X_i \mid \mathbf{Pa}(X_i))$ associated to a node X_i in $\mathcal{B}$, we have that $\Pr_{\mathcal{B}}(X_i \mid \mathbf{Pa}(X_i)) = p(X_i \mid \mathbf{Pa}(X_i))$. For all nodes X_i in $\mathcal{B}$ and $X_j \in \mathbf{Pa}(X_i)$, by construcction, $\Pr_{\mathcal{B}}$, when viewed as a distribution on $\mathcal{N}$, is also with the cpd on the edge from $\mathbf{Pa}(X_i)$ to X_j. Thus, $\Pr_{\mathcal{B}}$ is consistent with all the cpds in $\boldsymbol{pdg}(\mathcal{B},\beta)$; so$\mathit{Inc}_{\boldsymbol{pdg}(\mathcal{B},\beta)}(\Pr_{\mathcal{B}}) = 0$.

We next want to show that $\mathit{IDef}_{\boldsymbol{pdg}(\mathcal{B},\beta)}(\mu) \geq 0$ for all distributions μ. To do this, we first need some definitions. Let ρ be a permutation of $1, \ldots, n$. 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 $X_j \in \mathbf{Pa}(X_i)$ 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 μ 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.A.3 that the cpd on the edge in $\boldsymbol{pdg}(\mathcal{B}, \beta)$ from $\mathbf{Pa}(X_i)$ to X_i is just the cpd associated with X_i in $\mathcal{B}$, while the cpd on the edge in $\boldsymbol{pdg}(\mathcal{B}, \beta)$ from $\mathbf{Pa}(X_i)$ to $X_j \in \mathbf{Pa}(X_i)$ consists only of deterministic distributions (i.e., ones that put probability 1 on one element), which all have entropy 0. Thus,

$$\sum_{X \xrightarrow{a} Y \in \mathcal{A}^{\boldsymbol{pdg}(\mathcal{B})}} \mathrm{H}_\mu(Y \mid X) = \sum_{i=1}^{n} \mathrm{H}_\mu(X_i \mid \mathbf{Pa}(X_i)). \tag{3.7}$$

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

$$\begin{aligned}
\mathit{IDef}_{\boldsymbol{pdg}(\mathcal{B},\beta)}(\mu) &= \left[\sum_{X \xrightarrow{a} Y \in \mathcal{A}^{\boldsymbol{pdg}(\mathcal{B})}} \mathrm{H}_\mu(Y \mid X) \right] - \mathrm{H}(\mu) \\
&= \sum_{i=1}^{n} \mathrm{H}_\mu(X_i \mid \mathbf{Pa}(X_i)) - \sum_{i=1}^{n} \mathrm{H}_\mu(X_i \mid \mathbf{X}_{\prec_\rho i}) \quad \text{[by Fact 3.A.3 and (3.7)]} \\
&= \sum_{i=1}^{n} \Big[\mathrm{H}_\mu(X_i \mid \mathbf{Pa}(X_i)) - \mathrm{H}_\mu(X_i \mid \mathbf{X}_{\prec_\rho i}) \Big] \\
&= \sum_{i=1}^{n} \mathrm{I}_\mu \left(X_i \,;\, \mathbf{X}_{\prec_\rho i} \setminus \mathbf{Pa}(X_i) \;\middle|\; \mathbf{Pa}(X_i) \right). \quad \text{[by Fact 3.A.4]}
\end{aligned}$$

Using Fact 3.A.4, it now follows that, for all distributions μ, $\mathit{IDef}_{\boldsymbol{pdg}(\mathcal{B})}(\mu) \geq 0$. Furthermore, for all μ and permutations ρ,

$$\mathit{IDef}_{\boldsymbol{pdg}(\mathcal{B})}(\mu) = 0 \quad \text{iff} \quad \forall i.\ X_i \perp\!\!\!\perp_\mu \mathbf{X}_{\prec_\rho i}. \tag{3.8}$$

Since the left-hand side of (3.8) is independent of ρ, it follows that X_i is independent of $\mathbf{X}_{\prec_\rho i}$ for some permutation ρ iff X_i is independent of $\mathbf{X}_{\prec_\rho i}$ for every permutation ρ. Since there is a permutation compatible with $\mathcal{B}$, we get that $\mathit{IDef}_{\boldsymbol{pdg}(\mathcal{B},\beta)}(\Pr_{\mathcal{B}}) = 0$. We have now shown that that $\mathit{IDef}_{\boldsymbol{pdg}(\mathcal{B},\beta)}$ and Inc are non-negative functions of μ, and both are zero at $\Pr_{0\mathcal{B}}$. Thus, for all $\gamma \geq 0$ and all vectors β, we have that $[\![\boldsymbol{pdg}(\mathcal{B},\beta)]\!]_\gamma(\Pr_\mathcal{B}) \leq [\![\boldsymbol{pdg}(\mathcal{B},\beta)]\!]_\gamma(\mu)$ for all distributions μ. We complete the proof by showing that if $\mu \neq \Pr_\mathcal{B}$, then $[\![\boldsymbol{pdg}(\mathcal{B},\beta)]\!]_\gamma(\mu) > 0$ for $\gamma > 0$.

So suppose that $\mu \neq \Pr_\mathcal{B}$. Then μ must also match each cpd of $\mathcal{B}$, for otherwise $\mathit{Inc}_{\boldsymbol{pdg}(\mathcal{B},\beta)}(\mu) > 0$, and we are done. Because $\Pr_\mathcal{B}$ is the *unique* distribution that matches the both the cpds and independencies of $\mathcal{B}$, μ 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 μ. There must be some permutation ρ 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 (3.8), $\mathit{IDef}_{\boldsymbol{pdg}(\mathcal{B})}(\mu) > 0$. This completes the proof. □

3.A.3 Factor Graph Proofs

Theorems 3.4.2 and 3.4.3 are immediate corolaries of their more general counterparts, Theorems 3.4.4 and 3.4.5, which we now prove.

Theorem 3.4.5. *For all unweighted PDGs $\mathcal{N}$ and non-negative vectors $\mathbf{v}$ over $\mathcal{A}^{\mathcal{N}}$, and all $\gamma > 0$, we have that $[\![(\mathcal{N},\mathbf{v},\gamma\mathbf{v})]\!]_\gamma = \gamma\, \mathit{GFE}_{(\Phi_{\mathcal{N}},\mathbf{v})}$; consequently, $[\![(\mathcal{N},\mathbf{v},\gamma\mathbf{v})]\!]^*_\gamma = \{\Pr_{(\Phi_{\mathcal{N}},\mathbf{v})}\}$.*

Proof. Let $\mathcal{M} := (\mathcal{N},\mathbf{v},\gamma\mathbf{v})$ be the PDG in question. Explicitly, $\alpha^{\mathcal{M}}_L = v_L$ and $\beta^{\mathcal{M}}_L = \gamma v_L$. By Proposition 3.4.6,

$$[\![\mathcal{M}]\!]_\gamma(\mu) = \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu}\left\{ \sum_{X \xrightarrow{L} Y} \left[\beta_L \log\frac{1}{\mathbb{P}_a(y \mid x)} + (\alpha_L\gamma - \beta_L)\log\frac{1}{\mu(y \mid x)}\right] - \gamma\log\frac{1}{\mu(\mathbf{w})}\right\}.$$

Let $\{\phi_L\}_{L\in\mathcal{A}} := \Phi_{\mathcal{M}}$ denote the factors of the factor graph associated with $\mathcal{M}$. Because we have $\alpha_L\gamma = \beta_L$, the middle term cancels, leaving us with

$$\begin{aligned}
[\![\mathcal{M}]\!]_\gamma(\mu) &= \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu}\left\{ \sum_{X\xrightarrow{L}Y}\left[\beta_L \log\frac{1}{\mathbb{P}_a(y\mid x)}\right] - \gamma\log\frac{1}{\mu(\mathbf{w})}\right\} \\
&= \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu}\left\{ \sum_{X\xrightarrow{L}Y}\left[\gamma v_L \log\frac{1}{\phi(x,y)}\right] - \gamma\log\frac{1}{\mu(\mathbf{w})}\right\} \qquad [\text{as } \beta_L = v_L\gamma] \\
&= \gamma \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu}\left\{ \sum_{X\xrightarrow{L}Y}\left[v_L \log\frac{1}{\phi(x,y)}\right] - \log\frac{1}{\mu(\mathbf{w})}\right\} \\
&= \gamma\, \mathit{GFE}_{(\Phi_{\mathcal{M}},\mathbf{v})}.
\end{aligned}$$

It immediately follows that the associated factor graph has $[\![\mathcal{M}]\!]^*_\gamma = \{\Pr_{\Phi(\mathcal{M})}\}$, because the free energy is clearly a constant plus the KL divergence from its associated probability distribution. □

Theorem 3.4.2. $\Pr_\Phi = [\![\mathit{updg}(\Phi)]\!]^*_1$ *for all factor graphs* Φ.[4]

Proof. In $\mathit{pdg}(\Psi,\gamma)$, there is an edge $1 \to X_J$ for every $J \in \mathcal{J}$, and also edges $X_J \twoheadrightarrow X_j$ for each $X_j \in X_J$. Because the latter edges are deterministic, a distribution μ that is not consistent with one of the edges, say $X_J \twoheadrightarrow X_j$, has $\mathit{Inc}_{\mathcal{M}}(\mu) = \infty$. 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 μ places positive probability on their co-occurance (i.e., $\mu(j^*,\mathbf{z}^*) > 0$), then we would have

$$\mathop{\mathbb{E}}_{\mathbf{z}\sim\mu_J} \boldsymbol{D}\Big(\mu(X_j \mid X_J = \mathbf{z}) \,\Big\|\, \mathbb{1}[X_j = \mathbf{z}_j]\Big) = \sum_{\substack{\mathbf{z}\in\mathcal{V}(X_J),\\ \iota\in\mathcal{V}(X_j)}} \mu(\mathbf{z},\iota)\log\frac{\mu(\iota\mid\mathbf{z})}{\mathbb{1}[\mathbf{z}_j=\iota]} \geq \mu(\mathbf{z}^*,j^*)\log\frac{\mu(j^*\mid\mathbf{z})}{\mathbb{1}[\mathbf{z}^*_j = j_*]} = \infty$$

Consequently, a distribution μ that does not satisfy the the projections has $[\![\mathcal{M}_{\Psi,\gamma}]\!]_\gamma(\mu) = \infty$ for every γ. Thus, a distribution that has a finite score must

[4] Recall that we identify the unweighted PDG $(\mathcal{G},\mathbf{p})$ with the weighted PDG $(\mathcal{G},\mathbb{P},\mathbf{1},\mathbf{1})$.

match the constraints, so we can identify such a distribution with its restriction to the original variables of Φ. Moreover, for all distributions μ with finite score and projections $X_J \twoheadrightarrow X_j$, the conditional entropy $\mathrm{H}(X_j \mid X_J) = -\mathbb{E}_\mu \log(\mu(x_j \mid x_J))$ and divergence from the constraints are both zero. Therefore the per-edge terms for both IDef_{m} and Inc_{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 3.4.4, we have $\boldsymbol{pdg}(\Psi, \gamma) = (\boldsymbol{updg}(\Phi), \theta, \gamma\theta)$, so by Proposition 3.4.6,

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

which differs from GFE_Ψ by the value $\sum_J \theta_J \log Z_J$, which is constant in μ.

□

CHAPTER 4

QUALITATIVE MECAHNISM INDEPENDENCE

RELATIVE ENTROPY SOUP

Oliver E. Richardson, Ph.D.
Cornell University 2024

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 $\mathcal{A}$, which is the qualitative structure of a probabilistic dependency graph (PDG). When $\mathcal{A}$ represents a qualitative Bayesian network, 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 information theory. Applying compatibility to cyclic structures helps to clarify a longstanding conceptual issue in information theory.

4.1 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) [28]. And in deep learning, a popular class of models called *normalizing flows* [35, 18] specify a distribution by composing a fixed distribution over some latent space, say a standard normal distribution, 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 [?]. 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 ($B \perp\!\!\!\perp C$).[1] Moreover, dependence can be written in terms of independence: Y is a function

[1] This well-known fact (Lemma 4.A.1) is formalized and proved in Section 5.C, where all proofs can be found.

of X if and only if Y is conditionally independent of itself given X (i.e., $X \twoheadrightarrow Y$ iff $Y \perp\!\!\!\perp Y \mid X$). 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) [29] and axioms for conditional independence [26], 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).

So the paper 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 is the structure of another type of probabilistic graphical model: a *probabilistic dependency graph (PDG)* [32, 33, 31].

Although the qualitative aspects of PDGs were characterized by Richardson and Halpern [32] using a scoring function, that scoring function does not seem to get at the qualitative aspects that we are most interested in here. In this work, we develop from first principles an alternate qualtiative semantics for directed hypergraphs. 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 4.2.1).

But QIM-compatibility can do much more than represent conditional independencies in acyclic networks. For one thing, it can encode arbitrary functional dependencies (Theorem 4.2.2); for another, it gives meaningful semantics to cyclic models. Indeed, compatibility lets us go well beyond capturing dependence and independence. The fact that Pearl [28] also views causal models as representing independent mechanisms suggests that there might be a connection between causality and comptability. In fact, there is. A *witness* that a distribution μ 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 μ 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 4.4). 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 10 and 11). It also gracefully handles incomplete fragments of a causal picture, as well as "over-determined" ones.

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 paper justifying it.

4.2 Qualitative Independent-Mechanism (QIM) Compatibility

In this section, we present the central definition of our paper: a way of making precise Pearl's notion of "independent mechanisms", used to motivate Bayesian Networks from a causal perspective. Pearl [30, 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* [7], that can have multiple sources.

Definition 4.2.1. A *directed hypergraph* (or simply a hypergraph, since all our hypergraphs will be directed) 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 $T = T_a$. 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 $\mathcal{N} := \bigcup_{a \in \mathcal{A}} S_a \cup T_a$). Thus, we often leave $\mathcal{N}$ implicit, referring to the hypergraph simply as $\mathcal{A}$. □

Following the graphical models literature, we are interested in hypergraphs whose nodes represent variables, so that each $X \in \mathcal{N}$ will ultimately be associated with a (for simplicity, finite) set $\mathcal{V}(X)$ of possible values. However, one should not think of $\mathcal{V}$ as part of the information carried by the hypergraph. 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 μ on a set of variables $\mathcal{X} \cong (\mathcal{N}, \mathcal{V})$, those variables must have possible values—but the *qualitative* properties of μ, 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 4.2.2 (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 μ (i.e., $\bar{\mu}(\mathcal{Y}) = \mu(\mathcal{Y})$),

(b) the variables $\mathcal{U}_{\mathcal{A}}$ are mutually independent (i.e., $\bar{\mu}(\mathcal{U}_{\mathcal{A}}) = \prod_{a \in \mathcal{A}} \bar{\mu}(U_a)$), and

(c) 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., $\forall a \in \mathcal{A}.\ \bar{\mu} \models (S_a, U_a) \twoheadrightarrow T_a$).

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

While Definition 4.2.2 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 6. $\mu(X, Y)$ is compatible with $\mathcal{A} = \{\emptyset \xrightarrow{1} \{X\}, \emptyset \xrightarrow{2} \{Y\}\}$ (depicted in PDG notation as $\rightarrow\boxed{X}\ \ \boxed{Y}\leftarrow$) 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. □

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* [22], *factoring* as a product of probability tables, or *d-separation* [8]). To state the general result (Theorem 4.2.1), we must first clarify how the graphs of standard graphical and causal models give rise to directed hypergraph s.

Suppose that $G = (V, E)$ is a graph, whose edges may be directed or undirected. Given a vertex $u \in V$, write $\mathbf{Pa}_G(u) := \{v : (v, u) \in E\}$ 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 $\mathcal{A}_G := \left\{ \mathbf{Pa}_G(u) \xrightarrow{u} \{u\} \right\}_{u \in V}$ be the hypergraph *corresponding* to the graph G.

link to proof

Theorem 4.2.1. *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 μ satisfies $\mathcal{I}(G)$.*

Theorem 4.2.1 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 [1]. Moreover, QIM-compatibility can capture properties other than independence. As the following example shows, it can capture determinism.

Example 7. If $\mathcal{A} = \{\xrightarrow{1} X,\ \xrightarrow{2} X\}$ consists of just two hyperarcs pointing to a single variable X, then a distribution $\mu(X)$ is QIM-compatible with $\mathcal{A}$ iff μ places all mass on a single value $x \in \mathcal{V}(X)$. □

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., $Y \to X$ and $Z \to X$) represent a single mechanism by which X is jointly determined (by Y and Z), rather than two distinct mechanisms. A central thrust of Richardson and Halpern's 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 $\mathcal{A} \sqcup^{(+n)}_{X \to Y}$ denote the hypergraph that results from augmenting $\mathcal{A}$ with n additional (distinct) hyperarcs from X to Y.

link to proof

Theorem 4.2.2. *(a)* $\mu \models X \twoheadrightarrow Y \land \Diamond\mathcal{A}$ *if and only if* $\forall n \geq 0.\ \mu \models \Diamond\mathcal{A} \sqcup^{(+n)}_{X\to Y}$.

(b) if $\mathcal{A} = \mathcal{A}_G$ *for a dag* G*, then* $\mu \models X \twoheadrightarrow Y \land \Diamond\mathcal{A}$ *if and only if* $\mu \models \Diamond\mathcal{A} \sqcup^{(+1)}_{X\to Y}$.

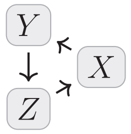
(c) if $\exists a \in \mathcal{A}$ *such that* $S_a = \emptyset$ *and* $X \in T_a$*, then* $\mu \models X \twoheadrightarrow Y \land \Diamond\mathcal{A}$ *iff* $\mu \models \Diamond\mathcal{A} \sqcup^{(+2)}_{X\to Y}$.

Based on the intuition given after Example 7, 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 4.E for counterexamples.

Finally, as alluded to above, QIM-compatibility gives meaning to cyclic structures, a topic that we will revisit often in Sections 4.3 and 4.4. We start with some simple examples.

Example 8. Every $\mu(X, Y)$ is compatible with $\boxed{X}\rightleftarrows\boxed{Y}$, because every distribution is compatible with $\to\boxed{X}\to\boxed{Y}$, and a mechanism with no inputs is a special case of one that can depend on Y. □

The logic above is an instance of an important reasoning principle, which we develop in Section 4.C. Although the 2-cycle in Example 8 is straightforward, generalizing it even slightly to a 3-cycle raises a not-so-straightforward question, whose answer will turn out to have surprisingly broad implications.



Example 9. 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 $\to X \to Y \to Z$. 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 4.4. △

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 $\{A{\to}B, B{\to}A\}$ 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 [20]. This may appear to be a flaw in Definition 4.2.2, but it is unavoidable (see Section 4.C) if we wish to also capture causality, as we do in the next section.

4.3 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 4.3.1). Moreover, such causal models and QIM-compatibility witnesses are themselves closely related (Section 4.3.2). In this section, we establish a causal grounding for QIM-compatibility. To do so, we must first review some standard definitions.

Definition 4.3.1 (Pearl [30]). 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* $f_Y : \mathcal{V}(\mathcal{U} \cup \mathcal{V} - Y) \to \mathcal{V}(Y)$ that determines its value as a function of the other variables.

$\square$

In a SEM M, a variable $X \in \mathcal{V}$ *does not depend* on $Y \in \mathcal{V} \cup \mathcal{U}$ if $f_X(\ldots, y, \ldots) = f_X(\ldots, y', \ldots)$ for all $y, y' \in \mathcal{V}(Y)$. Let the parents $\mathbf{Pa}_M(X)$ of X be the set of variables on which X depends. M is *acyclic* iff $\mathbf{Pa}_M(X) \cap \mathcal{V} = \mathbf{Pa}_G(X)$ for some dag G with vertices $\mathcal{V}$. In an acyclic SEM, it is easy to see that a setting of the exogneous variables determines the values of the endogenous variables (symbolically: $M \models \mathcal{U} \twoheadrightarrow \mathcal{V}$). A *probabilistic SEM* (PSEM) $\mathcal{M} = (M, P)$ is a SEM, together with a probability P over the exogenous variables. When $\mathcal{M} \models \mathcal{U} \twoheadrightarrow \mathcal{V}$ (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 cylic PSEM, however, may induce more than one such distribution, or none at all. In general, a PSEM $\mathcal{M}$ induces a (possiby 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 $\mathcal{M} = (M, P)$, define $\{\!\!\{\mathcal{M}\}\!\!\} :=$

$$\left\{ \nu \in \Delta\mathcal{V}(\mathcal{V} \cup \mathcal{U}) \;\middle|\; \begin{array}{l} \forall Y \in \mathcal{V}.\ \nu\Big(f_Y(\mathcal{U}, \mathcal{V}-Y) = Y\Big) = 1, \\ \nu(\mathcal{U}) = P(\mathcal{U}) \end{array} \right\}$$

and define $\{\!\!\{M\}\!\!\}$ 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. Given a PSEM $\mathcal{M}$, let $\{\!\!\{\mathcal{M}\}\!\!\}$ consist of all joint distributions $\nu(\mathcal{U}, \mathcal{V})$ that satisfy the two constraints above (or just the first of them, in the case of a non-probabilistic SEM). $\{\!\!\{\mathcal{M}\}\!\!\}$ can be thought of as the set of distributions compatible wth $\mathcal{M}$. It ; this set captures the behavior of $\mathcal{M}$ in the absence of interventions. 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 μ.

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 φ 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 φ 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 $\mathrm{s} \mapsto \mathbf{y}[Y]$, 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 formualas by structural induction in the usual way.[2] 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 [10]. 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 ϵ_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 are clearly a special class of PSEMs but any PSEM can be converted to an equivalent

[2] $M \models \varphi_1 \land \varphi_2$ iff $M \models \varphi_1$ and $M \models \varphi_2$; $M \models \neg\varphi$ iff $M \not\models \varphi$.

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 PSEMs. In fact, *qualitatively*, randomized PSEMs are more expressive: they can encode independence. It should come as no surprise that randomized PSEMs and QIM-compatibility are related.

4.3.1 The Equivalence Between QIM-Compatibility and Randomized PSEMs

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 4.3.1). We then give a natural generalization of a causal model that exactly captures QIM-compatibility with an arbitrary hypergraph (Proposition 4.3.2). 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 μ.

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 $\mathbf{Pa}_{\mathcal{M}}(Y) \subseteq S_Y \cup \{U_Y\}$ for all $Y \in \mathcal{V}$. We start by presenting the result in the case where $\mathcal{A}$ corresponds to a directed graph.

link to proof

Proposition 4.3.1. *Given a graph G and a distribution μ, $\mu \models \Diamond\mathcal{A}_G$ iff there exists a fully randomized PSEM of dependency structure $\mathcal{A}_G$ from which μ can arise.*

Proposition 4.3.1 shows that, for those hypergraphs induced by graphs, QIM-compatibility means arising from a fully randomized PSEM of the appropriate dependency structure. Theorem 4.2.1 makes precise a phenomenon that seems to be almost universally impilictly 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 4.3.2. 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 $\mathcal{F} = \{f_a : \mathcal{V}(S_a) \times \mathcal{V}(U_a) \to \mathcal{V}(T_a)\}_{a\in\mathcal{A}}$, 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. □

link to proof

Proposition 4.3.2. *$\mu \models \Diamond\mathcal{A}$ iff there exists a generalized randomized PSEM with structure $\mathcal{A}$ from which μ can arise.*

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* [3], a discussion we defer to future work.

4.3.2 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 4.F). 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 4.3.3. 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 $\bar{\mu} \models (U_{a_X}, S_{a_X}) \twoheadrightarrow T_{a_X}$ (this is just property (c) in Definition 4.2.2), $X \in T_{a_X}$ 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 $\mathrm{s} \in \mathcal{V}(S_{a_X})$ for which $\bar{\mu}(U_{a_X}{=}u, S_{a_X}{=}\mathrm{s}) > 0$, there is a unique $t \in \mathcal{V}(T_{a_X})$ such that $\bar{\mu}(u, \mathrm{s}, t) > 0$. In this case, set $f_X(u, \mathrm{s}, \ldots) := t[X]$. If $\bar{\mu}(U_{a_X}{=}u, S_{a_X}{=}\mathrm{s}) = 0$, $f_X(u, \mathrm{s}, \ldots)$ can be an arbitrary function of u and s. Let $\mathrm{PSEMs}_{\mathcal{A}}(\bar{\mu})$ denote the set of PSEMs that can result. □

It's clear from Construction 4.3.3 that $\mathrm{PSEMs}_{\mathcal{A}}(\bar{\mu})$ 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 μ 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 acylic 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 ω 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 ω 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.

link to proof

Proposition 4.3.3. *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 $\mathcal{M} \in \mathrm{PSEMs}_{\mathcal{A}}(\bar{\mu})$.*

Equivalently, this means that $\mathrm{PSEMs}_{\mathcal{A}}(\bar{\mu})$, the possible outputs of Construction 4.3.3, 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 $\mathcal{M} \in \mathrm{PSEMs}_{\mathcal{A}}(\bar{\mu})$ 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 whch 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

$$\mathrm{do}_M(\mathbf{X}{=}\mathbf{x}) := \bigcap_{X \in \mathbf{X}} \bigcap_{\mathbf{s} \in \mathcal{V}(\mathbf{Pa}(X))} f_X(U_X, \mathbf{s}) = \mathbf{x}[X], \quad \text{where } \mathbf{x}[X] \text{ is the value of } X \text{ in } \mathbf{x}. \tag{4.1}$$

This is intuitively the event in which the randomness is such that $\mathbf{X} = \mathbf{x}$ regardless of the values of the parent variables. [3]As we now show, conditioning on

[3] This is essentially the event in which, for each $X \in \mathbf{X}$, the *response variable* $\hat{U}_X := \lambda\mathbf{s}. f_X(\mathbf{s}, U_X)$, whose possible values $\mathcal{V}(\hat{U}_X)$ are functions from $\mathcal{V}(\mathbf{Pa}_M(X))$ to $\mathcal{V}(X)$ [34, 2], takes on the constant function $\lambda\mathbf{p}.\, x$.

$\mathrm{do}_M(\mathbf{X}{=}\mathbf{x})$ has the effect of intervention.

link to proof

Theorem 4.3.4. *Suppose that $\bar{\mu}$ is a witness to $\mu \models \Diamond\mathcal{A}$, $\mathcal{M} \in \mathrm{PSEMs}_{\mathcal{A}}(\bar{\mu})$, $\mathbf{X} \subseteq \mathcal{X}$ and $\mathbf{x} \in \mathcal{V}(\mathbf{X})$. If $\bar{\mu}(\mathrm{do}_{\mathcal{M}}(\mathbf{X}{=}\mathbf{x})) > 0$, then:*

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

(b) for all events $\varphi \subseteq \mathcal{V}(\mathcal{X})$, $\Pr_{\mathcal{M}}\left([\mathbf{X}{\leftarrow}\mathbf{x}]\varphi\right) \leq \bar{\mu}\left(\varphi \,\middle|\, \mathrm{do}_{\mathcal{M}}(\mathbf{X}{=}\mathbf{x})\right) \leq \Pr_{\mathcal{M}}\left(\langle\mathbf{X}{\leftarrow}\mathbf{x}\rangle\varphi\right)$
and all three are equal when $\mathcal{M} \models \mathcal{U} \twoheadrightarrow \mathcal{X}$ (such as when $\mathcal{M}$ is acyclic).

Theorem 4.3.4 shows that the relationship between witnesses and causal models extends to interventions. Even when $\mathrm{do}_M(\mathbf{X}{=}\mathbf{x})$ has probability zero, it is always possible to find a nearly equivalent setting where the bounds of the theorem apply.[4] 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 $\mathrm{do}(\,\cdot\,)$ notation actually corresponds to an event [13]). We emphasize that the conditioning (on $\mathrm{do}_M(\mathbf{X}{=}\mathbf{x})$) is on the randomness U_X and not X itself; intervening on $\mathbf{X}{=}\mathbf{x}$ is indeed fundamentally different from conditioning on $\mathbf{X}{=}\mathbf{x}$.

4.4 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 causal-

[4] 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 ϵ, and a distribution $\bar{\mu}'$ that is ϵ-close to $\bar{\mu}$, such that $\bar{\mu}'(\mathrm{do}_{\mathcal{M}'}(\mathbf{X}{=}\mathbf{x})) > 0$. As a result, Theorem 4.3.4 places bounds on the conditional probabilities that are possible limits of sequences of distributions $(\nu_k)_{k\geq 0}$ where $\nu_k(\mathrm{do}_M(\mathbf{X}{=}\mathbf{x})) > 0$, i.e., the possible outcomes of conditioning a *non-standard* probability measure [11] on this probability-zero event.

ity and probability. It makes explicit the distributional consequences of BN structure. Yet, despite substantial interest [1], generalizing the BN case to more complex (e.g., cyclic) dependency structures remains largely an open problem. In Section 4.4.1, 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 4.4.2, 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 fucntion IDef, which we use to show that our information-theoretic constraints degrade gracefully on "near-compatible" distributions.

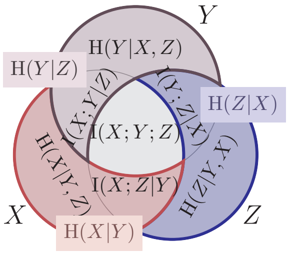


Figure 1: $\mathbf{I}_\mu$.

We now review the critical information theoretic concepts and their relationships to (in)dependence (see Section 4.D.1 for a full primer). Conditional entropy $\mathrm{H}_\mu(Y|X)$ measures how far μ is from satisfying the functional dependency $X \twoheadrightarrow Y$. Conditional mutual information $\mathrm{I}_\mu(Y;Z|X)$ measures how far μ is from satisfying the conditional independence $Y \perp\!\!\!\perp Z \mid X$. Linear combinations of these quantities (for $X, Y, Z \subseteq \mathcal{X}$) can be viewed as the inner product between a coefficient vector $\mathbf{v}$ and a $2^{|\mathcal{X}|} - 1$ dimensional vector $\mathbf{I}_\mu$ that we will call the *information profile* of μ. For three variables, the components of this vector are illustrated in Figure 1 (right). It is not hard to see that an arbitrary conjunction of (conditional) (in)dependencies can be expressed as a constraint $\mathbf{I}_\mu \cdot \mathbf{v} \geq 0$, for some appropriate vector $\mathbf{v}$.

We now formally introduce the qualitative PDG scoring function IDef, which interprets a hypergraph structure $\mathcal{A}$ as a function of the form $\mathbf{I}_\mu \cdot \mathbf{v}_\mathcal{A}$. This *information deficiency*, given by

$$\mathit{IDef}_{\mathcal{A}}(\mu) = \mathbf{I}_\mu \cdot \mathbf{v}_\mathcal{A} := -\operatorname{H}_\mu(\mathcal{X}) + \sum_{a \in \mathcal{A}} \operatorname{H}_\mu(T_a \mid S_a), \tag{4.2}$$

is the difference between the number of bits needed to (independently) specify the randomness in μ along the hyperarcs of $\mathcal{A}$, and the number of bits needed to specify a sample of μ according to its own structure ($\emptyset \to \mathcal{X}$). While IDef has some nice properties[5], it can also behave unintuitively in some cases; for instance, it can be negative. Clearly, it does not measure how close μ is to being structurally compatible with $\mathcal{A}$, in general. Nevertheless, there is still a fundamental relationship between IDef and QIM-compatibility, as we now show.

4.4.1 A Necessary Condition for QIM-Compatibility

What constraints does QIM-compatibility with $\mathcal{A}$ place on a distribution μ? When G is a dag, we have seen that if $\mu \models \Diamond\mathcal{A}_G$, then μ must satisfy the independencies

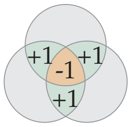
[5] It captures BN independencies and the dependencies of Theorem 4.2.2, reduces to maximum entropy for the empty hypergraph, and combines with the quantitative PDG scoring function [32] to capture factor graphs.

of the corresponding Bayesian network (Theorem 4.2.1); we have also seen that additional hyperarcs impose functional dependencies (Theorem 4.2.2). But these results apply only when $\mathcal{A}$ is of a very special form. More generally, $\mu \models \Diamond\mathcal{A}$ implies that μ can arise from some randomized causal model whose equations have dependency structure $\mathcal{A}$ (Propositions 4.3.1 and 4.3.2). Still, unless $\mathcal{A}$ has a particularly special form, it is not obvious whether or not this says something about μ. The primary result of this section is an information-theoretic bound (Theorem 4.4.1) that generalizes most of the concrete consequences of QIM-compatibility we have seen so far (Theorems 4.2.1 and 4.2.2). 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.

[link to proof]

Theorem 4.4.1. *If $\mu \models \Diamond\mathcal{A}$, then $\mathit{IDef}_{\mathcal{A}}(\mu) \leq 0$.*

Theorem 4.4.1 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 4.2.1 and 4.2.2 are immediate consequences of it. To illustrate some of its subtler implications, let's return to the 3-cycle in Example 9.



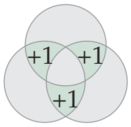
Example 10. It is easy to see (e.g., by inspecting Figure 1) that $\mathit{IDef}_{\text{3-cycle}}(\mu) = \mathrm{H}_\mu(Y|X) + \mathrm{H}_\mu(Z|Y) + \mathrm{H}_\mu(X|Z) - \mathrm{H}_\mu(XYZ) = -\,\mathrm{I}_\mu(X;Y;Z)$. Theorem 4.4.1 therefore tells us that a distribution μ that is QIM-compatible with the 3-cycle cannot have negative interaction information $\mathrm{I}_\mu(X;Y;Z)$. What does this mean? Overall, conditioning on the value of one variable can only reduce the amount of remaining information in other variables (in expectation). When $\mathrm{I}(X;Y;Z) < 0$, conditioning on one variable causes the other two to share more information than they did

before. The most extreme instance is μ_{xor}, the distribution in which two variables are independent and the third is their parity (illustrated on the right). It seems intuitively clear that μ_{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 4.4.1. $\triangle$

For many, there is an intuition that $\mathrm{I}(X;Y;Z) < 0$ should require a fundementally "3-way" interaction between the variables, and should not arise through pairwise interactions alone [14]. This has been a source of conflict [37, 24, 23, 5], because traditional ways of making precise "pairwise interactions" (e.g., maximum entropy subject to pairwise marginal constraints and pairwise factorization) do not ensure that $\mathrm{I}(X;Y;Z) \geq 0$. 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 $\mathbf{I}_\mu$. 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 [15]. 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 11. 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 μ_{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 $\mathrm{I}(A; B; C) = 0$.

This example has been used to argue that multivariate Shannon information does not take into account important structural differences between distributions [15]. We are now in a position to give a novel and particularly persuasive response, by appealing to QIM-compatibility.[6] 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 μ_{xor} (which is certainly not compatible with the 3-cycle) together with a seemingly irrelevant random bit X_1. By the results of Section 4.3, 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. $\triangle$

[6] 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.

In light of Example 11, one might reasonably conjecture that the converse of Theorem 4.4.1 holds. Unfortunately, it does not (see Section 4.D.4); the quantity $\mathit{IDef}_{\mathcal{A}}(\mu)$ 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 4.4.1 to distributions that are only "near-compatible" with $\mathcal{A}$.

4.4.2 A Scoring Function for QIM-Compatibility

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

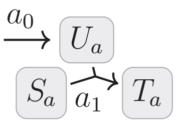
$$\mathrm{QIM}\mathit{Inc}_{\mathcal{A}}(\mu) := \inf_{\substack{\nu(\mathcal{U},\mathcal{X})\\ \nu(\mathcal{X})=\mu(\mathcal{X})}} -\mathrm{H}_\nu(\mathcal{U}) + \sum_{a\in\mathcal{A}} \mathrm{H}_\nu(U_a) + \sum_{a\in\mathcal{A}} \mathrm{H}_\nu(T_a|S_a,U_a). \tag{4.3}$$

$\mathrm{QIM}\mathit{Inc}$ is a direct translation of Definition 4.2.2 (a-c); it measures the (optimal) quality of an extended distribution ν as a witness. The infimum restricts the search to ν satisfying (a), the first two terms measure ν's discrepancy of with (b), and the last term measures ν's discrepancy with (c). Therefore:

link to proof

Proposition 4.4.2. $\mathrm{QIM}\mathit{Inc}_{\mathcal{A}}(\mu) \geq 0$, *with equality iff* $\mu \models \mathcal{A}$.

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

$\boxed{S_a} \xrightarrow{a} \boxed{T_a}$ with the pair of hyperarcs 

.

Finally, add one additional hyperarc $\mathcal{U} \to \mathcal{X}$. (Intuitively, this hyperarc creates functional dependencies in the spirit of Theorem 4.2.2.) With these definitions in place, we can state a theorem that bounds $\mathrm{QIM}\mathit{Inc}$ above and below with information deficiencies. The lower bound generalizes Theorem 4.4.1 by giving an upper limit on $\mathit{IDef}_{\mathcal{A}}(\mu)$ even for distributions μ that are not QIM-compatible with $\mathcal{A}$. The upper bound is tight in general, and shows that $\mathrm{QIM}\mathit{Inc}_{\mathcal{A}}$ can be equivalently defined as a minimization over $\mathit{IDef}_{\mathcal{A}^\dagger}$.

link to proof

Theorem 4.4.3. (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 ν to additional variables $\mathcal{U} = \{U_a\}_{a \in \mathcal{A}}$ indexed by $\mathcal{A}$, then:*

$$\mathit{IDef}_{\mathcal{A}}(\mu) \le \mathrm{QIM}\mathit{Inc}_{\mathcal{A}}(\mu) \le \mathit{IDef}_{\mathcal{A}^\dagger}(\nu).$$

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

$$\mathrm{QIM}\mathit{Inc}_{\mathcal{A}}(\mu) = \min \left\{ \mathit{IDef}_{\mathcal{A}^\dagger}(\nu) : \begin{array}{c} \nu \in \Delta\mathcal{V}(\mathcal{X},\mathcal{U}) \\ \nu(\mathcal{X}) = \mu(\mathcal{X}) \end{array} \right\}.$$

where the minimization is over all possible ways of assigning values to the variables in $\mathcal{U}$. The minimum is achieved when $|\mathcal{V}(U_a)| \le |\mathcal{V}(T_a)|^{|\mathcal{V}(S_a)|}$.

4.5 Discussion

We have shown how directed hypergraph s 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 4

4.A Proofs

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

4.A.1 From CPDs to Distributions over Functions

Compare two objects:

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

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

$$p(Y{=}y \mid X{=}x) := q(Y^X(x) = y).$$

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:

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

Think of Y^X as a collection of variables $\{Y^x : x \in \mathcal{V}X\}$ 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 $x \in VX$, and hence is essentially a funciton of type $\mathcal{V}X \to \Delta\mathcal{V}Y$, the same as p. In addition, if we apply the previous construction, we recover p, since:

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

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.

4.A.2 Results on (In)dependence

Lemma 4.A.1. *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 $f_i : \mathcal{V}(X_i) \to Y_i$. Then if $X_1, \dots, X_n$ are mutually independent (according to a joint distribution μ), 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 $\mathbf{y} = (y_1, \dots, y_n) \in Y_1 \times \cdots \times Y_n$, we have

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

□

Lemma 4.A.2 (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 $f : V(A) \to V(B)$ such that $\nu(f(Y) = X) = 1$, i.e., the event $f(A) = B$ occurs with probability 1.

1. Let $f : \mathcal{V}(A) \to \mathcal{V}(B)$ and $g : \mathcal{V}(A) \to \mathcal{V}(C)$ 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 $f : \mathcal{V}(A) \to \mathcal{V}(B)$ as before, and $g : \mathcal{V}(B) \to \mathcal{V}(C)$, such that the events $f(A) = B$ and $g(B) = C$ occur with proability 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 4.2.1. *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 μ satisfies $\mathcal{I}(G)$.*

Proof. Label the vertics of $G = (\mathcal{N}, E)$ by natural numbers so that they are a topological sort of G—that is, without loss of generality, suppose $\mathcal{N} = [n] := \{1, 2, \ldots, n\}$, and $i < j$ whenever $i \to j \in E$. 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 $\mathcal{X} = (X_1, \ldots, X_n)$ for the variables $\mathcal{X}$ corresponding to $\mathcal{N}$ over which μ 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 $\mathcal{U} = (U_1, \ldots, U_n)$, 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 $\bar{\mu} \models (U_1, \ldots, U_k) \twoheadrightarrow (X_1, \ldots, X_k)$ 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 $\bar{\mu} \models (X_1, \ldots, X_k)$ for some $k < n$. Now, conditon (c) of Definition 4.2.2 says $\bar{\mu} \models (S_{k+1}, U_{k+1}) \twoheadrightarrow X_{k+1}$. Because the varaibles are sorted in topological order, the parent variables S_{k+1} are a subset of $\{X_1, \ldots, X_n\}$, which are determined by $\mathcal{U}$ by the induction hypothesis; at the same time clearly $\bar{\mu} \models (U_1, \ldots, U_{k+1}) \twoheadrightarrow U_{k+1}$ as well. So, by two instances of Lemma 4.A.2, $\bar{\mu} \models (U_1, \ldots U_{k+1}) \twoheadrightarrow X_{k+1}$. Combining with our inductive hypothesis, we find that $\bar{\mu} \models (U_1, \ldots U_{k+1}) \twoheadrightarrow (X_1, \ldots, X_{k+1})$. So, by induction, $\bar{\mu} \models (U_1, \ldots, U_k) \twoheadrightarrow (X_1, \ldots, X_k)$ for $k \in [n]$, and in particular, $\bar{\mu} \models \mathcal{U} \twoheadrightarrow \mathcal{X}$.

With this in mind, we now return to proving that μ 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 $k \in [n]$, $\mu(X_1, \ldots, X_k) = \mu(X_1, \ldots, X_{k-1}) \mu(X_k \mid S_k)$. By QIM-compatibility witness condition (c), we know that $\bar{\mu} \models (S_k, U_k) \twoheadrightarrow X_k$, and so there exists a function $f_k : \mathcal{V}(S_k) \times \mathcal{V}(U_k) \to \mathcal{V}(X_k)$ for which the event $f_k(S_k, U_k) = X_k$ occurs with probability 1. Since $\bar{\mu} \models (U_1, \ldots, U_{k-1}) \twoheadrightarrow (X_1, \ldots, X_{k-1})$, and U_k is independent of $(U_1, \ldots, U_{k-1})$, it follows from Lemma 4.A.1 that $\bar{\mu} \models (X_1, \ldots, X_{k-1}) \perp\!\!\!\perp U_k$. Thus

$$\begin{aligned}\mu(X_1, \ldots, X_{k-1}, X_k) &= \sum_{u \in \mathcal{V}(U_k)} \mu(X_1, \ldots, X_{k-1}) \bar{\mu}(U_k = u) \cdot \mathbb{1}[X_k = f_k(S_k, u)] \\ &= \mu(X_1, \ldots, X_{k-1}) \sum_{u \in \mathcal{V}(U_k)} \bar{\mu}(U_k = u) \cdot \mathbb{1}[X_k = f_k(S_k, u)]\end{aligned}$$

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 μ is a probability distribution, know that $\varphi(X_k, S_k)$ must be the conditional probability of X_k given $X_1, \ldots, X_{k-1}$, and it depends only on the variables S_k. Thus $\mu(X_1, \ldots, X_k) = \mu(X_1, \ldots, X_{k-1})\mu(X_k \mid S_k)$.

Therefore $\nu(\mathcal{X}) = \mu(\mathcal{X})$ factors as required by the BN G, meaning that μ has the independencies specified by G. (See Koller & Friedman Thm 3.2, for instance.)

($\Longleftarrow$). Suppose μ satiesfies the independencies of G, meaning that each node is conditionally independent of its non-descendents given its parents. We now repeatedly apply the construction Section 4.A.1 to construct a QIM-compatibility witness. Specifically, for $k \in \{1, \ldots, n\}$, 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 $(U_1, \ldots, U_n)$, and observe that a setting $\mathbf{g} = (g_1, \ldots, g_n)$ of $\mathcal{U}$ uniquely picks out a value of $\mathcal{X}$, by evaluating each function in order. Let's call this function $f : \mathcal{V}(\mathcal{U}) \to \mathcal{V}(\mathcal{X})$.

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 $X_1, \ldots, X_{i-1}$, 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 g $\in \mathcal{V}(\mathcal{U})$ and x $\in \mathcal{V}(\mathcal{X})$, the function f is characterized by the property

$$f(\mathbf{g}) = \mathbf{x} \quad\iff\quad \bigwedge_{i=1}^{n} g_i(\mathbf{x}[S_i]) = \mathbf{x}[X_i]. \tag{4.4}$$

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 (4.4) 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 4.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 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 4.2.2(b)), and satisfy $(S_k, U_k) \twoheadrightarrow X_k$ for all $k \in [n]$ (Definition 4.2.2(c)). It remains only to verify that the marginal of $\bar{\mu}$ on the variables $\mathcal{X}$ is the original distribution μ (Definition 4.2.2(a)). Here is where we rely on the fact that μ satisfies the independencies of G, which means that we can factor $\mu(\mathcal{X})$ as

$\mu(\mathcal{X}) = \prod_{i=1}^{n} \mu(X_i \mid S_i)$.

$$
\begin{aligned}
\bar{\mu}(\mathcal{X}{=}\mathbf{x}) &= \sum_{\mathbf{g}\in\mathcal{V}(\mathcal{U})} \bar{\mu}(\mathcal{U}{=}\mathbf{g}) \cdot \delta f(\mathbf{x} \mid \mathbf{g}) \\
&= \sum_{(g_1,\ldots,g_n)\in\mathcal{V}(\mathcal{U})} \mathbb{1}\big[\mathbf{x} = f(\mathbf{g})\big] \; \prod_{i=1}^{n} \bar{\mu}(U_i{=}g_i) \\
&= \sum_{(g_1,\ldots,g_n)\in\mathcal{V}(\mathcal{U})} \mathbb{1}\Big[\bigwedge_{i=1}^{n} g_i(\mathbf{x}[S_i]) = \mathbf{x}[X_i]\Big] \; \prod_{i=1}^{n} \bar{\mu}(U_i{=}g_i) \qquad \text{[by (4.4)]} \\
&= \prod_{i=1}^{n} \; \sum_{g\in\mathcal{V}(U_i)} \mathbb{1}\big[g(\mathbf{x}[S_i]) = \mathbf{x}[X_i]\big] \cdot \bar{\mu}(U_i = g) \\
&= \prod_{i=1}^{n} \bar{\mu}\Big(\Big\{ g \in \mathcal{V}(U_i) \;\Big|\; g(\mathbf{s}_i) = x_i \Big\}\Big) \quad \text{where} \quad \begin{array}{l} x_i := \mathbf{x}[X_i], \\ \mathbf{s}_i := \mathbf{x}[S_i] \end{array} \\
&= \prod_{i=1}^{n} \bar{\mu}\big(U_i(\mathbf{s}_i) {=} x_i\big) \\
&= \prod_{i=1}^{n} \mu(X_i = x_i \mid S_i = \mathbf{s}_i) \\
&= \mu(\mathcal{X} = \mathbf{x}).
\end{aligned}
$$

Therefore, when μ satisfies the independencies of a BN G, it is QIM-compatible with $\mathcal{A}_G$. □

Before we move on to proving the other results in the paper, we first illustrate how this relatively substantial first half of the proof of Theorem 4.2.1 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 4.4.1, $\mathit{IDef}_{\mathcal{A}_G}(\mu) \leq 0$. In the appendix of [32], it is shown that $\mathit{IDef}_{\mathcal{A}_G}(\mu) \geq 0$ with equality iff μ satisfies the BN's independencies. Thus μ must satisfy the appropriate independencies. □

Theorem 4.2.2.

(a) *$\mu \models X {\twoheadrightarrow} Y \land \Diamond\mathcal{A}$ if and only if $\forall n \geq 0.\ \mu \models \Diamond\mathcal{A} \sqcup^{(+n)}_{X\to Y}$.*

(b) *if $\mathcal{A} = \mathcal{A}_G$ for a dag G, then $\mu \models X {\twoheadrightarrow} Y \land \Diamond\mathcal{A}$ if and only if $\mu \models \Diamond\mathcal{A} \sqcup^{(+1)}_{X\to Y}$.*

(c) *if $\exists a \in \mathcal{A}$ such that $S_a = \emptyset$ and $X \in T_a$, then $\mu \models X {\twoheadrightarrow} Y \land \Diamond\mathcal{A}$ iff $\mu \models \Diamond\mathcal{A} \sqcup^{(+2)}_{X\to Y}$.*

Proof. **(a).** The forward direction is straightforward. Suppose that $\mu \models \mathcal{A}$ and $\mu \models X \twoheadrightarrow Y$. 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" ν in the unique way to n additional constant variables $U_1, \ldots, U_n$, each of which can only take on one value. We claim that this "extended" distribution ν', which we conflate with ν because it is not meaningfully different, is a witness to $\mu \models \Diamond\mathcal{A} \sqcup^{(+n)}_{X\to Y}$. Since $\mu \models X \twoheadrightarrow Y$ it must also be that $\nu \models X \twoheadrightarrow Y$, and it follows that $\nu \models (X, U_i) \twoheadrightarrow Y$ for all $i \in \{1, \ldots, n\}$, demonstrating that the new requirements of ν' imposed by Definition 4.2.2(c) hold. (The remainder of the requirements for condition (c), namely that $\nu' \models (S_a, U_a) \twoheadrightarrow T_a$ for $a \in \mathcal{A}$, still hold because ν' is an extension of ν, 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 ν (or, more precisely, an isomorphic "extension" of it to additional trivial variables) is a witness of $\mu \models \Diamond\mathcal{A} \sqcup^{(+n)}_{X\to Y}$.

The reverse direction is difficult to prove directly, yet it is a straightforward application of Theorem 4.4.1. Suppose that $\mu \models \mathcal{A} \sqcup^{(+n)}_{X\to Y}$ for all $n \geq 0$. By

Theorem 4.4.1, we know that

$$0 \geq \mathit{SDef}_{\mathcal{A} \sqcup {}^{(+n)}_{X \to Y}}(\mu) = \mathit{SDef}_{\mathcal{A}}(\mu) + n \, \mathrm{H}_\mu(Y|X).$$

Because $\mathit{SDef}_{Ar}(\mu)$ is bounded below (by $-\log|\mathcal{V}(\mathcal{X})|$), it cannot be the case that $\mathrm{H}_\mu(Y|X) > 0$; otherwise, the inequality above would not hold for large n (specifically, for $n > \log|\mathcal{V}(\mathcal{X})| / \mathrm{H}_\mu(Y|X)$). By Gibbs inequailty, $\mathrm{H}_\mu(Y|X)$ is non-negative, and thus it must be the case that $\mathrm{H}_\mu(Y|X) = 0$. Thus $\mu \models X \twoheadrightarrow Y$. It is also true that $\mu \models \Diamond\mathcal{A}$ by monotonicity (Theorem 4.C.2), which is itself a direct application of Theorem 4.4.1

(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 $\mu \models \Diamond\mathcal{A} \sqcup {}^{(+n)}_{X \to Y}$ 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 4.4.1. Suppose $\mu \models \Diamond\mathcal{A} \sqcup {}^{(+1)}_{X \to Y}$. In this case where $\mathcal{A} = \mathcal{A}_G$, it was shown by Richardson and Halpern [32] that $\mathit{SDef}_{\mathcal{A}}(\mu) \geq 0$. It follows that

$$0 \overset{\text{(Theorem 4.4.1)}}{\geq} \mathit{SDef}_{\mathcal{A} \sqcup {}^{(+n)}_{X \to Y}}(\mu) = \mathit{SDef}_{\mathcal{A}}(\mu) + \mathrm{H}_\mu(Y|X) \;\geq\; 0,$$

and thus $\mathrm{H}_\mu(Y|X) = 0$, 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 $\mathcal{A} \rightsquigarrow \{\to \{X\}\}$, suppose that $\mu \models \Diamond\mathcal{A} \sqcup {}^{(+2)}_{X \to Y}$. By monotonicity, this means $\mu \models \mathcal{A}$ and also that $\mu \models \to \boxed{X} \rightrightarrows \boxed{Y}$. Let $\mathcal{A}'$ denote this hypergraph. Once again by appeal to Theorem 4.4.1, we have that

$$0 \geq \mathit{SDef}_{\mathcal{A}'} = -\mathrm{H}_\mu(X,Y) + \mathrm{H}(X) + 2\,\mathrm{H}_\mu(Y|X) = \mathrm{H}_\mu(Y|X) \geq 0.$$

It follows that $\mathrm{H}_\mu(Y|X) = 0$, and thus $\mu \models X \twoheadrightarrow Y$. As mentioned above, we also know that $\mu \models \mathcal{A}$, and thus $\mu \models \Diamond\mathcal{A} \ \land \ X\twoheadrightarrow Y$ as promised. □

4.A.3 Causality Results of Section 4.3

Proposition 4.3.1. *Given a graph G and a distribution μ, $\mu \models \Diamond\mathcal{A}_G$ iff there exists a fully randomized PSEM of dependency structure $\mathcal{A}_G$ from which μ can arise.*

Proof. **($\Longrightarrow$).** 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 4.2.2. Because $\mathcal{A}_G$ is partitional, the elements of $\mathrm{PSEMs}_{\mathcal{A}_G}(\bar{\mu})$ are ordinary (i.e., not generalized) randomized PSEMs. We claim that every $\mathcal{M} = (M, P) \in \mathrm{PSEMs}_{\mathcal{A}_G}(\bar{\mu})$ that is a randomized PSEM from which μ can arise, and also has the property that $\mathbf{Pa}_M(Y) \subseteq \mathbf{Pa}_G(Y) \cup \{U_Y\}$ 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 4.2.2), these variables $\{U_X\}_{X\in\mathcal{X}}$ are mutually independent according to $\bar{\mu}$. Furthermore, because $\mathcal{M} = (M, P) \in \mathrm{PSEMs}_{\mathcal{A}_G}(\bar{\mu})$, 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 $\mathcal{M} \in \mathrm{PSEMs}_{\mathcal{A}_G}(\bar{\mu})$ the equation f_Y can depend only on $S_Y = \mathbf{Pa}_G(Y) \subseteq \mathcal{X}$ 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 4.2.2), we know that $\bar{\mu}(\mathcal{X}) = \mu$. By Proposition 4.3.3(a), we know that $\mu \in \{\!\{\mathcal{M}\}\!\}$. Together, the previous two sentences mean that μ can arise from $\mathcal{M}$.

- Finally, as mentioned in the first bullet item, the equation f_Y in M can depend only on $S_Y = \mathbf{Pa}_G(Y)$ and on U_Y. Thus $\mathbf{Pa}_M(Y) \subseteq \mathbf{Pa}_G(Y) \cup \{U_Y\}$ 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 μ can arise, with the property that $\mathbf{Pa}_{\mathcal{M}}(Y) \subseteq \mathbf{Pa}_G(Y) \cup \{U_Y\}$ for all $Y \in \mathcal{X}$.

($\Longleftarrow$). Conversely, suppose there is a randomized PSEM $\mathcal{M} = (M = (\mathcal{Y}, \mathcal{U}, \mathcal{F}), P)$ with the property that $\mathbf{Pa}_M(Y) \subseteq \mathbf{Pa}_G(Y) \cup \{U_Y\}$ for all Y, from which μ can arise. The last clause means there exists some $\nu \in \{\!\!\{\mathcal{M}\}\!\!\}$ such that $\nu(\mathcal{X}) = \mu$. We claim that this ν 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 $Y \in \mathcal{X}$, $\nu \models \mathbf{Pa}_G(Y) \cup \{U_Y\} \twoheadrightarrow Y$. Since $\nu \in \{\!\!\{\mathcal{M}\}\!\!\}$, we know that M's equation holds with probability 1 in ν, and so it must be the case that $\nu \models \mathbf{Pa}_M(Y) \twoheadrightarrow Y$. Note that, in general, if $\mathbf{A} \subseteq \mathbf{B}$ and $\mathbf{A} \twoheadrightarrow \mathbf{C}$, then $\mathbf{B} \twoheadrightarrow \mathbf{C}$. By assumption, $\mathbf{Pa}_M(Y) \subseteq \mathbf{Pa}_G(Y) \cup \{U_Y\}$, and thus $\nu \models \mathbf{Pa}_G(Y) \cup \{U_Y\} \twoheadrightarrow Y$.

Thus ν satisfies all conditions (a-c) for a QIM-compatibility witness, and hence $\mu \models \mathcal{A}_G$. □

Proposition 4.3.2. *$\mu \models \Diamond\mathcal{A}$ iff there exists a generalized randomized PSEM with structure $\mathcal{A}$ from which μ can arise.*

Proof. **($\Longrightarrow$).** Suppose $\mu \models \mathcal{A}$, meaning there exists a witness $\nu(\mathcal{X}, \mathcal{U})$ with property Definition 4.2.2(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 ν. Meanwhile, by Definition 4.2.2(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 4.2.2(a), we know that μ can arise from $\mathcal{M}$.

($\Longleftarrow$). 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 μ. We claim that this ν is also a witness that $\mu \models \mathcal{A}$. The marginal constraint from Definition 4.2.2(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. □

Proposition 4.3.3. *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 $\mathcal{M} \in \mathrm{PSEMs}_{\mathcal{A}}(\bar{\mu})$.*

Proof. (a) is straightforward. Suppose $\mathcal{M} \in \mathrm{PSEMs}(\nu)$. By construction, the equations of $\mathcal{M}$ reflect functional dependencies in ν, and hence hold with probability 1.[7] Furthermore, the distribution $P(\mathcal{U})$ in all $\mathcal{M} \in \mathrm{PSEMs}(\nu)$ is equal to $\nu(\mathcal{U})$. These two facts, demonstrate that ν satisfies the two constraints required for membership in $\{\!\!\{\mathcal{M}\}\!\!\}$.

[7] 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 $\mathcal{M} \in \mathrm{PSEMs}(\nu)$.

(b). We do the two directions separately. First, suppose $\mathcal{M} \in \mathrm{PSEMs}(\nu)$. We have already shown (in part (a)) that $\nu \in \{\!\!\{\mathcal{M}\}\!\!\}$. The construction of $\mathrm{PSEMs}(\nu)$ 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, $\mathbf{Pa}_{\mathcal{M}}(X) \subseteq S_{a_X} \cup \{U_{a_X}\}$.

Conversely, suppose $\mathcal{M} = (\mathcal{X}, \mathcal{U}, \mathcal{F})$ is a PSEM satisfying $\nu \in \{\!\!\{\mathcal{M}\}\!\!\}$ and $\mathbf{Pa}_{\mathcal{M}}(X) \subseteq S_{a_X} \cup \{U_{a_X}\}$. We would like to show that $\mathcal{M} \in \mathrm{PSEMs}(\nu)$. 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 $\mathbf{s} \in \mathcal{V}(S_{a_X})$ and $u \in \mathcal{V}(U_{a_X})$. If $\nu(\mathbf{s}, u) > 0$, then we know there is a unique value of $x \in \mathcal{V}(X)$ such that $\nu(\mathbf{s}, u, x) > 0$. Since $\mathcal{M}$'s equation for X, f_X, depends only on s and u, and holds with probability 1, we know that $f_X(\mathbf{s}, u) = t$, as required. On the other hand, if $\nu(\mathbf{s}, u) = 0$, then any choice of $f_X(\mathbf{s}, u)$ 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 $\mathcal{M} \in \mathrm{PSEMs}(\nu)$. $\square$

Theorem 4.3.4. *Suppose that $\bar{\mu}$ is a witness to $\mu \models \Diamond\mathcal{A}$, $\mathcal{M} \in \mathrm{PSEMs}_{\mathcal{A}}(\bar{\mu})$, $\mathbf{X} \subseteq \mathcal{X}$ and $\mathbf{x} \in \mathcal{V}(\mathbf{X})$. If $\bar{\mu}(\mathrm{do}_{\mathcal{M}}(\mathbf{X}{=}\mathbf{x})) > 0$, then:*

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

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

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 $\nu(\mathrm{do}_M(\mathbf{X}{=}\mathbf{x})) > 0$, the conditional distribution

$$\nu \mid \mathrm{do}_M(\mathbf{X}{=}\mathbf{x}) = \nu(\mathcal{U}, \mathcal{X}) \cdot \prod_{X \in \mathbf{X}} \mathbb{1}\big[\forall \mathbf{s}. f_X(U_X, \mathbf{s}) = \mathbf{x}[X]\big] \Big/ \nu(\mathrm{do}_M(\mathbf{X}{=}\mathbf{x}))$$

is defined. By assumption, $\mathcal{M} \in \mathrm{PSEMs}(\nu)$ and ν is a witness to $\mu \models \mathcal{A}$. Thus, by Proposition 4.3.3, 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 ν. But the support of the conditional distribution $\nu \mid \mathrm{do}_M(\mathbf{X}{=}\mathbf{x})$ is a subset of the support of ν, so all equations of $\mathcal{M}$ also hold in the conditioned distribution. Furthermore, the event $\mathrm{do}_M(\mathbf{X}{=}\mathbf{x})$ is the event in which, for all $X \in \mathbf{X}$, the variable U_X takes on a value such that $f_X(\ldots, U_X, \ldots) = \mathbf{x}[X]$. Thus the equations corresponding to $\mathbf{X} = \mathbf{x}$ also hold with probability 1 in $\nu \mid \mathrm{do}_M(\mathbf{X}{=}\mathbf{x})$.

This shows that all equations of $\mathcal{M}_{\mathbf{X} \leftarrow \mathbf{x}}$ hold with probability 1 in $\nu \mid \mathrm{do}_M(\mathbf{X}{=}\mathbf{x})$. However, the marginal distribution $\nu(\mathcal{U} \mid \mathrm{do}_M(\mathbf{X}{=}\mathbf{x}))$ 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, $\nu \mid \mathrm{do}_M(\mathbf{X}{=}\mathbf{x}) \notin \{\!\!\{\mathcal{M}_{\mathbf{X} \leftarrow \mathbf{x}}\}\!\!\}$. Our objective, therefore, is to show that there is a *different* distribution $\nu' \in \{\!\!\{\mathcal{M}_{\mathbf{X} \leftarrow \mathbf{x}}\}\!\!\}$ such that $\nu'(\mathcal{X}) = \nu(\mathcal{X} \mid \mathrm{do}_M(\mathbf{X}{=}\mathbf{x}))$. Let $\mathbf{Z} := \mathcal{X} \setminus \mathbf{X}$, and $\mathcal{U}_{\mathbf{Z}} := \{U_Z : Z \in \mathbf{Z}\}$. We can define ν' according to

$$\nu'(\mathcal{X}, \mathcal{U}_{\mathbf{X}}, \mathcal{U}_{\mathbf{Z}}) := \nu(\mathcal{X}, \mathcal{U}_{\mathbf{Z}} \mid \mathrm{do}_M(\mathbf{X}{=}\mathbf{x})) P(\mathcal{U}_{\mathbf{X}}).$$

This distribution satisfies three critical properties:

1. Clearly ν' has the appropriate marginal $\nu'(\mathcal{X}) = \nu(\mathcal{X} \mid \mathrm{do}_M(\mathbf{X}{=}\mathbf{x}))$ on exogenous variables $\mathcal{X}$, by construction.

2. At the same time, the marginal on exogenous variables is

$$\begin{aligned}
\nu'(\mathcal{U}) &= \nu'(\mathcal{U}_{\mathbf{X}}, \mathcal{U}_{\mathbf{Z}}) \\
&= \int_{\mathcal{V}(\mathcal{X})} \nu(\omega, \mathcal{U}_{\mathbf{Z}} \mid \mathrm{do}_M(\mathbf{X}{=}\mathbf{x})) P(\mathcal{U}_{\mathbf{X}}) \,\mathrm{d}\omega \\
&= P(\mathcal{U}_{\mathbf{X}}) \nu(\mathcal{U}_{\mathbf{Z}} \mid \mathrm{do}_M(\mathbf{X}{=}\mathbf{x})) \\
&= P(\mathcal{U}_{\mathbf{X}}) P(\mathcal{U}_{\mathbf{Z}} \mid \mathrm{do}_M(\mathbf{X}{=}\mathbf{x})) && \left[\, \text{since } \mathrm{do}_M(\mathbf{X}{=}\mathbf{x}) \text{ depends only on } \mathcal{U} \,\right] \\
&= P(\mathcal{U}_{\mathbf{X}}) P(\mathcal{U}_{\mathbf{Z}}) && \left[\begin{array}{l} \text{since } \mathrm{do}_M(\mathbf{X}{=}\mathbf{x}) \text{ depends} \\ \text{only on } \mathcal{U}_{\mathbf{X}}\text{, while } \mathcal{U}_{\mathbf{X}} \text{ and} \\ \mathcal{U}_{\mathbf{Z}} \text{ are independent in } \nu \text{ (by} \\ \text{the witness condition).} \end{array} \right] \\
&= P(\mathcal{U}_{\mathbf{X}}, \mathcal{U}_{\mathbf{Z}}) && \left[\, \text{same reason as above} \,\right]
\end{aligned}$$

3. Finally, ν' 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 $\nu \mid \mathrm{do}_M(\mathbf{X}{=}\mathbf{x})$—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 $\nu(\mathcal{X} \mid \mathrm{do}_M(\mathbf{X}{=}\mathbf{x}))$ can arise from $\mathcal{M}_{\mathbf{X}\leftarrow\mathbf{x}}$.

(part b). We will again make use of the distribution ν' 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

$$\mathcal{F}_{\mathbf{X}\leftarrow\mathbf{x}}(\mathbf{u}) := \left\{ \omega \in \mathcal{V}(\mathcal{X}) \;\middle|\; \begin{array}{c} \forall X \in \mathbf{X}. \quad \omega[X] = \mathbf{x}[X] \\ \forall Y \in \mathcal{X} \setminus \mathbf{X}. \quad \omega[Y] = f_X(\omega[\mathcal{X} \setminus Y], \mathbf{u}) \end{array} \right\}$$

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

$$\begin{aligned}(M, \mathbf{u}) \models [\mathbf{X}{\leftarrow}\mathbf{x}]\varphi \quad &\iff \quad (M_{\mathbf{X}\leftarrow\mathbf{x}}, \mathbf{u}) \models \varphi \\ &\iff \forall \omega \in \mathcal{F}_{\mathbf{X}\leftarrow\mathbf{x}}(\mathbf{u}).\ \omega \in \varphi \\ &\iff \mathcal{F}_{\mathbf{X}\leftarrow\mathbf{x}}(\mathbf{u}) \ \subseteq \ \varphi,\end{aligned}$$

then ϕ holds at all points that satisfy the equations of $M_{\mathbf{X}\leftarrow\mathbf{x}}$. So, since ν' is supported only on such points (property 3), it must be that $\nu'(\varphi) = 1$. By property 1, $\nu'(\varphi) = \nu(\varphi \mid \mathrm{do}_M(\mathbf{X}{=}\mathbf{x}))$.

Furthermore, if $\nu'(\varphi) > 0$, then there must exist some $\omega \in \mathcal{F}_{\mathbf{X}\leftarrow\mathbf{x}}(\mathbf{u})$ satisfying φ, 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:

$$\begin{aligned}\Pr_{\mathcal{M}}([\mathbf{X}{\leftarrow}\mathbf{x}]\varphi) &= P(\{\mathbf{u} \in \mathcal{V}(\mathcal{U}) : (M, \mathbf{u}) \models [\mathbf{X}{\leftarrow}\mathbf{x}]\varphi\}) \\ &= \sum_{\mathbf{u}\in\mathcal{V}(\mathcal{U})} P(\mathbf{u}) \mathbb{1}\big[\mathcal{F}_{\mathbf{X}\leftarrow\mathbf{x}}(\mathbf{u}) \subseteq \varphi\big] \\ &\leq \sum_{\mathbf{u}\in\mathcal{V}(\mathcal{U})} P(\mathbf{u}) \nu'(\varphi \mid \mathbf{u}) \qquad = \nu'(\varphi) = \nu(\varphi \mid \mathrm{do}_M(\mathbf{X}{=}\mathbf{x})) \\ &\leq \sum_{\mathbf{u}\in\mathcal{V}(\mathcal{U})} P(\mathbf{u}) \mathbb{1}\big[\mathcal{F}_{\mathbf{X}\leftarrow\mathbf{x}}(\mathbf{u}) \cap \varphi \neq \emptyset\big] \\ &= P(\{\mathbf{u} \in \mathcal{V}(\mathcal{U}) : (M, \mathbf{u}) \models \langle \mathbf{X}{\leftarrow}\mathbf{x} \rangle \varphi\}) \\ &= \Pr_{\mathcal{M}}(\langle \mathbf{X}{\leftarrow}\mathbf{x} \rangle \varphi), \qquad \text{as desired.}\end{aligned}$$

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 φ 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,

$$(M, \mathbf{u}) \models [\mathbf{X}{\leftarrow}\mathbf{x}]\varphi \qquad \iff \qquad (M, \mathbf{u}) \models \langle \mathbf{X}{\leftarrow}\mathbf{x} \rangle \varphi,$$

and thus the probability of both formulas are the same—and it must also equal $\nu(\varphi \mid \mathrm{do}_M(\mathbf{X}{=}\mathbf{x}))$ which we have shown lies between them. $\square$

Proposition 4.C.4. *

Proof. 1. Suppose $\mu \models \mathcal{A}_G$, $\bar{\mu}(\mathcal{U}, \mathcal{X})$ is a witness to this, and $\mathcal{M} \in \mathrm{PSEMs}_{\mathcal{A}_G}(\bar{\mu})$.

($\Longrightarrow$). Suppose $\mathbf{X} \perp\!\!\!\perp_G \mathbf{Y} \mid \mathbf{Z}$, meaning every path from $\mathbf{X}$ to $\mathbf{Y}$ in G goes through $\mathbf{Z}$. We now show that

$\langle$ INCOMPLETE $\rangle$

$\square$

4.A.4 Information Theoretic Results of Section 4.4

To prove Theorem 4.4.1 and Theorem 4.4.3(a), we will need the following Lemma.

Lemma 4.A.3. *Consider a set of variables $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$, and another (set of) variable(s) X. Every joint distribution $\mu(X, \mathbf{Y})$ over the values of X and $\mathbf{Y}$ satisfies*

$$\sum_{i=1}^{n} \mathrm{I}_\mu(X\,;\,Y_i) \;\leq\; \mathrm{I}_\mu(X\,;\,\mathbf{Y}) + \sum_{i=1}^{n} \mathrm{H}_\mu(Y_i) - \mathrm{H}_\mu(\mathbf{Y}).$$

Proof. Since there is only one joint distribution in scope, we omit the subscript μ, writing $\mathrm{I}(-)$ instead of $\mathrm{I}_\mu(-)$ and $\mathrm{H}(-)$ instead of $\mathrm{H}_\mu(-)$, in the body of this proof. The following fact will also be very useful:

$$\mathrm{I}(A; B, C) = \mathrm{I}(A; C) + \mathrm{I}(A; B \mid C) \quad \text{(the chain rule for mutual information).} \tag{4.5}$$

We prove this by induction on n. In the base case ($n = 1$), we must show that $\mathrm{I}(X;Y) \le \mathrm{I}(X;Y) + \mathrm{H}(Y) - \mathrm{H}(Y)$, which is an obvious tautology. Now, suppose inductively that

$$\sum_{i=1}^{k} \mathrm{I}(X\,;\,Y_i) \;\le\; \mathrm{I}(X\,;\,\mathbf{Y}_{1:k}) + \sum_{i=1}^{k} \mathrm{H}(Y_i) - \mathrm{H}(\mathbf{Y}_{1:k}) \tag{IH$_k$}$$

for some $k < n$, where $\mathbf{Y}_{1:k} = (Y_1, \ldots, Y_k)$. We now prove that the analogue for $k + 1$ also holds. Some calculation reveals:

$$\begin{aligned}
\mathrm{I}(X;Y_{k+1}) &= \mathrm{I}(X;\mathbf{Y}_{1:k+1}) - \mathrm{I}(X;\mathbf{Y}_{1:k} \mid Y_{k+1}) && \Big[\,\text{by MI chain rule (4.5)}\,\Big] \\
&\le \mathrm{I}(X;\mathbf{Y}_{1:k+1}) && \Big[\,\text{since }\ \mathrm{I}(X;\mathbf{Y}_{1:k} \mid Y_{k+1}) \ge 0\,\Big] \\
&= \mathrm{I}(X;Y_{k+1} \mid \mathbf{Y}_{1:k}) + \mathrm{I}(\mathbf{Y}_{1:k};Y_{k+1}) && \Big[\,\text{by MI chain rule (4.5)}\,\Big] \\
&= \begin{pmatrix} \mathrm{I}(X;\mathbf{Y}_{1:k+1}) + \mathrm{H}(Y_{k+1}) - \mathrm{H}(\mathbf{Y}_{1:k+1}) \\ -\,\mathrm{I}(X;\mathbf{Y}_{1:k}) \qquad\qquad\qquad + \mathrm{H}(\mathbf{Y}_{1:k}) \end{pmatrix} && \left[\begin{array}{l} \text{left: one more MI chain rule (4.5);} \\ \text{right: defn of mutual information} \end{array}\right].
\end{aligned}$$

Observe: adding this inequality to our inductive hypothesis (IH$_k$) yields (IH$_{k+1}$)! So, by induction, the lemma holds for all k. □

Theorem 4.4.1. *If $\mu \models \Diamond\mathcal{A}$, then IDef$_{\mathcal{A}}(\mu) \le 0$.*

Proof. Suppose that $\mu \models \mathcal{A}$, meaning that there is a witness $\nu(\mathcal{X},\mathcal{U})$ that extends μ, and has properties (a-c) of Definition 4.2.2. For each hyperarc a, since $\nu \models (S_a, U_a) \twoheadrightarrow T_a$, we have $\mathrm{H}_\nu(T_a \mid S_a, U_a) = 0$, and so

$$\mathrm{H}_\mu(T_a \mid S_a) = \mathrm{H}_\nu(T_a \mid S_a, U_a) + \mathrm{I}_\nu(T_a; U_a \mid S_a) = \mathrm{I}_\nu(T_a; U_a \mid S_a).$$

Thus, we compute

$$\begin{aligned}
\sum_{a\in\mathcal{A}} \mathrm{H}_\mu(T_a \mid S_a) &= \sum_{a\in\mathcal{A}} \mathrm{I}_\nu(U_a;T_a \mid S_a) \\
&= \sum_{a\in\mathcal{A}} \mathrm{I}_\nu(U_a;T_a,S_a) - \mathrm{I}_\nu(U_a;S_a) && \text{by MI chain rule (4.5)} \\
&\le \sum_{a\in\mathcal{A}} \mathrm{I}_\nu(U_a;T_a,S_a) && \text{since } \mathrm{I}_\nu(U_a\,;\,S_a) \ge 0 \\
&\le \sum_{a\in\mathcal{A}} \mathrm{I}_\nu(U_a;\mathcal{X}) && \text{since } \mathcal{X} \twoheadrightarrow (S_a,T_a) \\
&\le \mathrm{I}_\nu(\mathcal{X};\mathcal{U}) + \sum_{a\in\mathcal{A}} \mathrm{H}_\nu(U_a) - \mathrm{H}_\nu(\mathcal{U}) && \text{by Lemma 4.A.3} \\
&= \mathrm{I}_\nu(\mathcal{X};\mathcal{U}) && \begin{array}{r}\text{since } \mathcal{U} \text{ are independent} \\ \text{(per condition (b) of Definition 4.2.2)}\end{array} \\
&\le \mathrm{H}_\nu(\mathcal{X}) = \mathrm{H}_\mu(\mathcal{X}). && \text{(per condition (a) of Definition 4.2.2)}
\end{aligned}$$

Thus, $\mathit{IDef}_{\mathcal{A}}(\mu) \le 0$. □

Proposition 4.4.2. $\mathrm{QIM}\mathit{Inc}_{\mathcal{A}}(\mu) \ge 0$, *with equality iff* $\mu \models \mathcal{A}$.

Proof. The first term in the definition of $\mathrm{QIM}\mathit{Inc}$ be written as

$$\Big(- \mathrm{H}_\nu(\mathcal{U}) + \sum_{a\in\mathcal{A}} \mathrm{H}_\nu(U_a) \Big) = \mathop{\mathbb{E}}_{\nu} \Big[\log \frac{\nu(\mathcal{U})}{\prod_a \nu(U_a)} \Big]$$

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 $\mathrm{QIM}\mathit{Inc}_{\mathcal{A}}(\mu)$, are all conditional entropies, and hence non-negative as well. Thus $\mathrm{QIM}\mathit{Inc}_{\mathcal{A}}(\mu) \ge 0$.

Now, suppose μ 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) $\mathrm{H}_\nu(T_a|S_a,U_a) = 0$, and (d) $\{U_a\}_{a\in\mathcal{A}}$ are mutually independent. Then clearly ν satisfies the condition under the infemum, every

$\mathrm{H}_\nu(T_a|S_a,U_a)$ is zero. It is also immediate that the final term is zero as well, because it equals $\boldsymbol{D}(\nu(\mathcal{U}) \parallel \prod_a \nu(U_a))$, and $\nu(\mathcal{U}) = \prod_a \nu(U_a)$, per the definition of mutual independence. Thus, ν witnesses that $\mathrm{QIM}\mathit{Inc}_{(\mathcal{A},\lambda)} = 0$.

Conversely, suppose $\mathrm{QIM}\mathit{Inc}_{(\mathcal{A},\lambda)} = 0$. Because the feasible set is closed and bounded, as is the function, the infemum is achieved by some joint distribution $\nu(\mathcal{X},\mathcal{A})$ with marginal $\mu(\mathcal{X})$. In this distribution ν, we know that every $\mathrm{H}_\nu(T_a|S_a,U_a) = 0$ and $\boldsymbol{D}(\nu(\mathcal{U}) \parallel \prod_a \nu(U_a)) = 0$— because if any of these terms were positive, then the result would be positive as well. So ν 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 ν satisfies (d) as well. □

Theorem 4.4.3.

(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 ν to additional variables $\mathcal{U} = \{U_a\}_{a\in\mathcal{A}}$ indexed by $\mathcal{A}$, then:*

$$\mathit{IDef}_{\mathcal{A}}(\mu) \leq \mathrm{QIM}\mathit{Inc}_{\mathcal{A}}(\mu) \leq \mathit{IDef}_{\mathcal{A}^\dagger}(\nu).$$

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

$$\mathrm{QIM}\mathit{Inc}_{\mathcal{A}}(\mu) = \min\left\{ \mathit{IDef}_{\mathcal{A}^\dagger}(\nu) : \begin{array}{c} \nu \in \Delta\mathcal{V}(\mathcal{X},\mathcal{U}) \\ \nu(\mathcal{X}) = \mu(\mathcal{X}) \end{array} \right\}.$$

where the minimization is over all possible ways of assigning values to the variables in $\mathcal{U}$. The minimum is achieved when $|\mathcal{V}(U_a)| \leq |\mathcal{V}(T_a)|^{|\mathcal{V}(S_a)|}$.

Proof. Part (a). The left hand side of the theorem ($\mathit{IDef}_{\mathcal{A}}(\nu) \leq \mathrm{QIM}\mathit{Inc}_{\mathcal{A}}(\mu)$) is a strengthening of the argument used to prove Theorem 4.4.1. Specifically, let ν^*

be a minimizer of the optimization problem defining $\mathrm{QIM}\mathit{Inc}$ We calculate

$$\begin{aligned}
&\mathrm{QIM}\mathit{Inc}_{\mathcal{A}}(\mu) - \mathit{IDef}_{\mathcal{A}}(\mu)\\
&= \left(\sum_{a\in\mathcal{A}} \mathrm{H}_{\nu^*}(T_a \mid S_a, U_a) - \mathrm{H}_{\nu^*}(\mathcal{U}) + \sum_{a\in\mathcal{A}} \mathrm{H}_{\nu^*}(U_a) \right) - \left(\sum_{a\in\mathcal{A}} \mathrm{H}_{\mu}(T_a \mid S_a) - \mathrm{H}_{\mu}(\mathcal{X}) \right)\\
&= \sum_{a\in\mathcal{A}} \Big(\mathrm{H}_{\nu^*}(T_a \mid S_a, U_a) - \mathrm{H}_{\nu^*}(T_a \mid S_a) \Big) + \mathrm{H}_{\mu}(\mathcal{X}) - \mathrm{H}_{\nu^*}(\mathcal{U}) + \sum_{a\in\mathcal{A}} \mathrm{H}_{\nu^*}(U_a)\\
&= -\sum_{a\in\mathcal{A}} \mathrm{I}_{\nu^*}(T_a; U_a \mid S_a) \qquad + \mathrm{H}_{\mu}(\mathcal{X}) - \mathrm{H}_{\nu^*}(\mathcal{U}) + \sum_{a\in\mathcal{A}} \mathrm{H}_{\nu^*}(U_a).
\end{aligned}$$

The argument given in the first five lines of the proof of Theorem 4.4.1, gives us a particularly convenient bound for the first group of terms on the left:

$$\sum_{a\in\mathcal{A}} \mathrm{I}_{\nu^*}(U_a; T_a \mid S_a) \le \mathrm{I}_{\nu^*}(\mathcal{X};\mathcal{U}) + \sum_{a\in\mathcal{A}} \mathrm{H}_{\nu^*}(U_a) - \mathrm{H}_{\nu^*}(\mathcal{U}).$$

Substituting this into our previous expression, we have:

$$\begin{aligned}
&\mathrm{QIM}\mathit{Inc}_{\mathcal{A}}(\mu) - \mathit{IDef}_{\mathcal{A}}(\mu)\\
&\ge -\Big(\mathrm{I}_{\nu^*}(\mathcal{X};\mathcal{U}) + \sum_{a\in\mathcal{A}} \mathrm{H}_{\nu^*}(U_a) - \mathrm{H}_{\nu^*}(\mathcal{U}) \Big) + \mathrm{H}_{\mu}(\mathcal{X}) - \mathrm{H}_{\nu^*}(\mathcal{U}) + \sum_{a\in\mathcal{A}} \mathrm{H}_{\nu^*}(U_a)\\
&= \mathrm{H}_{\mu}(\mathcal{X}) - \mathrm{I}_{\nu^*}(\mathcal{X};\mathcal{U})\\
&\ge 0.
\end{aligned}$$

The final inequality holds because of our assumption that the marginal $\nu^*(\mathcal{X})$ equals $\mu(\mathcal{X})$. Thus, $\mathrm{QIM}\mathit{Inc}_{\mathcal{A}}(\mu) \ge \mathit{IDef}_{\mathcal{A}}(\mu)$, as proimised.

We now turn to the right hand inequality, and part (b) of the theorem. Recall that ν^* is defined to be a minimizer of the optimization problem defining $\mathrm{QIM}\mathit{Inc}$. For the right inequality $(\mathrm{QIM}\mathit{Inc}_{\mathcal{A}}(\mu) \le \mathit{IDef}_{\mathcal{A}^\dagger}(\nu))$ of part (a), observe that

$$\begin{aligned}
\mathit{IDef}_{\mathcal{A}^\dagger}(\nu) &= -\mathrm{H}_{\nu}(\mathcal{X},\mathcal{U}) + \sum_{a\in\mathcal{A}} \mathrm{H}_{\nu}(U_a) + \sum_{a\in\mathcal{A}} \mathrm{H}_{\nu}(T_a|S_a, U_a) + \mathrm{H}_{\nu}(\mathcal{X} \mid \mathcal{U})\\
&= \Big(-\mathrm{H}_{\nu}(\mathcal{U}) + \sum_{a\in\mathcal{A}} \mathrm{H}_{\nu}(U_a) \Big) + \sum_{a\in\mathcal{A}} \mathrm{H}_{\nu}(T_a|S_a, U_a)
\end{aligned}$$

$$
\begin{aligned}
&\geq \Big(-\mathrm{H}_{\nu^*}(\mathcal{U}) + \sum_{a\in\mathcal{A}} \mathrm{H}_{\nu^*}(U_a) \Big) + \sum_{a\in\mathcal{A}} \mathrm{H}_{\nu^*}(T_a|S_a, U_a) \\
&= \mathrm{QIM}\mathit{Inc}(\mu).
\end{aligned}
$$

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.

□

4.B SIM-Equivalence

[now that we know this version of Theorem 2 is false, much of this discussion doesn't make sense.] Applying Theorem 4.2.2 with $\mathcal{A}' = \emptyset$ yeilds a quintessential special case: $\mu \models \boxed{X}\rightrightarrows\boxed{Y}$ iff Y is a function of X according to μ. At first glance, this already seems to capture the essence of Theorem 4.2.2; is it really meaningfully weaker? In fact it is; to illustrate, our next example is another graph that behaves the same way—but not in all contexts.

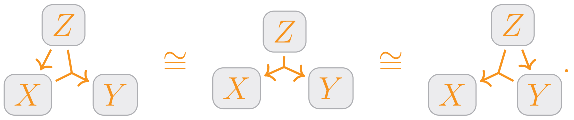
(unproven!)

Example 12. In the appendix, we prove $\mu \models \boxed{X}\!\rightarrow\!\boxed{Y}\!\leftarrow$ iff Y is a function of X (according to μ). But, in general, this graph says something distinct from (and stronger than, as we will see in Section 4.C) the example above. After an adding the hyperarc $\emptyset \to \{X\}$ to both graphs, for example, they behave differently: every distribution μ satisfying $X \twoheadrightarrow Y$ also satisfies $\mu \models \to\boxed{X}\rightrightarrows\boxed{Y}$, but only when Y is a constant can it be the case that $\mu \models \to\boxed{X}\!\rightarrow\!\boxed{Y}\!\leftarrow$. □

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 *(structurally) equivalent* ($\mathcal{A} \cong \mathcal{A}'$) if for every context $\mathcal{A}''$ and distribution μ, 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 4.B.1. *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 4.B.1 provides a simple and useful way to relate QIM-compatibility of different hypergraphs. If we restrict to acyclic structures, for instance, we find:

Theorem 4.B.2 (Chickering 4). *Any two qualitative Bayesian Networks that represent the same independencies can be proven equivalent using only instances of* Proposition 4.B.1 *(in which X, Y, Z may be sets of variables).*

Theorem 4.B.2 is essentially a restatement of main result of Chickering [4], 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 $X \to Y$ to be *covered* in a graph G iff $\mathbf{Pa}_G(Y) = \mathbf{Pa}_G(X) \cup \{X\}$; 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 4.B.2. Theorem 4.B.2 is one demonstration of the usefulness of Proposition 4.B.1, 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 Section 4.C, which is an analogue of implication.

4.C Monotonicity and Undirected Graphical Models

Monotonicity of PDG inconsistency [31] is a powerful reasoning principle. Many important inequalities (e.g., the data processing inequality, relationships between statistical distances, the evidence lower bound, ...) can be proved using only a simple inference rule: "more beliefs can only increase inconsistency". In this section, we develop and apply an anlogous principle for QIM-compatibility. But first, we start with something simple. The fact that (quantitative) PDG inconsistency is monotonic is a powerful reasoning principle that can be used to prove many important inequalities [31]. In this section, we develop a related principle for QIM-compatibility. 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 a nice property: learning an additional formula ϕ_{n+1} can only narrow the set of worlds one considers possible. The same is true of QIM-compatibility.

Proposition 4.C.1. *If $\mathcal{A} \subseteq \mathcal{A}'$ and $\mu \models \Diamond\mathcal{A}'$, then $\mu \models \Diamond\mathcal{A}$.*

Here is a direct but not very useful analague: if $\mathcal{A} \subseteq \mathcal{A}'$ and $\mu \models \Diamond\mathcal{A}'$, conclude $\mu \models \Diamond\mathcal{A}$. After all, if μ 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.

$$\begin{tikzcd}
A \arrow[rr, "f"] & {} \arrow[d, squiggly] & B \arrow[d, hook] \\
A' \arrow[u, hook] \arrow[rr, dashed] & {} & B'
\end{tikzcd}$$

Since we use *directed* hypergraphs, there is actually a finer notion of monotonicity at play. Inputs and ouputs 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 $f : A \to B$ can be regarded as one of type $A' \to B'$. This is depicted to the right. The same principile applies in our setting. If $\mathbf{X}$ and $\mathbf{Z}$ are sets of variables and $\mathbf{X} \subseteq \mathbf{Z}$, then $\mathcal{V}(\mathbf{Z}) \hookrightarrow \mathcal{V}(\mathbf{X})$, 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.

Definition 4.C.1. If $\mathcal{A} = \{S \xrightarrow{a} T\}_a$, $\mathcal{A}' = \{S' \xrightarrow{a'} T'\}_{a'}$, and there is an injective map $\iota : \mathcal{A}' \to \mathcal{A}$ 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 $\mathcal{A} \rightsquigarrow \mathcal{A}'$). □

QIM-compatibility is monotonic with respect to weakening ($\rightsquigarrow$).

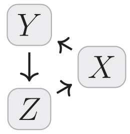
Theorem 4.C.2. *If $\mathcal{A} \rightsquigarrow \mathcal{A}'$ and $\mu \models \Diamond \mathcal{A}$, then $\mu \models \Diamond \mathcal{A}'$.*

Theorem 4.C.2 is strictly stronger than Proposition 4.C.1 because a hyperarc with no targets is vacuous, so removing all targets of a hyperarc is equivalent to deleting it.

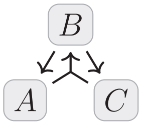
Theorem 4.C.2 explains why BNs and MRFs are arguably *anti*-monotonic: adding $X \to Y$ to a graph G means adding X to the *sources* the hyperarc whose target is Y, in $\mathcal{A}_G$.

As mentioned in the main body of the paper, 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 8. Every $\mu(X, Y)$ is compatible with $\boxed{X}\rightleftarrows\boxed{Y}$. This is because this cycle is weaker than a hypergraph that can already represent any distribution, i.e., $\to\boxed{X}\to\boxed{Y} \quad \rightsquigarrow \quad \boxed{X}\rightleftarrows\boxed{Y}$. △.



Example 9. 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 $\to X \to Y \to Z$. 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 4.4. △



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 (cylic) pair of directed edges $\{A \to B, B \to A\}$, as we have implicitly done in defining $\mathcal{A}_G$. In this way, undirected graphs, too, naturally correspond to directed hypergraph s. 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 [20]. This may appear to be a flaw in Definition 4.2.2 (QIM-compatibility), but it is unavoidable. While both BNs and MRFs are monotonic, it is impossible to capture both classes with a monotonic definition.

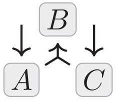
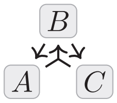
Theorem 4.C.3. *It is possible to define a relation $\overset{\bullet}{\models}$ between distributions μ and directed hypergraph s $\mathcal{A}$ satisfying any two, but not all three, of the following.*

(monotonicity) *If $\mu \overset{\bullet}{\models} \mathcal{A}$ and $\mathcal{A} \rightsquigarrow \mathcal{A}'$, then $\mu \overset{\bullet}{\models} \mathcal{A}'$.*

(positive BN capture) *If μ satisfies the independencies $\mathcal{I}(G)$ of a dag G, then $\mu \overset{\bullet}{\models} \mathcal{A}_G$.*

(negative MRF capture) *If $\mu \overset{\bullet}{\models} \mathcal{A}_G$ for an undirected directed graph G, then μ has one of the Markov properties with respect to G.*

The proof is a direct and easy-to-visualize application of monotonicity (Theorem 4.C.2). Assume montonicity 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:

$\mu_{xor} \models$  $\rightsquigarrow$  $= \mathcal{A}_{A-B-C}$. But $\mu_{xor} \not\models A \perp\!\!\!\perp C \mid B$. □

We emphasize that Theorem 4.C.3 has implications for the qualitative semantics of *any* graphical model (even if one were to reject the definition QIM-compatibility). We now look into the implications for some lesser-known graphical models, which may appear not to comply with Theorem 4.C.3.

Dependecny Networks To readers familiar with *dependency networks (DNs)* [12], Theorem 4.C.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.

Theorem 4.C.3 shows that QIM-compatibility does not capture MRFs (at least, in the obvious way) at a purely observational level. Nevertheless, there is still a sense in which QIM-compatibility captures MRFs *causally*—that is, if we *intervene* instead of conditioning.

[link to proof]

Proposition 4.C.4. *Let G be an undirected graph whose vertices correspond to variables $\mathcal{X}$.*

1. *Let $\mu(\mathcal{X})$ be a positive distribution (i.e., $\forall \mathbf{x} \in \mathcal{V}(\mathcal{X}).\ \mu(\mathcal{X}{=}\mathbf{x}) > 0$). If $\mu \models \mathcal{A}_G$, then for every witness $\bar{\mu}$ and causal model $\mathcal{M} \in \mathrm{PSEMs}_{\mathcal{A}_G}(\bar{\mu})$, whenever $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathcal{X}$ are such that $\mathbf{X} \perp\!\!\!\perp_G \mathbf{Y} \mid \mathbf{Z}$, it is the case that $\bar{\mu} \models \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathrm{do}_{\mathcal{M}}(\mathbf{Z} = \mathbf{z})$.*
2. *Convesely, there exists some distribution $\mu(\mathcal{X})$ If $\mu \not\models \mathcal{A}_G$, then*

4.D Information Theory, PDGs, and QIM-Compatibility

4.D.1 More Detailed Primer on Information Theory

We now expand on the fundemental information quantities introduced at the beginning of Section 4.4. Let μ be a probability distribution, and be X, Y, Z be (sets of) discrete random variables. The *entropy* of X is the uncertainty in X, when it is distributed according to μ, as measured by the number of bits of information needed (in expectation) needed to determine it, if the distribution μ is known. It is given by

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

and a few very important properties; chief among them, $\mathrm{H}_\mu(X)$ is non-negative, and equal to zero iff X is a constant according to μ. The "joint entropy" $\mathrm{H}(X, Y)$ 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:

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

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

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

The former is non-negative and equal to zero iff $\mu \models X \perp\!\!\!\perp Y$, and the latter is non-negative and equal to zero iff $\mu \models X \perp\!\!\!\perp Y \mid Z$. All of these quantities are purely "structural" or "qualitative" in the sense that they are invariant to relabelings of values, and

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: $\mathrm{I}(X;Y|Z) = \mathrm{I}(X;(Y,Z)) - \mathrm{I}(X;Z)$. 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, $\mathrm{H}_\mu(X)$ is clearly the special case of $\mathrm{H}(X|Z)$ when Z is a constant (e.g., $Z = \emptyset$). Furthermore, entropy and mutual information are also interdefinable and generated by linear combinations of one another. It is easy to verify that $\mathrm{I}_\mu(X;Y) = \mathrm{H}_\mu(X) + \mathrm{H}_\mu(Y) - \mathrm{H}(X,Y)$ and $\mathrm{I}_\mu(X;Y|Z) = \mathrm{H}_\mu(X|Z) + \mathrm{H}_\mu(Y|Z) - \mathrm{H}(X,Y|Z)$, and thus mutual information is derived from entropy. Yet on the other hand, $\mathrm{I}_\mu(Y;Y) = \mathrm{H}_\mu(Y)$ and $\mathrm{I}_\mu(Y;Y|X) = \mathrm{H}_\mu(Y|X)$—thus entropy is a special case of mutual information.

4.D.2 Structural Deficiency: More Motivation, and Examples

To build intuition for IDef, which characterizes our bounds in Section 4.4, we now visualize the vector $\mathbf{v}_\mathcal{A}$ for various example hypergraphs.

- Subfigures 4.D.1a, 4.D.1b, and 4.D.1c show how adding hyperarcs makes distriutions more deterministic. When $\mathcal{A}$ is the empty hypergraph, IDef reduces to negative entropy, and so prefers distributions that are "maximally uncertain" (e.g., Subfigures 4.D.1a and 4.D.1d). For this empty but all distributions μ have negative $\mathit{IDef}_\mathcal{A}(\mu) \leq 0$. In the definition of IDef, each hyperarc $X \to Y$ is compiled to a "cost" $H(Y|X)$ for uncertainty in Y given X. One can see this visually in Figure 4.D.1 as a red crescent that's added to the information profile as we move from 4.D.1d to 4.D.1e to 4.D.1f.
- Some hypergraphs (see Figures 4.D.1b and 4.D.1h) are *indiscriminate*, in the sense that every distribution gets the same score (of zero, because a point mass δ always has $\mathit{SDef}_\mathcal{A}(\delta) = 0$). Such a graph has a structure such that *any* distribution can be precisely encoded by the process in (b). As

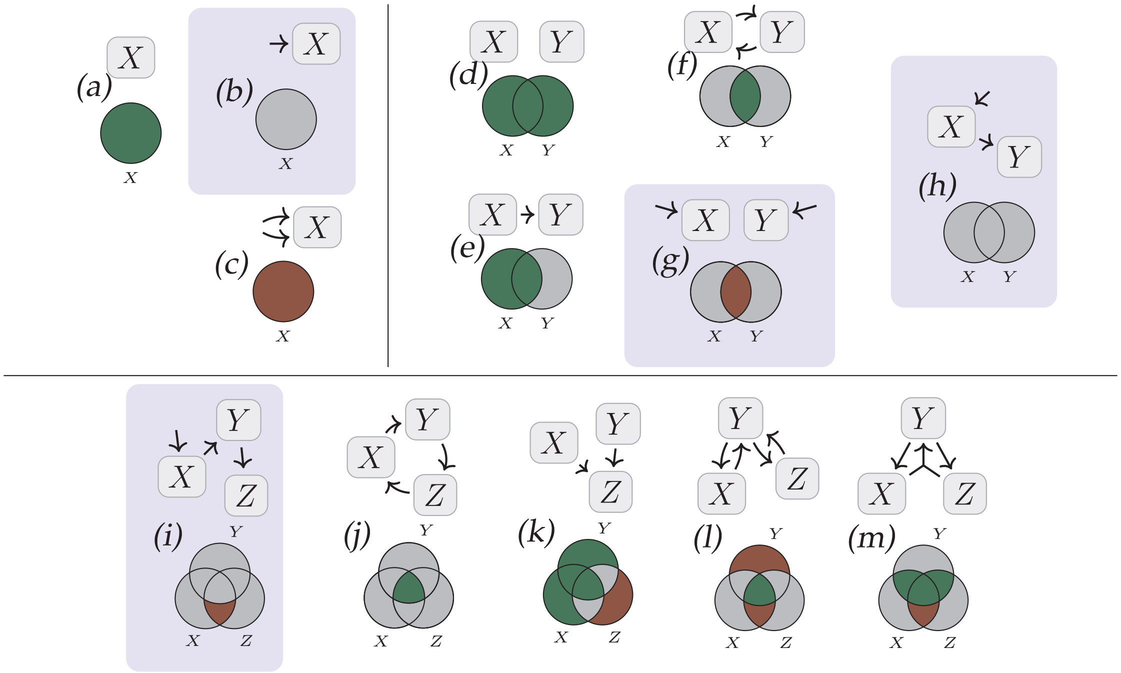


Figure 4.D.1: Illustrations of the structural deficiency $IDef_{\mathcal{A}}$ underneath drawn underneath their associated hypergraphs $\{G_i\}$. Each circle represents a variable; an area in the intersection of circles $\{C_j\}$ but outside of circles $\{D_k\}$ corresponds to information that is shared between all C_j's, but not in any D_k. Variation of a candidate distribution μ in a green area makes its qualitative fit better (according to $IDef$), while variation in a red area makes its qualitative fit worse; grey is neutral. Only the boxed structures in blue, whose $IDef$ can be seen as measuring distance to a particular set of (conditional) independencies, are expressible as BNs.

shown here and also in Richardson and Halpern [32], $IDef$ can also indicate independencies and conditional independencies, illustrated respectively in Subfigures 4.D.1g and 4.D.1i.

- For more complex structures, structural information deficiency $IDef$ can represent more than independence and dependence. The cyclic structures in Examples 8 and 9, correspond to the structural deficiencies pictured in Subfigures 4.D.1f and 4.D.1j, respectively, which are functions that encourage shared information between the three variables.

4.D.3 Weights for SIM-Inc

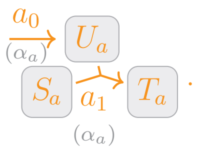
Given $\mathcal{A}$ and $|\mathcal{A}|+2$ positive weights $\boldsymbol{\lambda} = (\lambda^{(\mathrm{a})}, \lambda^{(\mathrm{b})}, \{\lambda^{(\mathrm{c})}_a\}_{a\in\mathcal{A}})$, define the function

$$\mathrm{QIM}\mathit{Inc}_{\mathcal{A},\boldsymbol{\lambda}}(\mu) := \inf_{\nu(\mathcal{U},\mathcal{X})} \left\{ \begin{aligned} &\lambda^{(\mathrm{a})} \boldsymbol{D}\big(\nu(\mathcal{X}) \,\big\|\, \mu(\mathcal{X})\big) \\ +\;& \lambda^{(\mathrm{b})} \Big(-\mathrm{H}_\nu(\mathcal{U}) + \sum_{a\in\mathcal{A}} \mathrm{H}_\nu(U_a) \Big) \\ +\;& \sum_{a\in\mathcal{A}} \lambda^{(\mathrm{c})}_a \,\mathrm{H}_\nu(T_a | S_a, U_a) \end{aligned} \right. . \tag{4.6}$$

When $\boldsymbol{\lambda} = (\infty, 1, \mathbf{1})$, we get the analogous quantity defined in (4.3) in the main text.

Here are some analogous results for this generalized version with weights. For a weighted hypergraph $(\mathcal{A}, \boldsymbol{\alpha})$, here is a strengthening of Theorem 4.4.3, and the appropriate translateion of the hypergraph. Given $(\mathcal{A}, \boldsymbol{\alpha})$ translate it to a new derandomized hypergraph $(\mathcal{A}, \boldsymbol{\alpha})^\dagger$ by replacing each weighted hyperarc

$\boxed{S_a} \xrightarrow[(\alpha_a)]{a} \boxed{T_a}$ with the pair of weighted hyperarcs



.

⟨ INCOMPLETE ⟩

4.D.4 Counter-Examples to the Converse of Theorem 4.4.1

In light of Example 11 and its connections to *SDef* through Theorem 4.4.1, 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 4.4.1 is false.

Example 13. Suppose $\mu(X, Y, Z) = \text{Unif}(X, Z)\delta\text{id}(Y|X)$ and $\mathcal{A} = \{\to X, \to Y\}$, where all variables are binary. Then $\mathit{SDef}_{\mathcal{A}}(\mu) = 0$, but X and Y are not independent. □

Here is another counter-example, of a very different kind.

Example 14. 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 of $\mathcal{V}(A, B, C)$ can be QIM-compatible with the 3-cycle $\mathcal{A}_{3\circ}$. Yet it is easy to find examples of such distributions μ that have positive interaction information $\mathrm{I}(A; B; C)$, and thus $\mathit{SDef}_{\mu}(\mathcal{A}_{3\circ}) \leq 0$ for such distributions. □

4.E QIM-Compatibility Constructions and Counterexamples

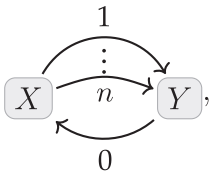
We now give a counterexample to a simpler previously conjectured strengthening of Theorem 4.2.2, in which part (a) is an if-and-only-if. This may be surprising. In the unconditional case, it is true that, two arcs $\{\xrightarrow{1} X, \xrightarrow{2} X\}$ precisely encode that X is a constant, as illustrated by Example 7. The following, slightly more general result, is an immediate correlary of Theorem 4.2.2(c).

Proposition 4.E.1. $\mu \models \mathcal{A} \sqcup \{\xrightarrow{1} X, \xrightarrow{2} X\}$ *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 μ is even) are enough to ensure this. That is, there are partitional graphs $\mathcal{A}$ such that $\mu \overset{e}{\models} \mathcal{A}$ but $\mu \not\models \mathcal{A} \sqcup \{X \xrightarrow{1} Y, X \xrightarrow{2} Y\}$.

Example 15. We will construct a witness of SIM-compatibility for the hypergraph

$\mathcal{A} :=$ 

,

in which Y is *not* a function of X, which for $n = 3$ will disprove the analogue of Theorem 4.2.2 for the partitional context $\mathcal{A}'$ equal to the 2-cycle.

Let $\mathcal{U} = (U_0, U_1, \ldots, U_n)$ be a vector of n mutually independent random coins, and A is one more independent random coin. For notational convenience, define the random vector $\mathbf{U} := (U_0, \ldots, U_n)$ consisting of all variables U_i except for U_0. Then, define variables X and Y according to:

$$\begin{aligned} X &:= (A \oplus U_1, \ldots, A \oplus U_n, \;\; U_0 \oplus U_1, U_0 \oplus U_2, \ldots, U_0 \oplus U_n) \\ &= (A \oplus \mathbf{U}, \;\; U_0 \oplus \mathbf{U}) \\ Y &:= (A, U_0 \oplus \mathbf{U}) = (A, \; U_0 \oplus U_1, U_0 \oplus U_2, \ldots, U_0 \oplus U_n), \end{aligned}$$

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 $\nu(X, Y, \mathcal{U})$.

It we now show that ν 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:

$$\begin{aligned} g(A, \mathbf{B}, U_0) &= (A \oplus U_0 \oplus \mathbf{B},\ \mathbf{B}) \\ &= (A \oplus U_0 \oplus U_0 \oplus \mathbf{U},\ U_0 \oplus \mathbf{U}) \qquad \text{since } \mathbf{B} = U_0 \oplus \mathbf{U} \\ &= (A \oplus \mathbf{U}, U_0 \oplus \mathbf{U}) = X \end{aligned}$$

(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 $\mu(X, Y) \models^{e} \mathcal{A}$.

Yet, 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 those who are interested, observe that the bound of Theorem 4.4.1 tells that we must satisfy

$$\begin{aligned} 0 \geq \mathit{IDef}_{\mathcal{A}}(\mu) &= -\,\mathrm{H}_\mu(X, Y) + n\,\mathrm{H}_\mu(Y \mid X) + \mathrm{H}_\mu(X \mid Y) \\ &= -\,\mathrm{I}_\mu(X; Y) + (n-1)\,\mathrm{H}_\mu(Y \mid X) \end{aligned}$$

Indeed, this distribution has information profile

$$\mathrm{H}(X \mid Y) = 1\,\text{bit}, \qquad \mathrm{I}(X; Y) = n\,\text{bits}, \qquad \mathrm{H}(Y \mid X) = 1\,\text{bit},$$

and so $\mathit{IDef}_{\mathcal{A}}(\mu) = -1\,\text{bit}$. Intuitively, this one missing bit corresponds to the value of A that is not determined by the structure of $\mathcal{A}$. □

4.F 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 4.3.2. Given a (generalized) randomized PSEM $\mathcal{M}$, we now show that distributions $\nu \in \{\!\!\{\mathcal{M}\}\!\!\}$, are QIM-compatibility witness showing that the marginals of ν are QIM-compatible with the hypergraph $\mathcal{A}_{\mathcal{M}}$. More formally:

Proposition 4.F.1. *If $\mathcal{M} = (M{=}(\mathcal{U}, \mathcal{V}, \mathcal{F}), P)$ 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 ν. Finally, (b) is also satisfied by construction: we assumed that $\mathcal{U}_{\mathcal{A}} = \{U_a\}_{a \in \mathcal{A}}$ are independent.

4.G An Algorithm for Finding Witnesses: The Null Value Construction

We have now built up a body of examples, but it is still not clear how to compute QIM-compatibility. In other words, it is still not clear how to solve the decision problem: given μ and $\mathcal{A}$, determine whether or not $\mu \models \mathcal{A}$. In this section, we discuss one approach to this problem.

If you start with a distribution $\nu(\mathcal{X})$, it's not at all obvious how to extend it with a conditional distribution $\nu(\mathcal{U}|\mathcal{X})$ such that the variables $\mathcal{U}$ are *unconditionally* dependent, given that they cannot be independent of $\mathcal{X}$. It seems that the only way to ensure this unconditional independence is to start with a distribution $\nu(\mathcal{U}) = \prod_a \nu(U_a)$ and then figure out how to extend it to the variables $\mathcal{X}$.

To begin, for each $a \in \mathcal{A}$, take U_a to be a response variable, taking values $\mathcal{V}(U_a) = (\mathcal{V}T_a)^{(\mathcal{V}S_a)}$, just as in Section 4.3. But now we run into a problem: without carefully selecting the supports of the distributions over $\mathcal{U}$, it is entirely possible that there will be some joint setting $\mathbf{u} \in \mathcal{V}\mathcal{U}$ occurs with positive problability, but represents a collection of functions that has no fixed point. For example, take the graph

Example 16 (5, continued)**.** By randomly selecting distributions $\Pr(U_1), \Pr(U_2)$, and $\Pr(U_3)$ (see Section 4.G), one finds that the set of distributions that are consistent with this 3-cycle has larger dimension than the set of distributions that factorize according to $\Pr(X,Y,Z) \propto \phi_1(X,Y)\phi_2(Y,Z)\phi_3(Z,X)$.[8] Thus, our definition does not coincide with the corresponding factor graph. □

Conjecture 4.G.1. *If $\mu_0 \models \mathcal{A}$, and μ' lies on the path $\mu(t)$ of gradient flow minimizing SDef$_{\mathcal{A}}(\mu')$, starting at $\mu(0) = \mu_0$, then $\mu' \models \mathcal{A}$.*

The following has emperical support.

Conjecture 4.G.2. *The distribution (s?) $\hat\mu := \arg\min_{\mu:\mu\models\mathcal{A}} \boldsymbol{D}(\mu \parallel \hat\mu)$ have the same information profile as μ.*

[8] see appendix for details.

4.H Even Structural Compatibility

4.H.1 Even QIM-Compatibility

If $\mathcal{M}$ is a cyclic or subpartitional PSEM, then $\{\!\!\{\mathcal{M}\}\!\!\}$ may contain many distributions. Still, so long as it is non-empty, there is still a unique distribution that, arguably, best describes the distribution of the PSEM (in the absence of interventions)—namely, the one that, for any given value $\mathbf{u} \in \mathcal{V}(\mathcal{U})$, treats all "fixed-points" $\mathbf{x} \in \mathcal{F}(\mathbf{u})$ of the equations $\mathcal{F}$ symmetrically.

$$\Big(\text{Recall that} \qquad \mathcal{F}(\mathbf{u}) := \{\mathbf{x} \in \mathcal{V}(\mathcal{X}) : \forall a \in \mathcal{A}. f_a(S_a(\mathbf{x}), u_a) = S_a(\mathbf{x})\}.\Big)$$

For example, if $\mathcal{M}$ has no exogenous variables ($\mathcal{U} = \emptyset$), endogenous variables $\mathcal{X} = [X_1, \ldots, X_n]$ that are all binary, and equations $f_{X_i}(\mathcal{X} \setminus X_i) = X_{(i+1)\%n}$, then $\{\!\!\{\mathcal{M}\}\!\!\}$ is a 1-dimensional specturm of distributions supported on the two points $\{(0, \ldots, 0), (1, \ldots, 1)\}$. The distribution that gives the two an equal weight of $\frac{1}{2}$ is somehow special, in that it is the unique one that does not break the symmetry by preferring either $(0, \ldots, 0)$ or $(1, \ldots, 1)$. This intuition is made precise, and generalized to QIM-compatibility witnesses (rather than PSEMs), by the following definition.

Definition 4.H.1. We say a witness $\nu(\mathcal{U}, \mathcal{X})$ to $\mu \models \mathcal{A}$ is *even*, iff it satisfies properties (a,b) of Definition 4.2.2, and also the following strengthening of property (c):

(d) $\nu(\mathcal{X} \mid \mathcal{U}) \propto \mathbb{1}\Big[\bigwedge_{a \in \mathcal{A}} f_a(S_a, U_a) = T_a\Big],$

for some set $\mathcal{F} = \{f_a : \mathcal{V}(S_a, U_a) \to \mathcal{V}(T_a)\}_{a \in \mathcal{A}}$ of equations. In this case, we say μ is *evenly* QIM-compatible (EQIM-compatible) with $\mathcal{A}$, write $\mu \overset{e}{\models} \mathcal{A}$, and call the pair $(\nu, \mathcal{F})$ a witness of EQIM-compatibility. □

EQIM-compatibility clearly implies QIM-compatibility, and thus is a stricter notion. Furthermore, EQIM-compatibility witnesses have an even sharper relationship to causal models: A witness $(\bar{\mu}(\mathcal{U}, \mathcal{X}), \mathcal{F})$ to EQIM-compatibility, can be equivalently viewed as a PSEM $\mathcal{M} = (\mathcal{U}, \mathcal{X}, \mathcal{F}, \nu(\mathcal{U}))$, because the rest of the distribution $\nu(\mathcal{X} \mid \mathcal{U})$ is determined by $\mathcal{F}$ and property (d).

Proposition 4.H.1. • *There is a 1-1 correspondence between EQIM-compatibility witnesses and GRPSEMs $\mathcal{M}$ in which $\{\!\!\{\mathcal{M}\}\!\!\} \neq \emptyset$.*

Proof. 1. Given a EQIM-compatibility witness $(\nu(\mathcal{X}, \mathcal{U}_{\mathcal{A}}), \mathcal{F})$,

by Proposition 4.3.2 $\mathrm{PSEMs}_{\mathcal{A}}(\nu)$ □

Thus, PSEMs are in direct 1-1 correspondence with EQIM-compatibility witnesses when hypergraphs $\mathcal{A} = \mathcal{A}_G$ for some graph G.

We now verify that various results from the main text extend to EQIM-compatibility.

[Theorem 4.2.1] When G is acyclic (and, more generally, when $\mathcal{M} \models \mathcal{U} \twoheadrightarrow \mathcal{X}$), the extra condition (d) holds trivially, and so EQIM-compatibility coincides with QIM-compatibility, and the analogue of Theorem 4.2.1 in which $\models$ is replaced with $\overset{e}{\models}$ also holds.

⋮

[Theorem 4.C.2] As we show in the the proof of Theorem 4.C.2, EQIM-compatibility is also monotonic with respect to weakening ($\rightsquigarrow$).

The original scoring function IDef is related to even QIM-compatibility.

Conjecture 4.H.2. *There is a constant $\kappa = \kappa(\mathcal{A}, \mathcal{V})$ depending on the hypergraph $\mathcal{A}$ and the possible values $\mathcal{V}$ that the variables can take, such that*

$$\mathit{IDef}_{\mathcal{A}}(\mu) \le \kappa \qquad \iff \qquad \mu \overset{e}{\models} \mathcal{A}$$

Proof. **($\Longleftarrow$).** Suppose $\mu \overset{e}{\models} \mathcal{A}$. Then there is some witness $\bar{\mu}$ extending μ to independent variables $\mathcal{U}_{\mathcal{A}} = \{U_a\}_{a \in \mathcal{A}}$.

($\Longrightarrow$). Suppose that $\mathit{IDef}_{\mathcal{A}}(\mu) \le \kappa$. □

4.H.2 ESIM Compatibility Scoring Rules

We have now seen that, $\mathit{IDef}_{\mathcal{A}^\dagger}$ measures distance from being a witness to QIM-compatibility (Theorem 4.4.3(b)). Modulo a constant offset or limit, $\mathit{IDef}_{\mathcal{A}}$, i.e., the original scoring function applied to the original hypergraph

Let's now repeat the same approach as the previous section, by explicitly constructing a scoring function for EQIM-compatibility. Extend our previous weight vector by one entry, so that $\boldsymbol{\lambda} = (\lambda^{(\mathrm{a})}, \lambda^{(\mathrm{b})}, \{\lambda^{(\mathrm{c})}_a\}_{a \in \mathcal{A}}, \lambda^{(\mathrm{d})})$.

$$\begin{aligned}\mathrm{EQIM}\mathit{Inc}_{\mathcal{A},\boldsymbol{\lambda}}(\mu) := \inf_{\nu(\mathcal{X},\mathcal{U})} & \ \blacksquare \\ & + \lambda^{(\mathrm{d})} \inf_{\mathcal{F}} \mathop{\mathbb{E}}_{\mathbf{u} \sim \nu(\mathcal{U})} \Big[\boldsymbol{D}\big(\nu(\mathcal{X} \mid \mathcal{U} = \mathbf{u}) \,\big\|\, \mathrm{Unif}[\mathcal{F}(\mathbf{u})]\big) \Big],\end{aligned}$$

where $\blacksquare$ consists of everything but the infemum from Equation (4.3), $\mathcal{F}$ ranges over sets of equations along $\mathcal{A}$ (as in Definitions 4.3.1 and 4.H.1), and $\mathrm{Unif}[\mathcal{F}(\mathbf{u})]$ is the uniform distribution over joint settings of $\mathcal{X}$ that are consistent with the equations after fixing context $\mathcal{U} = \mathbf{u}$. Recall that the key step of constructing $\mathcal{A}^\dagger$

was to add the hyperarc $\mathcal{U} \to \mathcal{X}$. But for even compatbility, we want to effectively do the opposite—that is, subject to satisfying the other constraints, we want to incentivize, rather than penalize the conditional entropy $\mathrm{H}(\mathcal{X}|\mathcal{U})$. This is made precise by the following proposition.

Proposition 4.H.3. (a) *For all* $\boldsymbol{\lambda} > \mathbf{0}$, $\mathrm{EQIM}\mathit{Inc}_{\mathcal{A},\boldsymbol{\lambda}}(\mu) \geq 0$ *with equality iff* $\mu \overset{e}{\models} \mathcal{A}$.

In other words, *IDef* itself already measures distance from *even QIM-compatibility*, once we find the appropriate constant to make it non-negative. Although has an enormous benefit of not requiring an infemum.

$\langle$ `TODO: There's an issue here I still need to finish working out.` $\rangle$

$$\mathit{IDef}_{\mathcal{A}}(\mu) \leq \mathrm{QIM}\mathit{Inc}_{\mathcal{A}}(\mu) \leq \mathit{IDef}_{\mathcal{A}^\dagger}(\nu^*) \leq \mathit{IDef}_{\mathcal{A}}(\mu) + \kappa().$$

$\kappa_{\mathcal{A}}$ is a (possibly infinite) piecewise constant function of μ with finitely many pieces, and finite when $\mu \models \mathcal{A}$.

4.H.3 Complete Derandomization for Cyclic Models

We have seen that a number of properties of causal models are simpler when $\mathcal{M} \models \mathcal{U} \twoheadrightarrow \mathcal{X}$. In some sense, the job of a causal model is model $\mathcal{X}$ by adding variables $\mathcal{U}$ that account for any uncertainty. When $\mathcal{M} \not\models \mathcal{U} \twoheadrightarrow \mathcal{X}$, this job is in some sense incomplete; cycles can create a new source of uncertainty. In this subsection, we explore the effects of adding one more variable U_0 to account for the remaining uncertainty, by explaining it as randomenss. Technically speaking, this means looking into one way of converting a randomized PSEM $\mathcal{M}$ to one in which $\mathcal{U} \twoheadrightarrow \mathcal{X}$.

Our construction is parameterized by a PSEM $\mathcal{M}$ and a choice of $\nu \in \{\!\{\mathcal{M}\}\!\}$. In brief, we use a natural construction explained in the next subsection (4.A.1) to obtain a "maximally independent" derandomization of $\nu(\mathcal{X} \mid \mathcal{U})$. The result is a new generalized randomized PSEM, which we call $\mathcal{M}^{+\tilde{\nu}(\mathcal{X}^{\mathcal{U}})}$, differing from $\mathcal{M}$ in that it has one extra hyperarc $\mathcal{U} \to \mathcal{X}$, and, correspondingly, an extra variable U_0, and an extra equation $f_0 : \mathcal{U} \to \mathcal{X}$. This new causal model has two important properties:

1. Only ν can arise from $\mathcal{M}^{+\tilde{\nu}(\mathcal{X}^{\mathcal{U}})}$ (i.e., $\{\!\{\mathcal{M}^{+\tilde{\nu}(\mathcal{X}^{\mathcal{U}})}\}\!\} = \{\nu\}$), and
2. the exogenous variables determine the values endogenous ones (i.e., $\mathcal{M}^{+\tilde{\nu}(\mathcal{X}^{\mathcal{U}})} \models (\mathcal{U}, U_0) \twoheadrightarrow \mathcal{X}$).

Constructing the Causal Model $\mathcal{M}^{+\tilde{\nu}(\mathcal{X}^{\mathcal{U}})}$

We now apply the general technique above to obtain the causal model $\mathcal{M}^{+\tilde{\nu}(\mathcal{X}^{\mathcal{U}})}$ discussed in Section 4.H.3. Concretely, let $\mathcal{V}(U_0) := \prod_{\mathbf{u} \in \mathcal{V}(\mathcal{U})} \mathcal{F}(\mathbf{u})$ consist of all functions from $\mathcal{V}(\mathcal{U})$ to $\mathcal{V}(\mathcal{X})$ consistent with the equations $\mathcal{F}$, and add an equation $\mathcal{X} = U_0(\mathbf{u}) = f_0(U_0, \mathbf{u})$ along the hyperarc $\mathcal{U} \to \mathcal{X}$.

Given $\nu \in \{\!\{\mathcal{M}\}\!\}$, define $P(U_0{=}f) := \prod_{\mathbf{u} \in \mathcal{V}(\mathcal{U})} \nu(f(\mathbf{u}) \mid \mathbf{u})$. As shown more generally in **??**,

1. this indeed a probability distribution, and

2. the joint distribution $P(\mathcal{U}, U_0) = P(\mathcal{U})P(U_0)$ extends uniquely along the function defined by U_0, to a distribution $P(\mathcal{U}, U_0, \mathcal{X})$, and the marginal of that distribution on $(\mathcal{U}, \mathcal{X})$ equals ν.

We must also be careful about how to respect interventions. In the most general form, an intervention of the form $f_a \leftarrow g$, for some function $g : \mathcal{V}(S_a) \to \mathcal{V}(T_a)$, not only modifies the equation f_a by setting it equal to g, but also modifies the equation f_0 according to:

$$f_0(\mathbf{u}) := f_0(\mathbf{u})[]$$

⟨ INCOMPLETE ⟩

For instance, if $\mathcal{M}$ is a PSEM, then to perform an intervention $\mathbf{X}{\leftarrow}\mathrm{x}$ on the causal model $\mathcal{M}^{+\tilde{\nu}(\mathcal{X}^{\mathcal{U}})}$, we mean not only to replace the equation fo f_X, but also modify

$$f_0^{\text{new}}(\mathbf{u}) := f_0^{\text{old}}(\mathbf{u})[\mathbf{X} \mapsto \mathbf{x}]$$

⟨ INCOMPLETE ⟩

Proposition 4.H.4. *1.* $\Pr_{\mathcal{M}^{+\tilde{\nu}(\mathcal{X}^{\mathcal{U}})}}(\varphi) = \Pr_{\mathcal{M}}(\varphi)$

Part II

A Universal Objective

CHAPTER 5

LOSS AS THE INCONSISTENCY OF A PDG: CHOOSE YOUR MODEL, NOT YOUR LOSS

RELATIVE ENTROPY SOUP

Oliver E. Richardson, Ph.D.

Cornell University 2024

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. Probabilistic depencency graphs (PDGs) are probabilistic models that come equipped with a measure of “inconsistency”. 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.

5.1 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 pratice. Each implicitly represents different values and results in different behavior, so the choice between them can be quite important [**? ?**]. 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 probablistic 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 [**?**].

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 [17]. 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 $-\mathrm{ELBO}(x)$ 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 inequalitites that traditionally motivate the ELBO, and the monotonicity of Rényi divergence.

In the coming sections, 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.

5.2 Preliminaries

We generally use capital letters for variables, and lower case letters for their values. For variables X and Y, a conditional probability distribution (cpd) p on Y given X, written $p(Y|X)$, consists of a probability distribution on Y (denoted $p(Y|X{=}x)$ or $p(Y|\,x)$ for short), for each possible value x of X. If μ is a probability on outcomes that determine X and Y, then $\mu(X)$ denotes the marginal of μ on X, and $\mu(Y|X)$ denotes the conditional marginal of μ on Y given X. Depending on which we find clearer in context, we write either $\mathbb{E}_\mu f$ or $\mathbb{E}_{\omega\sim\mu} f(\omega)$ for expectation of $f : \Omega \to \mathbb{R}$ over a distribution μ with outcomes Ω. We write $\boldsymbol{D}(\mu \parallel \nu) = \mathbb{E}_\mu \log \frac{\mu}{\nu}$ for the relative entropy (KL Divergence) of ν with respect to μ, we write $\mathrm{H}(\mu) := \mathbb{E}_\mu \log \frac{1}{\mu}$ for the entropy of μ, $\mathrm{H}_\mu(X) := \mathrm{H}(\mu(X))$ for the marginal entropy on a variable X, and $\mathrm{H}_\mu(Y \mid X) := \mathbb{E}_\mu \log {}^1\!/\!_{\mu(Y|X)}$ for the conditional entropy of Y given X.

A *probabilistic dependency graph* (PDG) [**?**], like a Bayesian Network (BN), is a directed graph with cpds attached to it. While this data is attached to the *nodes* of a BN, it is attached to the *edges* of a PDG. For instance, a BN of shape $X \to Y \leftarrow Z$ contains a single cpd $\Pr(Y|X,Z)$ on Y given joint values of X and Z, while a PDG of the same shape contains two cpds $p(Y|X)$ and $q(Y|Z)$. The second approach is strictly more expressive, and can encode joint dependence with an extra variable. All information in a PDG can be expressed with variable confidence. We now restate the formal definition.

Definition 5.2.1. A Probabilistic Dependency Graph (PDG) is a tuple $\mathcal{m} = (\mathcal{N}, \mathcal{A}, \mathcal{V}, \mathbb{P}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, where

- $\mathcal{N}$ is a set of nodes, corresponding to variables;
- $\mathcal{V}$ associates each node $X \in \mathcal{N}$ with a set $\mathcal{V}(X)$ of possible values that the variable X can take;
- $\mathcal{A}$ is a set of labeled edges $\{X \overset{a}{\to} Y\}$, each with a source X and target Y from $\mathcal{N}$;
- $\mathbb{P}$ associates a cpd $p_a(Y|X)$ to each edge $\mathcal{A}LX\!\!\!Y \in \mathcal{A}$;
- $\boldsymbol{\alpha}$ associates to each edge $\mathcal{A}LXY$ a non-negative number α_L representing the modeler's confidence in the functional dependence of Y on X;
- $\boldsymbol{\beta}$ associates to each edge L a number β_L, the modeler's confidence in the reliability of the cpd p_a. □

How should one choose parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$? A choice of $\beta_{\!_L} = 0$ means that the cpd $p_{\!_L}$ is effectively ignored, in the sense that such a PDG is equivalent to one in which the edge is attached to a different cpd $q \neq p_{_L}$. On the other hand, a large value of $\beta_{_L}$ (or ∞) indicates high (or absolute) confidence in the cpd. By default, we suppose $\beta = 1$, which is just a convenient choice of units—what's important are the magnitudes of β relative to one another. The parameter $\boldsymbol{\alpha}$, typically in $[0, 1]$, represents certainty in the causal structure of the graph, and plays only a minor role in this paper.

Like other graphical models, PDGs have semantics in terms of joint distributions μ over all variables. Most directly, a PDG $\mathcal{M}$ determines two scoring functions on joint distributions μ. For the purposes of this paper, the more important of the two is the *incompatibility* of μ with respect to $\mathcal{M}$, which measures the quantitative discrepency between μ and $\mathcal{M}$'s cpds, and is given by

$$\mathit{Inc}_{\mathcal{M}}(\mu) := \sum_{X \xrightarrow{a} Y} \beta_{_L} \mathop{\mathbb{E}}_{x \sim \mu(X)} \boldsymbol{D}\Big(\mu(Y \,|\, x) \,\Big\|\, p_a(Y \,|\, x)\Big). \tag{5.1}$$

Relative entropy $\boldsymbol{D}(\mu \| p)$ measures divergence between μ and p, and can be viewed as the overhead (in extra bits per sample) of using codes optimized for p, when in fact samples are distributed according to μ [24]. But if one uses edges in proportion to the confidence one has in them, then inefficiencies for of high-confidence cpds are compounded, and hence more costly. So $\mathit{Inc}_{\mathcal{M}}(\mu)$ measures the total excess cost of using $\mathcal{M}$'s cpds in proportion to their confidences $\boldsymbol{\beta}$, when worlds are distributed according to μ.

The *inconsistency* of $\mathcal{M}$, denoted $\langle\!\langle \mathcal{M} \rangle\!\rangle$, is the smallest possible incompatibility of $\mathcal{M}$ with any distribution: $\langle\!\langle \mathcal{M} \rangle\!\rangle := \inf_\mu \mathit{Inc}_{\mathcal{M}}(\mu)$. This quantity, which does not depend on $\boldsymbol{\alpha}$, is the primary focus of this paper.

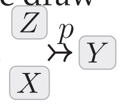
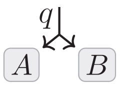
The second scoring function defined by a PDG $\mathcal{M}$, called the *I*nformation *Def*iciency, measures the qualitative discrepency between $\mathcal{M}$ and μ, and is given by

$$IDef_{\mathcal{M}}(\mu) := -\operatorname{H}(\mu) + \sum_{X \xrightarrow{a} Y} \alpha_{L} \operatorname{H}_{\mu}(Y \mid X).$$

$IDef_{\mathcal{M}}(\mu)$ can be thought of as the information needed to separately describe the target of each edge L given the value of its source (weighted by α_{L}) beyond the information needed to fully describe a sample from μ.

As shown by **?**], it is via these two scoring functions that PDGs capture other graphical models. The distribution specified by a BN $\mathcal{B}$ is the unique one that minimizes both $Inc_{\mathcal{B}}$ and $IDef_{\mathcal{B}}$ (and hence every positive linear combination of the two), while the distribution specfied by a factor graph Φ uniquely minimizes the sum $Inc_{\Phi} + IDef_{\Phi}$. In general, for any $\gamma > 0$, one can consider a weighted combination $[\![\mathcal{M}]\!]_{\gamma}(\mu) := Inc_{\mathcal{M}}(\mu) + \gamma\ IDef_{\mathcal{M}}(\mu)$, for which there is a corresponding γ-inconsistency $\langle\!\langle \mathcal{M} \rangle\!\rangle_{\gamma} := \inf_{\mu} [\![\mathcal{M}]\!]_{\gamma}(\mu)$. In the limit as $\gamma \to 0$, there is always a unique best distribution whose score is $\langle\!\langle \mathcal{M} \rangle\!\rangle$.

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 $\boxed{X}\xrightarrow{f}\boxed{Y}$. Definition 5.2.1 is equivalent to one in which edge sources and targets are both *sets* of variables. This allows us to indicate joint dependence with multi-tailed arrows, joint distributions with multi-headed arrows, and unconditional distributions with nothing at the tail. For instance, we draw

$p(Y|X, Z)$ as , and $q(A, B)$ as .

To emphasize that a cpd $f(Y|X)$ is degenerate (a function $f : X \to Y$), we will draw it with two heads, as in: $\boxed{X}\overset{f}{\twoheadrightarrow}\boxed{Y}$. 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 $\overset{x}{\twoheadrightarrow}\boxed{X}$. To specify a confidence $\beta \neq 1$, we place the value near the edge, lightly colored and parenthesized, as in: $\xrightarrow[(\beta)]{p}\boxed{X}$, and we write (∞) for the limit of high confidence $(\beta \to \infty)$.

Intuitively, believing more things can't make you any less inconsistent. Lemma 5.2.1 captures this formally: adding cpds or increasing confidences cannot decrease a PDG's inconsistency.

link to proof

Lemma 5.2.1 (Monotonicity of $\langle\!\langle \cdot \rangle\!\rangle$). *Suppose PDGs $\mathcal{M}$ and $\mathcal{M}'$ differ only in their edges (resp. $\mathcal{A}$ and $\mathcal{A}'$) and confidences (resp. $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$). If $\mathcal{A} \subseteq \mathcal{A}'$ and $\beta_L \leq \beta'_L$ for all $L \in \mathcal{A}$, then $\langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma \leq \langle\!\langle \mathcal{M}' \rangle\!\rangle_\gamma$ for all γ.*[1]

As we will see, this tool is sufficient to derive many interesting relationships between loss functions.

5.3 Standard Metrics as Inconsistencies

Suppose 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 $\mathrm{I}_p[X{=}x] := -\log p(X{=}x)$, or *surprisal* **[?]**, of the event $X{=}x$, according to p.[2] In machine learning, I_p is usually called "negative log likelihood", and is perhaps the most popular objective for training generative models **[? ?]**.

link to proof

[1] All proofs can be found in Section 5.C.

[2] This construction requires the event $X{=}x$ to be measurable. One can get similar, but subtler, results for densities, where this is not the case; see Section 5.A.

Proposition 5.3.1. *Consider a distribution $p(X)$. The inconsistency of the PDG comprising p and $X{=}x$ equals the surprisal $\mathrm{I}_p[X{=}x]$. That is,*

$$\mathrm{I}_p[X{=}x] = \left\langle\!\!\left\langle \xrightarrow{p} \boxed{X} \overset{x}{\twoheadleftarrow} \right\rangle\!\!\right\rangle.$$

(Recall that $\langle\!\langle \mathcal{m} \rangle\!\rangle$ is the inconsistency of the PDG $\mathcal{m}$.)

In some ways, this result is entirely unsurprising, given that (5.1) 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 discrepency 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 more than one sample. What if we replace x with an empirical distribution over many samples?

link to proof

Proposition 5.3.2. *If $p(X)$ is a probabilistic model of X, and $\mathcal{D} = \{x_i\}_{i=1}^m$ is a dataset with empirical distribution $\mathrm{Pr}_\mathcal{D}$, then* $\mathrm{CrossEntropy}(\mathrm{Pr}_\mathcal{D}, p) =$

$$\frac{1}{m}\sum_{i=1}^{m} \mathrm{I}_p[X{=}x_i] = \left\langle\!\!\left\langle \xrightarrow{p} \boxed{X} \xleftarrow[(\infty)]{\mathrm{Pr}_\mathcal{D}} \right\rangle\!\!\right\rangle + \mathrm{H}(\mathrm{Pr}_\mathcal{D}).$$

Remark. The term $H(\mathrm{Pr}_\mathcal{D})$ is a constant depending only on the data, so is irrelevant for optimizing p.

Essentially the only choices we've made in specifying the PDG of Proposition 5.3.2 are the confidences. But $\mathrm{CrossEntropy}(\mathrm{Pr}_\mathcal{D}, p)$ is the expected code length per sample from $\mathrm{Pr}_\mathcal{D}$, when using codes optimized for the (incorrect) distribution p. So implicitly, a modeler using cross-entropy has already articulated a belief the data distribution $\mathrm{Pr}_\mathcal{D}$ is the "true one". To get the same effect from a

PDG, the modeler must make this belief explicit by placing infinite confidence in $\Pr_{\mathcal{D}}$.

Now consider an orthogonal generalization of Proposition 5.3.1, in which the sample x is only a partial observation of (x, z) from a joint model $p(X, Z)$.

link to proof

Proposition 5.3.3. *If $p(X, Z)$ is a joint distribution, then the information content of the partial observation $X = x$ is given by*

$$\mathrm{I}_p[X{=}x] = \left\langle\!\!\!\left\langle \boxed{Z} \;\overset{p}{\swarrow\!\searrow}\; \boxed{X} \twoheadleftarrow \overset{x}{} \right\rangle\!\!\!\right\rangle . \tag{5.2}$$

Intuitively, the inconsistency of the PDG on the right side of (5.2) is localized to X, where the observation x conflicts with $p(X)$; other variables don't make a difference. The multi-sample partial-observation generalization also holds; see Section 5.B.3.

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 [?]. 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.

link to proof

Proposition 5.3.4 (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,*

$$\left\langle\!\!\left\langle \begin{tikzcd} & \mathrm{Pr}_{\mathcal{D}} \arrow[dl] \arrow[dr, "(\infty)"] & \\ \boxed{X} \arrow[rr, "h"'] & & \boxed{Y} \end{tikzcd} \right\rangle\!\!\right\rangle = \frac{1}{m}\sum_{i=1}^{m} \log \frac{1}{h(y_i \,|\, x_i)} - \mathrm{H}_{\mathrm{Pr}_{\mathcal{D}}}(Y|X).$$

Simple evaluation metrics, such as the accuracy of a classifier, and the mean squared error of a regressor, also arise naturally as inconsistencies.

link to proof

Proposition 5.3.5 (Log Accuracy as Inconsistency)**.** *Consider functions $f, h : X \to Y$ 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 β, is β times the log accuracy of h. That is,*

$$\left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "D", "(\beta)"'] & \boxed{X} \arrow[r, bend left, two heads, "h\ (r)"] \arrow[r, bend right, two heads, "f\ (s)"'] & \boxed{Y} \end{tikzcd} \right\rangle\!\!\right\rangle \begin{aligned} &= -\beta \log \Pr_{x\sim D}(f(x)\!=\!h(x)) \\ &= \beta\ \mathrm{I}_D[f = h]. \end{aligned} \tag{5.3}$$

One often speaks of the accuracy of a hypothesis h, leaving the true labels f and empirical distribution D implicit. Yet Proposition 5.3.5 suggests that there is a sense in which $D(X)$ plays the primary role: the inconsistency in (5.3) 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 μ 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 5.3.4: 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.

link to proof

Proposition 5.3.6 (MSE as Inconsistency)**.**

$$\left\langle\!\!\left\langle \begin{tikzcd} & & \mu_f \arrow[dr, "\mathcal{N}_1"] & \\ {} \arrow[r, "D", "(\infty)"'] & X \arrow[ur, two heads, "f"] \arrow[dr, two heads, "h"'] & & Y \\ & & \mu_h \arrow[ur, "\mathcal{N}_1"'] & \end{tikzcd} \right\rangle\!\!\right\rangle = \frac{1}{2}\,\mathbb{E}_D\big|f(X) - h(X)\big|^2 =: \mathrm{MSE}_D(f,h)\,,$$

where $\mathcal{N}_1(Y|\,\mu)$ is a unit Gaussian on Y with mean μ.

In the appendix, we treat general univariate Gaussian predictors, with arbitrary variances and confidences.

5.4 Regularizers and Priors as Inconsistencies

Regularizers are extra terms added to loss funtions, 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,[3] in which L2 regularization corresponds to a Gaussian prior [**?**], while L1 regularization corresponds to a Laplacian prior [**?**]. 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.

link to proof

Proposition 5.4.1. *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,*

$$\left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[dr, "q", "(\beta)"'] & & \\ & \boxed{\Theta} \arrow[r, "p"] & \boxed{Y} \\ {} \arrow[ur, two heads, "\theta"'] & & {} \arrow[u, "D", "(\infty)"'] \end{tikzcd} \right\rangle\!\!\right\rangle = \mathop{\mathbb{E}}_{y\sim D} \log \frac{1}{p(y\,|\,\theta)} + \beta \log \frac{1}{q(\theta)} - \mathrm{H}(D) \tag{5.4}$$

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 (5.4) becomes

$$\underbrace{\mathbb{E}_D \log \frac{1}{p(Y\,|\,\theta)}}_{\substack{\text{Cross entropy loss}\\ \text{(data-fit cost of }\theta)}} + \underbrace{\frac{\beta}{2}\theta_0}_{\substack{\text{L2 regularizer}\\ \text{(complexity cost of }\theta)}} + \underbrace{\beta \log k - \mathrm{H}(D)}_{\text{constant in } p \text{ and } \theta},$$

which is the L2 regularized version of Proposition 5.3.2. Moreover, the regularization strength corresponds exactly to the confidence β. What about other 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.

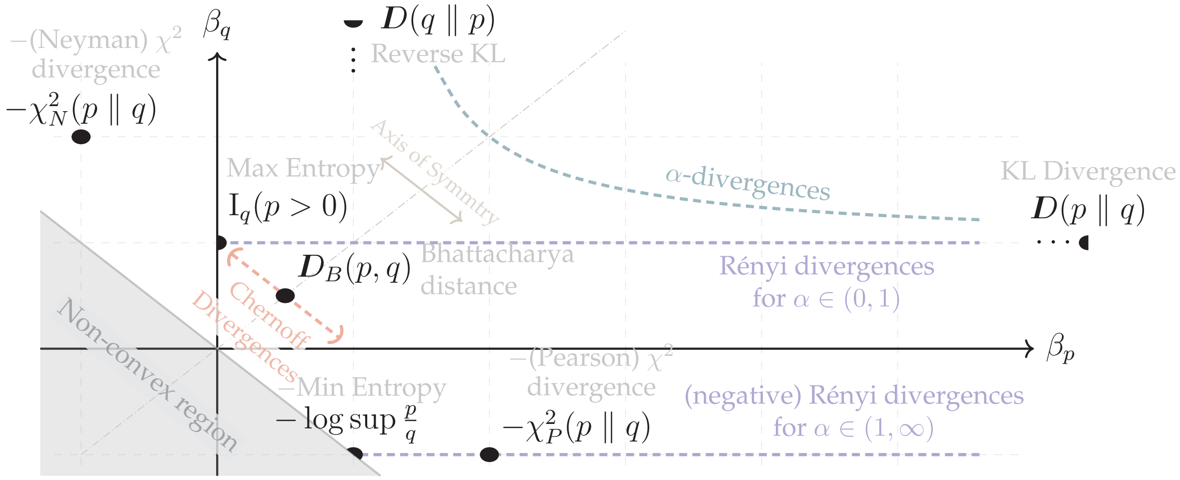

−(Neyman) χ^2 divergence
$-\chi^2_N(p \parallel q)$
β_q
$\boldsymbol{D}(q \parallel p)$
Reverse KL
Max Entropy
$\mathrm{I}_q(p > 0)$
Axis of Symmetry
α-divergences
KL Divergence
$\boldsymbol{D}(p \parallel q)$
$\boldsymbol{D}_B(p, q)$
Bhattacharya distance
Rényi divergences for $\alpha \in (0, 1)$
Chernoff Divergences
Non-convex region
β_p
−(Pearson) χ^2 divergence
−Min Entropy
$-\log \sup \frac{p}{q}$
(negative) Rényi divergences for $\alpha \in (1, \infty)$
$-\chi^2_P(p \parallel q)$

Figure 1: A map of the inconsistency of the PDG comprising $p(X)$ and $q(X)$, as we vary their respective confidences β_p and β_q. Solid circles indicate well-known named measures, semicircles indicate limiting values, and the heavily dashed lines are well-established classes.

5.5 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 β_p and β_q of each cpd, so we can form a PDG divergence $\boldsymbol{D}^{\mathrm{PDG}}_{(r,s)}(p \| q)$ for every setting (r, s) of (β_p, β_q). It turns out that a large class of statistical divergences arise in this way. We start with a familiar one.

Proposition 5.5.1 (KL Divergence as Inconsistency)**.** *The inconsistency of believing p with complete certainty, and also q with some finite certainty β, is β times the KL Divergence (or relative entropy) of q with respect to p. That is,*

$$\left\langle\!\!\left\langle \xrightarrow[(\infty)]{p} \boxed{X} \xleftarrow[(\beta)]{q} \right\rangle\!\!\right\rangle = \beta\, \boldsymbol{D}(p \parallel q).$$

[3] A full account can be found in the appendix.

This result gives us an intuitive interpretation of the asymmetry of relative entropy / KL divergence, and a prescription about when it makes sense to use it. $\boldsymbol{D}(p \parallel q)$ 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 $\boldsymbol{D}(p \parallel q)$ 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.

[link to proof]

Lemma 5.5.2. *The inconsistency $\boldsymbol{D}^{\mathrm{PDG}}_{(r,s)}(p\|q)$ of a PDG comprising $p(X)$ with confidence r and $q(X)$ with confidence s is given in closed form by*

$$\left\langle\!\!\left\langle \xrightarrow[(r)]{p} \boxed{X} \xleftarrow[(s)]{q} \right\rangle\!\!\right\rangle = -(r+s)\log \sum_x \Big(p(x)^r q(x)^s \Big)^{\frac{1}{r+s}} .$$

Of the many generalizations of KL divergence, Rényi divergences, first characterized by Alfréd Rényi **?** are perhaps the most significant, as few others have found either application or an interpretation in terms of coding theory [**?**]. The Rényi divergence of order α between two distributions $p(X)$ and $q(X)$ is given by

$$\boldsymbol{D}_\alpha(p \parallel q) := \frac{1}{1-\alpha} \log \sum_{x \in \mathcal{V}(X)} p(x)^\alpha q(x)^{1-\alpha}. \tag{5.5}$$

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 (5.5), because they satisfy a natural weakening of standard postulates for Shannon entropy due to **?**]. 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 (5.5) for some α [**?**]. It follows from Lemma 5.5.2

that every Rényi divergence is a PDG divergence, and every (non-limiting) PDG divergence is a (scaled) Rényi divergence.

Corollary 5.5.2.1 (Rényi Divergences)**.**

$$\left\langle\!\!\left\langle \xrightarrow[(r)]{p} \boxed{X} \xleftarrow[(s)]{q} \right\rangle\!\!\right\rangle = s \cdot \boldsymbol{D}_{\frac{r}{r+s}}(p \,\|\, q)$$
$$\textit{and} \qquad \boldsymbol{D}_\alpha(p \,\|\, q) = \left\langle\!\!\left\langle \xrightarrow[(\frac{\alpha}{1-\alpha})]{p} \boxed{X} \xleftarrow{q} \right\rangle\!\!\right\rangle$$

However, the two classes are not identical, because the PDG divergences have extra limit points. One big difference is that the reverse KL divergence $\boldsymbol{D}(q \,\|\, p)$ is not a Rényi divergence $\boldsymbol{D}_\alpha(p \,\|\, q)$ for any value (or limit) of α. This lack of symmetry has led others [e.g., **?**] to work instead with a symmetric variant called α-divergence, rescaled by an additional factor of $\frac{1}{\alpha}$. The relationships between these quantities can be seen in Figure 1.

The Chernoff divergence measures the tightest possible exponential bound on probability of error [**?**] in Bayesian hypothesis testing. It also happens to be the smallest possible inconsistency of simultaneously believing p and q, with total confidence 1.

Corollary 5.5.2.2. *The Chernoff Divergence between p and q equals*

$$\inf_{\beta \in (0,1)} \left\langle\!\!\left\langle \xrightarrow[(\beta)]{p} \boxed{X} \xleftarrow[(1-\beta)]{q} \right\rangle\!\!\right\rangle.$$

One significant consequence of representing divergences as inconsistencies is that we can use Lemma 5.2.1 to derive relationships between them. The following facts follow directly from Figure 1, by inspection.

Corollary 5.5.2.3. 1. *Rényi entropy is monotonic in its parameter α.*

2. $\boldsymbol{D}(p \,\|\, q) \geq 2\boldsymbol{D}_B(p, q) \leq \boldsymbol{D}(q \,\|\, p)$.

3. *If $q(p > 0) < 1$ (i.e., $q \not\ll p$), then $\boldsymbol{D}(q \,\|\, p) = \infty$.*

These divergences correspond to PDGs with only two edges and one variable. What about more complex graphs? For a start, conditional divergences

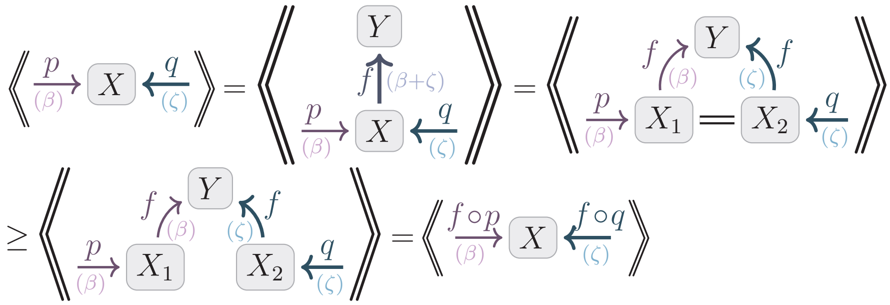


Figure 2: A visual proof of the data-processing inequality. In words: the cpd $f(Y|X)$ can always be satisfied, so adds no inconsistency. It is then equivalent to split f and the variable X into X_1 and X_2 with edges enforcing $X_1 = X_2$. But removing such edges can only decrease inconsistency. Finally, compose the remaining cpds to give the result. See the appendix for a full justification.

$$\boldsymbol{D}^{\mathrm{PDG}}_{(r,s)}\Big(p(Y|X)\,\Big\|\,q(Y|X)\,\Big|\,r(X)\Big) := \mathop{\mathbb{E}}_{x\sim r}\boldsymbol{D}^{\mathrm{PDG}}_{(r,s)}\Big(p(Y|x)\,\Big\|\,q(Y|x)\Big)$$

can be represented straightforwardly as

$$\boldsymbol{D}^{\mathrm{PDG}}_{(r,s)}(p \,\|\, q \,|\, r) = \left\langle\!\!\left\langle\; \begin{tikzcd} {} \arrow[r, "r", "(\infty)"'] & \boxed{X} \arrow[r, bend left, "p\,{\scriptstyle(r)}"] \arrow[r, bend right, "q\,{\scriptstyle(s)}"'] & \boxed{Y} \end{tikzcd} \;\right\rangle\!\!\right\rangle .$$

Other structures are useful intermediates. Lemma 5.2.1, plus some structural manipulation, gives visual proofs of many divergence properties; Figure 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.

5.6 Variational Objectives and Bounds

The fact that the incompatibility of $\mathcal{m}$ with a *specific* joint distribution μ 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 and provide a graphical proof language for it.

5.6.1 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 [**?** , §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 5.3.3). 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 $\mathrm{ELBO}_{p,q}(x) := \mathbb{E}_{z\sim q} \log \frac{p(x,z)}{q(z)}$. This is something we can estimate, since we can sample from q. By Jensen's inequality,

$$\operatorname*{ELBO}_{p,q}(x) = \mathop{\mathbb{E}}_{q} \log \frac{p(x,Z)}{q(Z)} \le \log \left[\mathop{\mathbb{E}}_{q} \frac{p(x,Z)}{q(Z)} \right] = \log p(x),$$

with equality if $q(Z) = p(Z)$. So to find p maximizing $p(x)$, it suffices to adjust p and q to maximize $\mathrm{ELBO}_{p,q}(x)$,[4] provided $\mathcal{Q}$ is expressive enough.

The formula for the ELBO is somewhat difficult to make sense of.[5] Nevertheless, it arises naturally as the inconsistency of the appropriate PDG.

link to proof

Proposition 5.6.1. *The negative ELBO of x is the inconsistency of the PDG containing p,q, and $X{=}x$, with high confidence in q. That is,*

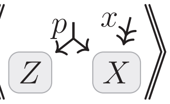
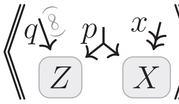
$$-\mathrm{ELBO}_{p,q}(x) = \left\langle\!\!\left\langle \begin{tikzcd}[row sep=small] & & p \arrow[dl] \arrow[dr] & & \\ {} \arrow[r, "q", "(\infty)"'] & \boxed{Z} & & \boxed{X} & {} \arrow[l, two heads, "x"'] \end{tikzcd} \right\rangle\!\!\right\rangle.$$

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

[4] or for many iid samples: $\max_{p,q} \sum_{x\in\mathcal{D}} \mathrm{ELBO}_{p,q}(x)$.

[5] Especially if p, q are densities. See Section 5.A.

visually intuitive proof of the bound traditionally used to motivate the ELBO, via Lemma 5.2.1:

$\log\frac{1}{p(x)} = \langle\!\langle$  $\rangle\!\rangle \le \langle\!\langle$  $\rangle\!\rangle = -\underset{p,q}{\mathrm{ELBO}}(x).$

The first and last equalities are Propositions 5.6.1 and 5.3.3 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 $q(Z)=p(Z)$: the best distribution for the left PDG has marginal $p(Z)$ anyway, so insisting on it incurs no further cost.

5.6.2 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* (5.6): when we decode an encoded image, we would like it to resemble the original.

$$\mathrm{Rec}(x) := \underset{z\sim e(Z|x)}{\mathbb{E}} \underbrace{\mathrm{I}_{d(X|z)}(x)}_{\left(\substack{\text{additional bits required to} \\ \text{decode } x \text{ from its encoding } z}\right)} = \sum_z e(z\,|\,x) \log \frac{1}{d(x\,|\,z)} \tag{5.6}$$

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 [17**?**], 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 5.6.1 only in that it can depend on X. VAEs are trained with the analogous form of the ELBO:

$$\begin{aligned}\mathrm{ELBO}_{p,e,d}(x) :=& \mathop{\mathbb{E}}_{z\sim e(Z|x)}\left[\log\frac{p(z)d(x\mid z)}{e(z\mid x)}\right]\\ =& -\mathrm{Rec}(x) - \boldsymbol{D}(e(Z|x)\parallel p).\end{aligned}$$

This gives us the following analog of Proposition 5.6.1.

link to proof

Proposition 5.6.2. *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,*

$$-\mathrm{ELBO}_{p,e,d}(x) = \left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "p"] & \boxed{Z} \arrow[r, bend left, "d"] & \boxed{X} \arrow[l, bend left, "e", "(\infty)"'] & {} \arrow[l, two heads, "x"'] \end{tikzcd} \right\rangle\!\!\right\rangle.$$

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:

$$\log\frac{1}{\underset{p,d}{\Pr}(x)} = \left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[d, "p"] & & {} \arrow[d, two heads, "x"] \\ \boxed{Z} \arrow[rr, bend left, "d"] & & \boxed{X} \end{tikzcd} \right\rangle\!\!\right\rangle \leq \left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[d, "p"] & & {} \arrow[d, two heads, "x"] \\ \boxed{Z} \arrow[rr, bend left, "d"] & & \boxed{X} \arrow[ll, "e", "(\infty)"'] \end{tikzcd} \right\rangle\!\!\right\rangle = \text{-}\underset{p,e,d}{\mathrm{ELBO}}(x).$$

The first and last equalities are Propositions 5.6.2 and 5.3.3, and the inequality is Lemma 5.2.1. See the appendix for multi-sample analogs of the bound and Proposition 5.6.2.

5.6.3 The β-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 **?**], one weights the reconstruction error (5.6) and the 'KL term' differently, resulting in a loss function of the form

$$\beta\text{-ELBO}_{p,e,d}(x) := -\text{Rec}(x) - \beta \boldsymbol{D}(e(Z|x) \parallel p),$$

which, when $\beta{=}1$, is the ELBO as before. The authors view β as a regularization strength, and argue that it sometimes helps to have a stronger prior. Sure enough:

link to proof

Proposition 5.6.3. *$-\beta$-ELBO$_{p,e,d}(x)$ is the inconsistency of the same PDG, but with confidence β in $p(Z)$.*

5.7 Free Energy as Factor Graph Inconsistency

A weighted factor graph $\Psi = (\phi_J, \theta_J)_{J\in\mathcal{J}}$, where each θ_J is a real-valued weight, J is associated with a subset of variables $\mathbf{X}_J$, and $\phi_J : \mathcal{V}(\mathbf{X}_J) \to \mathbb{R}$, 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_Ψ 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 [**?**]. Following **?**], let $\boldsymbol{pdg}(\Psi)$ be the PDG with edges $\{\xrightarrow{J}\mathbf{X}_J\}_{\mathcal{J}}$, 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 $[\![\boldsymbol{pdg}(\Psi)]\!]_1$. But what about the corresponding inconsistency, $\langle\!\langle \boldsymbol{pdg}(\Psi) \rangle\!\rangle_1$?

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 Ψ's inconsistency.

link to proof

Proposition 5.7.1. *For all weighted factor graphs Ψ, we have that $\langle\!\langle \boldsymbol{\mathit{pdg}}(\Psi) \rangle\!\rangle_1 = -\log Z_\Psi$.*

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 succesful in describing of not only chemical and biological systems [**?**], but also cognitive ones [6].

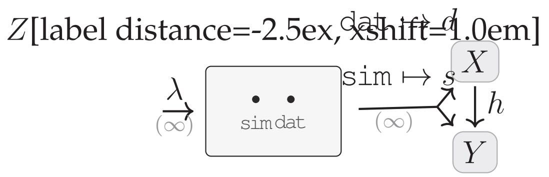
5.8 Beyond Standard Losses: A Concrete Example

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 unsalvagable, 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, $\mathcal{L}_{\mathtt{sim}} := \mathbb{E}_s[\log {}^1\!/\!{}_{h(Y|X)}]$. Conversely, if the simulation were bad and the data mostly intact, we would use $\mathcal{L}_{\mathtt{dat}}$, 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, $\mathcal{L}_1 := \lambda_{\texttt{s}}\mathcal{L}_{\texttt{sim}} + \lambda_{\texttt{d}}\mathcal{L}_{\texttt{dat}}$ for some $\lambda_{\texttt{s}}, \lambda_{\texttt{d}} > 0$ with $\lambda_{\texttt{s}} + \lambda_{\texttt{d}} = 1$. This amounts to training h with cross entropy with respect to the mixture $\lambda_{\texttt{s}} s + \lambda_{\texttt{d}} d$. 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

$\mathcal{L}_1 = \left\langle\!\!\left\langle\right.\right.$

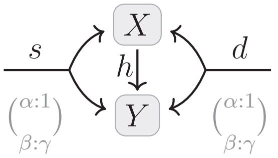


$\left.\left.\right\rangle\!\!\right\rangle$,

in which a swich variable Z with possible values $\{\texttt{sim}, \texttt{dat}\}$ controls whether samples come from s or d, and is distributed according to $\lambda(Z\!=\!\texttt{sim}) = \lambda_{\texttt{s}}$.

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 (unnormalized) product density ds, which in many ways is appropriate. However, by this metric, the optimal predictor $h^*(Y|x) \propto d(Y|x)s(Y|x)$ is *uncalibrated* [**?**]. 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 γ, it is essentially the γ-inconsistency

$\mathcal{L}_2 \approx C \left\langle\!\!\left\langle\right.\right.$



$\left.\left.\right\rangle\!\!\right\rangle_{\gamma} + \mathit{const},$

where C is the constant required to normalize the joint density sd, and const does not depend on h. However, the values of $\boldsymbol{\alpha}$ in this PDG indicate an overdetermination of XY (it is determined in two different ways), and so h^* is more deterministic than intended. By contrast,

$$\mathcal{L}_3 := \left\langle\!\!\left\langle \begin{tikzcd} & & \boxed{X} \arrow[dd, "h"'] & & \\ {} \arrow[r, dash, "s", "(\lambda_{\mathrm{s}})"'] & {} \arrow[ur, bend left] \arrow[dr, bend right] & & {} \arrow[ul, bend right] \arrow[dl, bend left] & {} \arrow[l, dash, "d"', "(\lambda_{\mathrm{d}})"] \\ & & \boxed{Y} & & \end{tikzcd} \right\rangle\!\!\right\rangle,$$

does not have this issue: the optimal predictor h^* according to $\mathcal{L}_3$ is proportional to the λ-weighted geometric mean of s and d. 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.

5.9 Reverse-Engineering a Loss Function?

Given an *arbitrary* loss function, can we find a PDG that gives rise to it? The answer appears to be yes—although not without making unsavory modeling choices. Without affecting its semantics, one may add the variable $\mathtt{T}$ that takes values $\{\mathtt{t}, \mathtt{f}\}$, and the event $\mathtt{T} = \mathtt{t}$, to any PDG. Now, given a cost function $c : \mathcal{V}(X) \to \mathbb{R}_{\geq 0}$, define the cpd $\hat{c}(\mathtt{T}|X)$ by $\hat{c}(\mathtt{t}|x) := e^{-c(x)}$. By threatening to generate the falsehood $\mathtt{f}$ with probability dependent on the cost of X, $\hat{c}$ ties the value of X to inconsistency.

link to proof

Proposition 5.9.1. $\left\langle\!\!\left\langle \xrightarrow[(\infty)]{p} \boxed{X} \xrightarrow{\hat{c}} \boxed{\mathtt{T}} \overset{\mathtt{t}}{\twoheadleftarrow} \right\rangle\!\!\right\rangle = \mathop{\mathbb{E}}_{x \sim p} c(x).$

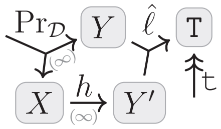
Setting confidence $\beta_p := \infty$ may not be realistic since we're still training the model p, but doing so is necessary to recover $\mathbb{E}_p\, c$.[6] Any mechanism that generates inconsistency based on the value of X (such as this one) also works in reverse: the PDG "squirms", contorting the probability of X to disperse the inconsistency. One cannot cannot simply "emit loss" without affecting the rest of

[6] If β_p were instead equal to 1, we would have obtained $-\log \mathbb{E}_p \exp(-c(X))$, with optimal distribution $\mu(X) \neq p(X)$.

the model, as one does with utility in an Influence Diagram [**?**]. 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 ℓ on pairs of output labels.

Concretely, define:

$\mathcal{S} :=$



and $$\mathcal{L} := \mathop{\mathbb{E}}_{\substack{(x,y)\sim\Pr_{\mathcal{D}} \\ y'\sim p(Y'|\,x)}} \big[\ell(y,y')\big].$$

Given Proposition 5.9.1, 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 ℓ. 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 $\lim\limits_{\gamma\to\infty} \langle\!\langle \mathcal{S} \rangle\!\rangle_\gamma = \mathcal{L}$.

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 γ, 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-enerineering 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.

5.10 Conclusions

We seen that that PDG semantics, in the same stroke by which they capture Bayesian Networks and Factor Graphs [**?**], 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.

Acknowledgements

Work supported in part by MURI grant W911NF-19-1-0217. Many thanks to my advisor, Joe Halpern, for his generous support, and for valuable critiques of many drafts. Thanks as well to my reviewers, who pushed me to better explain the confidence parameters, and to include a practical example (Section 5.8). Finally, thanks to my friends, particularly Varsha Kishore and Greg Yauney, for helping me to refine the presentation of these ideas.

APPENDICES FOR CHAPTER 5

5.A The Fine Print for Probability Densities

Densities and Masses. Many of our results (Propositions 5.3.1 to 5.B.5) technically require the distribution to be represented with a mass function (as opposed to a probability density function, or pdf). A PDG containg 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 discretization 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.

Markov Kernels. In the more general setting of measurable spaces, one may want to adjust the definition of a cpd that we gave, so that one instead works with *Markov Kernels*. This imposes an additional constraint: suppose the variable Y takes values in the measurable space $(\mathcal{V}(Y), \mathcal{B})$. If $p(Y|X)$ is to be a *Markov Kernel*, then for every fixed measurable subset $B \in \mathcal{B}$ of the measure space, the we must require that $x \mapsto \Pr(B|x)$ be a measurable function (with respect to the measure space in which X takes values). This too mostly does not bear on the present discussion, because the σ-algebras for all measure spaces of interest, are fine enough that one can get an arbitrarily close approximation of any cpd with a Markov Kernels. This means that the infemum defining the inconsistency of a PDG does not change.

5.B Further Results and Generalizations

5.B.1 Full Characterization of Gaussian Predictors

The inconsistency of a PDG containing two univariate Gaussian regressors of with arbitrary paremeters and confidences, is most cleanly articulated in terms of the geometric and quadratic means.

Definition 5.B.1 (Weighted Power Mean)**.** The weighted power mean $\mathrm{M}_p^w(\mathbf{r})$ of

Name	p	Formula
Harmonic	$(p = -1)$:	$\mathrm{HM}_w(\mathbf{r}) = {}^{1}\!/_{(\sum_{i=1}^{n} {}^{w_i}\!/_{r_i})}$
Geometric	$(\lim p \to 0)$:	$\mathrm{GM}_w(\mathbf{r}) = \prod_{i=1}^{n} r_i^{w_i}$
Arithmetic	$(p = 1)$:	$\mathrm{AM}_w(\mathbf{r}) = \sum_{i=1}^{n} w_i r_i$
Quadratic	$(p = 2)$:	$\mathrm{QM}_w(\mathbf{r}) = \sqrt{\sum_{i=1}^{n} w_i r_i^2}$

Table 5.B.1: special cases of the p-power mean $\mathrm{M}_p^w(\mathbf{r})$

the collection of real numbers $\mathbf{r} = r_1, \ldots, r_n$ with respect to the convex weights $w = w_1, \ldots, w_n$ satisfying $\sum_i w_i = 1$, is given by

$$\mathrm{M}_p^w(\mathbf{r}) := \Big(\sum_{i=1}^{n} w_i (r_i)^p \Big)^{\frac{1}{p}}.$$

We omit the superscript as a shorthand for the uniform weighting $w_i = {}^{1}\!/_{N}$. □

Many standard means, such as those in Table 5.B.1, are special cases. It is well known that $\mathrm{M}_p^w(\mathbf{r})$ is increasing in p, and strictly so if not all elements of $\mathbf{r}$ are identical. In particular, $\mathrm{QM}_w(a, b) > \mathrm{GM}_w(a, b)$ for all $a \neq b$ and positive weights w. We now present the result.

Proposition 5.B.1. *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]

$$\left\langle\!\!\left\langle \begin{tikzcd} & & \mu_1 \arrow[ddr, "\mathcal{N}", "(\beta_1)"'] & \\ & & \sigma_1 \arrow[dr] & \\ {} \arrow[r, "D", "(\infty)"'] & X \arrow[uur, two heads, "f"] \arrow[ur, two heads, "s"] \arrow[dr, two heads, "t"'] \arrow[ddr, two heads, "h"'] & & Y \\ & & \sigma_2 \arrow[ur] & \\ & & \mu_2 \arrow[uur, "\mathcal{N}"', "(\beta_2)"] & \end{tikzcd} \right\rangle\!\!\right\rangle = \mathop{\mathbb{E}}_{D}\left[(\beta_1{+}\beta_2)\log\frac{\mathrm{QM}_{\hat\beta}(\sigma_1,\sigma_2)}{\mathrm{GM}_{\hat\beta}(\sigma_1,\sigma_2)} + \frac{1}{2}\frac{\beta_1\beta_2}{\beta_1+\beta_2}\left(\frac{\mu_1-\mu_2}{\mathrm{QM}_{\hat\beta}(\sigma_1,\sigma_2)}\right)^2\right] \tag{5.7}$$

$$= \frac{1}{2}\mathop{\mathbb{E}}_{x\sim D}\left[\frac{\beta_1\beta_2}{2}\frac{\Big(f(x)-h(x)\Big)^2}{\beta_2 s(x)^2+\beta_1 t(x)^2} + \frac{\beta_1+\beta_2}{2}\log\frac{\beta_2 s(x)^2+\beta_1 t(x)^2}{\beta_1+\beta_2} \begin{array}{l} -\beta_2\log s(x \\ -\beta_1\log t(x \end{array}\right.$$

where $\hat\beta = (\frac{\beta_2}{\beta_1+\beta_2}, \frac{\beta_1}{\beta_1+\beta_2})$ represents the normalized and reversed vector of conficences $\beta = (\beta_1, \beta_2)$ 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

$$\begin{tikzcd} {} \arrow[r, "D", "(\infty)"'] & X \arrow[r, bend left, "\mathcal{N}(f(x),s(x))"] \arrow[r, bend right, "\mathcal{N}(h(x),t(x))"'] & Y \end{tikzcd} .$$

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 5.B.1 also gives a proof of the inequality of the weighted geometric and quadratic means.

Corollary 5.B.1.1. *For all σ_1 and σ_2, and all weight vectors β, $\mathrm{QM}_{\hat\beta}(\sigma_1,\sigma_2) \geq \mathrm{GM}_{\hat\beta}(\sigma_1,\sigma_2)$.*

5.B.2 Full-Dataset ELBO and Bounds

We now present the promised multi-sample analogs from Section 5.6.1.

[link to proof]

Proposition 5.B.2. *The following analog of Proposition 5.6.2 for a whole dataset $\mathcal{D}$ holds:*

$$-\mathop{\mathbb{E}}_{\Pr_{\mathcal{D}}} \mathrm{ELBO}_{p,e,d}(X) = \left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "p"] & \boxed{Z} \arrow[r, bend left, "d"] & \boxed{X} \arrow[l, bend left, "e", "(\infty)"'] & {} \arrow[l, "\Pr_{\mathcal{D}}"', "(\infty)"] \end{tikzcd} \right\rangle\!\!\right\rangle + \mathrm{H}(\Pr_{\mathcal{D}}).$$

Propositions 5.3.2 and 5.B.2 then give us an analog of the visual bounds in the body of the main paper (Section 5.6.1) for many i.i.d. datapoints at once, with only a single application of the inequality:

$$-\log \Pr(\mathcal{D}) = -\log \prod_{i=1}^{m} \left(\Pr(x^{(i)})\right) = -\frac{1}{m}\sum_{i=1}^{m} \log \Pr(x^{(i)}) =$$

$$\mathrm{H}(\Pr_{\mathcal{D}}) + \left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "p"] & \boxed{Z} \arrow[r, bend left, "d"] & \boxed{X} & {} \arrow[l, "\Pr_{\mathcal{D}}"', "(\infty)"] \end{tikzcd} \right\rangle\!\!\right\rangle \leq \left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "p"] & \boxed{Z} \arrow[r, bend left, "d"] & \boxed{X} \arrow[l, bend left, "e", "(\infty)"'] & {} \arrow[l, "\Pr_{\mathcal{D}}"', "(\infty)"] \end{tikzcd} \right\rangle\!\!\right\rangle + \mathrm{H}(\Pr_{\mathcal{D}})$$

$$= -\mathop{\mathbb{E}}_{\Pr_{\mathcal{D}}} \mathop{\mathrm{ELBO}}_{p,e,d}(X)$$

We also have the following formal statement of Proposition 5.6.3.

[link to proof]

Proposition 5.B.3. *The negative β-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 β. That is,*

$$-\beta\text{-ELBO}_{p,e,d}(x) = \left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "p", "(\beta)"'] & \boxed{Z} \arrow[r, bend left, "d"] & \boxed{X} \arrow[l, bend left, "e", "(\infty)"'] & {} \arrow[l, twoheadrightarrow, "x"'] \end{tikzcd} \right\rangle\!\!\right\rangle$$

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 5.B.3.1 (reconstruction error as inconsistency)**.**

$$-\text{Rec}_{ed,d}(x) := \mathop{\mathbb{E}}_{z\sim e(Z|x)} \mathrm{I}_{d(X|z)}(x) = \left\langle\!\!\left\langle \begin{tikzcd} \boxed{Z} \arrow[r, bend left, "d"] & \boxed{X} \arrow[l, bend left, "e", "(\infty)"'] & {} \arrow[l, twoheadrightarrow, "x"'] \end{tikzcd} \right\rangle\!\!\right\rangle$$

5.B.3 More Variants of Cross Entropy Results

First, we show that our cross entropy results hold for all γ, in the sense that γ contributes only a constant.

link to proof

Proposition 5.B.4. *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 $\mathrm{Pr}_\mathcal{D}$, 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 $\mathrm{Pr}_\mathcal{D}$*
3. $[\![\, p \,]\!]_\gamma(\mathrm{Pr}_\mathcal{D}) \quad + (1+\gamma)\,\mathrm{H}(\mathrm{Pr}_\mathcal{D})$
4. $\left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "p"] & \boxed{X} & {} \arrow[l, "\mathrm{Pr}_\mathcal{D}"', "(\infty)"] \end{tikzcd} \right\rangle\!\!\right\rangle_\gamma \quad + (1+\gamma)\,\mathrm{H}(\mathrm{Pr}_\mathcal{D})$

As promised, we now give the simultaneous generalization of the surprisal result (Proposition 5.3.1) to both multiple samples (like in Proposition 5.3.2) and partial observations (as in Proposition 5.3.3).

link to proof

Proposition 5.B.5. *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,*

$$\ell(p;\mathcal{D}) = \left\langle\!\!\left\langle \begin{tikzcd}[row sep=small] & {} \arrow[dl, "p"'] \arrow[dr] & & \\ \boxed{Z} & & \boxed{X} & {} \arrow[l, "\Pr_{\mathcal{D}}", "(\infty)"'] \end{tikzcd} \right\rangle\!\!\right\rangle + \mathrm{H}(\Pr_{\mathcal{D}}).$$

5.C PROOFS

Lemma 5.2.1. *Suppose PDGs $\mathcal{M}$ and $\mathcal{M}'$ differ only in their edges (resp. $\mathcal{A}$ and $\mathcal{A}'$) and confidences (resp. $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$). If $\mathcal{A} \subseteq \mathcal{A}'$ and $\beta_L \leq \beta'_L$ for all $L \in \mathcal{A}$, then $\langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma \leq \langle\!\langle \mathcal{M}' \rangle\!\rangle_\gamma$ for all γ.*

Proof. For every μ, adding more edges only adds non-negative terms to (5.1), while increasing β results in larger coefficients on the existing (non-negative) terms of (5.1). So for every fixed distribution μ, we have $[\![\mathcal{M}]\!]_\gamma(\mu) \leq [\![\mathcal{M}']\!]_\gamma(\mu)$. So it must also be the case that the infemum over μ, so we find that $\langle\!\langle \mathcal{M} \rangle\!\rangle \leq \langle\!\langle \mathcal{M}' \rangle\!\rangle$. □

Proposition 5.3.1. *Consider a distribution $p(X)$. The inconsistency of the PDG comprising p and $X{=}x$ equals the surprisal $\mathrm{I}_p[X{=}x]$. That is,*

$$\mathrm{I}_p[X{=}x] = \left\langle\!\!\left\langle \xrightarrow{p} \boxed{X} \overset{x}{\twoheadleftarrow} \right\rangle\!\!\right\rangle.$$

(Recall that $\langle\!\langle \mathcal{M} \rangle\!\rangle$ is the inconsistency of the PDG $\mathcal{M}$.)

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

$$\left\langle\!\!\left\langle \xrightarrow{p} \boxed{X} \overset{x}{\twoheadleftarrow} \right\rangle\!\!\right\rangle = \left[\!\!\left[\xrightarrow{p} \boxed{X} \overset{x}{\twoheadleftarrow} \right]\!\!\right](\delta_x) = \boldsymbol{D}(\delta_x \parallel p) = \log \frac{1}{p(x)} = \mathrm{I}_p(x).$$

□

Proposition 5.B.4 is a generalization of Proposition 5.3.2, so we prove them at the same time.

Proposition 5.3.2. *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* $\quad \mathrm{CrossEntropy}(\Pr_\mathcal{D}, p) =$

$$\frac{1}{m}\sum_{i=1}^{m} \mathrm{I}_p[X{=}x_i] = \left\langle\!\!\left\langle \xrightarrow{p} \boxed{X} \xleftarrow[(\infty)]{\Pr_\mathcal{D}} \right\rangle\!\!\right\rangle + \mathrm{H}(\Pr_\mathcal{D}).$$

Proposition 5.B.4. *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 $\gamma \geq 0$:*

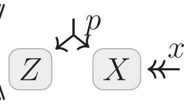
1. *The average negative log likelihood $\ell(p; \mathcal{D}) = -\frac{1}{m}\sum_{i=1}^m \log p(x_i)$*
2. *The cross entropy of p relative to $\Pr_\mathcal{D}$*
3. $[\![p]\!]_\gamma(\Pr_\mathcal{D}) \quad + (1+\gamma)\,\mathrm{H}(\Pr_\mathcal{D})$
4. $\left\langle\!\!\left\langle \xrightarrow{p} \boxed{X} \xleftarrow[(\infty)]{\Pr_\mathcal{D}} \right\rangle\!\!\right\rangle_\gamma \quad + (1+\gamma)\,\mathrm{H}(\Pr_\mathcal{D})$

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:

$$\begin{aligned}
[\![p]\!]_\gamma(\Pr_\mathcal{D}) + (1+\gamma)\,\mathrm{H}(\Pr_\mathcal{D}) &= \boldsymbol{D}(\Pr_\mathcal{D} \parallel p) - \gamma\,\mathrm{H}(\Pr_\mathcal{D}) + (1+\gamma)\,\mathrm{H}(\Pr_\mathcal{D}) \\
&= \boldsymbol{D}(\Pr_\mathcal{D} \parallel p) + \mathrm{H}(\Pr_\mathcal{D}) \\
&= \mathbb{E}_{\Pr_\mathcal{D}}\left[\log\frac{\Pr_\mathcal{D}}{p} + \log\frac{1}{\Pr_\mathcal{D}}\right] = \mathbb{E}_{\Pr_\mathcal{D}}\left[\log\frac{1}{p}\right],
\end{aligned}$$

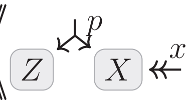
which is the cross entropy, as desired. □

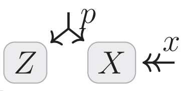
Proposition 5.3.3. *If $p(X, Z)$ is a joint distribution, then the information content of the partial observation $X = x$ is given by*

$$\mathrm{I}_p[X{=}x] = \left\langle\!\!\left\langle \vphantom{\int} \right.\right.$$


$$\left.\left. \vphantom{\int}\right\rangle\!\!\right\rangle. \tag{5.2}$$

Proof. As before, all mass of μ 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

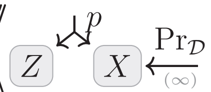


$$\left\langle\!\!\left\langle \cdots \right\rangle\!\!\right\rangle = \inf_{\mu(Z)} [\![\cdots]\!] \Big(\delta_x(X)\mu(Z)\Big)$$


$$\begin{aligned}
&= \inf_{\mu(Z)} \boldsymbol{D}\Big(\delta_x(X)\mu(Z) \,\Big\|\, p(X,Z)\Big) \\
&= \inf_{\mu(Z)} \mathop{\mathbb{E}}_{z\sim\mu} \log \frac{\mu(z)}{p(x,z)} \;\; = \inf_{\mu(Z)} \mathop{\mathbb{E}}_{z\sim\mu} \log \frac{\mu(z)}{p(x,z)}\frac{p(x)}{p(x)} \\
&= \inf_{\mu(Z)} \mathop{\mathbb{E}}_{z\sim\mu} \left[\log \frac{\mu(z)}{p(z \mid x)} + \log \frac{1}{p(x)}\right] \\
&= \inf_{\mu(Z)} \Big[\boldsymbol{D}(\mu(Z) \,\|\, p(Z \mid x))\Big] + \log \frac{1}{p(x)} \\
&= \log \frac{1}{p(x)} = \mathrm{I}_p(x) && \text{[Gibbs Inequality]}
\end{aligned}$$

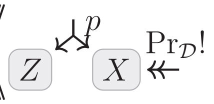
□

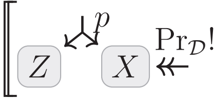
Proposition 5.B.5. *The average* marginal *negative log likelihood* $\ell(p;x) := -\frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}} \log \sum_z p(x,z)$ *is the inconsistency of the PDG containing* p *and the data distribution* $\Pr_{\mathcal{D}}$*, plus the entropy of the data distribution (which is constant in* p*). That is,*

$$\ell(p;\mathcal{D}) = \left\langle\!\!\left\langle \vphantom{\int} \right.\right.$$


$$\left.\left. \vphantom{\int}\right\rangle\!\!\right\rangle + \mathrm{H}(\Pr_{\mathcal{D}}).$$

Proof. The same idea as in Proposition 5.3.3, but a little more complicated.

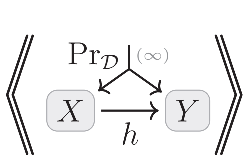


$$\left\langle\!\!\left\langle \cdots \right\rangle\!\!\right\rangle = \inf_{\mu(Z|X)} [\![\cdots]\!] \Big(\Pr_{\mathcal{D}}(X)\mu(Z \mid X)\Big)$$


$$
\begin{aligned}
&= \inf_{\mu(Z|X)} \boldsymbol{D}\Big(\Pr\nolimits_{\mathcal{D}}(X)\mu(Z \mid X) \;\Big\|\; p(X,Z)\Big) \\
&= \inf_{\mu(Z|X)} \mathop{\mathbb{E}}_{\substack{x\sim\Pr_{\mathcal{D}}\\ z\sim\mu}} \log \frac{\mu(z \mid x)\Pr_{\mathcal{D}}(x)}{p(x,z)} \\
&= \frac{1}{|\mathcal{D}|} \inf_{\mu(Z|X)} \sum_{x\in\mathcal{D}} \mathop{\mathbb{E}}_{z\sim\mu(Z|x)} \log \frac{\mu(z \mid x)\Pr_{\mathcal{D}}(x)}{p(x,z)} \frac{p(x)}{p(x)} \\
&= \frac{1}{|\mathcal{D}|} \inf_{\mu(Z|X)} \sum_{x\in\mathcal{D}} \left[\mathop{\mathbb{E}}_{z\sim\mu} \left[\log \frac{\mu(z \mid x)}{p(z \mid x)} \right] + \log \frac{1}{p(x)} - \log \frac{1}{\Pr_{\mathcal{D}}(x)} \right] \\
&= \frac{1}{|\mathcal{D}|} \sum_{x\in\mathcal{D}} \left[\inf_{\mu(Z|x)} \mathop{\mathbb{E}}_{z\sim\mu} \left[\log \frac{\mu(z \mid x)}{p(z \mid x)} \right] + \log \frac{1}{p(x)} - \log \frac{1}{\Pr_{\mathcal{D}}(x)} \right] \\
&= \frac{1}{|\mathcal{D}|} \sum_{x\in\mathcal{D}} \left[\inf_{\mu(Z)} \Big[\boldsymbol{D}(\mu(Z) \,\|\, p(Z \mid x)) \Big] + \log \frac{1}{p(x)} \right] - \mathrm{H}(\Pr\nolimits_{\mathcal{D}}) \\
&= \frac{1}{|\mathcal{D}|} \sum_{x\in\mathcal{D}} \log \frac{1}{p(x)} - \mathrm{H}(\Pr\nolimits_{\mathcal{D}}) \\
&= \frac{1}{|\mathcal{D}|} \sum_{x\in\mathcal{D}} \mathrm{I}_p(x) - \mathrm{H}(\Pr\nolimits_{\mathcal{D}}) \\
\Big(\quad &= \boldsymbol{D}(\Pr\nolimits_{\mathcal{D}} \,\|\, p) \quad \Big)
\end{aligned}
$$

□

Proposition 5.3.4. *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,*



$$
= \frac{1}{m} \sum_{i=1}^{m} \log \frac{1}{h(y_i \,|\, x_i)} - \mathrm{H}_{\Pr_{\mathcal{D}}}(Y|X).
$$

Proof. $\Pr_{\mathcal{D}}$ has high confidence, it is the only joint distribution μ with finite score. Since f is the only other edge, the inconsistency is therefore

$$
\mathop{\mathbb{E}}_{x\sim\Pr_{\mathcal{D}}} \boldsymbol{D}\Big(\Pr\nolimits_{\mathcal{D}}(Y \mid x) \;\Big\|\; f(Y \mid x)\Big) = \mathop{\mathbb{E}}_{x,y\sim\Pr_{\mathcal{D}}} \left[\log \frac{\Pr_{\mathcal{D}}(y \mid x)}{f(y \mid x)} \right]
$$

$$\begin{aligned} &= \mathop{\mathbb{E}}_{x,y\sim \Pr_{\mathcal{D}}} \left[\log \frac{1}{f(y \mid x)} - \log \frac{1}{\Pr_{\mathcal{D}}(y \mid x)} \right] \\ &= \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \left[\log \frac{1}{f(y \mid x)} \right] \quad - \mathrm{H}_{\Pr_{\mathcal{D}}}(Y \mid X) \end{aligned}$$

□

Proposition 5.3.5. *Consider functions $f, h : X \to Y$ 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 β, is β times the log accuracy of h. That is,*

$$\left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "D", "(\beta)"'] & \boxed{X} \arrow[r, bend left, "h\ (r)"] \arrow[r, bend right, "f\ (s)"'] & \boxed{Y} \end{tikzcd} \right\rangle\!\!\right\rangle \begin{aligned} &= -\beta \log \Pr_{x\sim D}(f(x) = h(x)) \\ &= \beta\, \mathrm{I}_D[f = h]. \end{aligned} \tag{5.3}$$

Proof. Becuase f is deterministic, for every x in the support of a joint distribution μ with finite score, we must have $\mu(Y \mid x) = \delta_f$, since if μ were to place any non-zero mass $\mu(x, y) = \epsilon > 0$ on a pont (x, y) with $y \neq f(x)$ results in an infinite contribution to the KL divergence

$$\boldsymbol{D}(\mu(Y \mid x) \parallel \delta_{f(x)}) = \mathop{\mathbb{E}}_{x,y\sim\mu} \log \frac{\mu(y \mid x)}{\delta_{f(x)}} \geq \mu(y, x) \log \frac{\mu(x, y)}{\mu(x) \cdot \delta_{f(x)}(y)} = \epsilon \log \frac{\epsilon}{0} = \infty.$$

The same holds for h. Therefore, for any μ with a finite score, and x with $\mu(x) > 0$, we have $\delta_{f(x)} = \mu(Y \mid x) = \delta_{h(x)}$, meaning that we need only consider μ 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 μ matches the conditional marginals exactly, and the total incompatibility of such a distribution μ is equal to the relative entropy $\boldsymbol{D}(\mu \parallel D)$, scaled by the confidence β 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 $\boldsymbol{D}(\mu \parallel D)$? It is well known that the conditional distribution $D \mid E \propto \delta_E(X)D(X) = \frac{1}{D(E)}\delta_E(X)D(X)$ satisfies this property uniquely (see, for instance, [11]). Let $f{=}h$ denote the event that f and h agree. Then we calculate

$$\left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "\overset{(\beta)}{D}"] & \boxed{X} \arrow[r, two heads, bend left, "h"] \arrow[r, two heads, bend right, "f"'] & \boxed{Y} \end{tikzcd} \right\rangle\!\!\right\rangle = \inf_{\substack{\mu(X) \text{ s.t.} \\ \operatorname{supp}(\mu)\subseteq[f=h]}} \beta \boldsymbol{D}\Big(\mu(X) \;\Big\|\; D(X)\Big)$$

$$\begin{aligned}
&= \beta \boldsymbol{D}\Big(D \mid [f{=}h] \;\Big\|\; D\Big) \\
&= \beta \mathop{\mathbb{E}}_{D|f=h} \log \frac{\delta_{f=h}(X)D(X)}{D(f{=}h)\cdot D(X)} \\
&= \beta \mathop{\mathbb{E}}_{D|f=h} \log \frac{1}{D(f{=}h)} && \left[\begin{array}{c}\text{since } \delta_{f=h}(x) = 1 \text{ for all } x \text{ that} \\ \text{contribute to the expectation}\end{array}\right] \\
&= -\beta \log D(f = h) && \Big[\text{ since } D(f = h) \text{ is a constant }\Big] \\
&= -\beta \log \Big(\text{accuracy}_{f,D}(h)\Big) \\
&= \beta \, \mathrm{I}_D[f = h].
\end{aligned}$$

□

Proposition 5.B.1. *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*

$$\left\langle\!\!\left\langle \begin{tikzcd} & & \boxed{\mu_1} \arrow[ddr, "\mathcal{N}", "(\beta_1)"'] & \\ & & \boxed{\sigma_1} \arrow[dr] & \\ {} \arrow[r, "D", "(\infty)"'] & \boxed{X} \arrow[uur, two heads, "f"] \arrow[ur, two heads, "s"] \arrow[dr, two heads, "t"'] \arrow[ddr, two heads, "h"'] & & \boxed{Y} \\ & & \boxed{\sigma_2} \arrow[ur] & \\ & & \boxed{\mu_2} \arrow[uur, "\mathcal{N}"', "(\beta_2)"] & \end{tikzcd} \right\rangle\!\!\right\rangle = \mathop{\mathbb{E}}_{D}\left[(\beta_1{+}\beta_2)\log\frac{\mathrm{QM}_{\hat\beta}(\sigma_1,\sigma_2)}{\mathrm{GM}_{\hat\beta}(\sigma_1,\sigma_2)} + \frac{1}{2}\frac{\beta_1\beta_2}{\beta_1+\beta_2}\left(\frac{\mu_1-\mu_2}{\mathrm{QM}_{\hat\beta}(\sigma_1,\sigma_2)}\right)^2\right] \tag{5.7}$$

$$= \frac{1}{2} \mathop{\mathbb{E}}_{x\sim D} \left[\frac{\beta_1\beta_2}{2} \frac{\Big(f(x)-h(x)\Big)^2}{\beta_2 s(x)^2+\beta_1 t(x)^2} + \frac{\beta_1+\beta_2}{2}\log\frac{\beta_2 s(x)^2+\beta_1 t(x)^2}{\beta_1+\beta_2} \begin{array}{l} -\beta_2 \log s(x \\ -\beta_1 \log t(x \end{array}\right.$$

where $\hat\beta = (\frac{\beta_2}{\beta_1+\beta_2}, \frac{\beta_1}{\beta_1+\beta_2})$ *represents the normalized and reversed vector of conficences* $\beta = (\beta_1, \beta_2)$ *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 $\mathcal{M}$ denote the PDG in question. Since D has high confidence, we know any joint distribution μ with a finite score must have $\mu(X) = D(X)$. Thus,

$$\begin{aligned}
\langle\!\langle \mathcal{M} \rangle\!\rangle &= \inf_{\mu} \mathop{\mathbb{E}}_{x\sim D} \mathop{\mathbb{E}}_{y\sim\mu|x} \left[\beta_1 \log \frac{\mu(y \mid x)}{\mathcal{N}(y \mid f(x), s(x))} + \beta_2 \log \frac{\mu(y \mid x)}{\mathcal{N}(y \mid h(x), t(x))} \right] \\
&= \inf_{\mu} \mathop{\mathbb{E}}_{x\sim D} \mathop{\mathbb{E}}_{y\sim\mu|x} \left[\beta_1 \log \frac{\mu(y \mid x)}{\frac{1}{s(x)\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y-f(x)}{s(x)}\right)^2\right)} + \beta_2 \log \frac{\mu(y \mid x)}{\frac{1}{t(x)\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y-h(x)}{t(x)}\right)^2\right)} \right] \\
&= \inf_{\mu} \mathop{\mathbb{E}}_{x\sim D} \mathop{\mathbb{E}}_{y\sim\mu|x} \left[\log \mu(y \mid x)^{\beta_1+\beta_2} \begin{array}{ll} +\frac{\beta_1}{2}\left(\frac{y-f(x)}{s(x)}\right)^2 & +\frac{\beta_2}{2}\left(\frac{y-h(x)}{t(x)}\right)^2 \\ +\beta_1 \log(s(x)\sqrt{2\pi}) & +\beta_2 \log(t(x)\sqrt{2\pi}) \end{array} \right].
\end{aligned} \tag{5.8}$$

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 (**??** 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 (5.8) leaves us with:

$$\langle\!\langle m \rangle\!\rangle = \inf_{\mu} \mathop{\mathbb{E}}_{x\sim D} \mathop{\mathbb{E}}_{y\sim\mu|x} \left[\begin{array}{lll} & +\frac{1}{2\,\tilde\sigma(x)^2}(y-g(x))^2 & +\frac{1}{2}\frac{\beta_1\beta_2}{\beta_1 t(x)^2+\beta_2 s(x)^2}(f(x)-h(x))^2 \\ \log\mu(y\mid x)^{\beta_1+\beta_2} & & \\ & +\beta_1\log(s(x)\sqrt{2\pi}) & +\beta_2\log(t(x)\sqrt{2\pi}) \end{array} \right]$$

Pulling the term on the top right, 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 φ by $-\log(\exp(-\varphi))$), we obtain:

$$\langle\!\langle m \rangle\!\rangle = \mathop{\mathbb{E}}_{x\sim D} \left[\begin{array}{l} \displaystyle\inf_{\mu(Y)} \mathop{\mathbb{E}}_{y\sim\mu} \left[\log\mu(y)^{\beta_1+\beta_2} - \log\left(\frac{1}{\sqrt{2\pi}^{\beta_1+\beta_2} s(x)^{\beta_1} t(x)^{\beta_2}} \exp\left\{ -\frac{1}{2}\Big(\frac{y-g(x)}{\tilde\sigma(x)}\Big)^2 \right\}\right)\right] \\ \quad + \displaystyle\frac{1}{2}\frac{\beta_1\beta_2}{\beta_1 t(x)^2+\beta_2 s(x)^2}\Big(f(x)-h(x)\Big)^2 \end{array} \right].$$

To simplify the presentation, let ψ be the term on the top right, and ξ be the term on the bottom. More explicitly, define

$$\psi(x,y) := \frac{1}{2}\frac{1}{\sqrt{2\pi}^{\beta_1+\beta_2} s(x)^{\beta_1} t(x)^{\beta_2}} \exp\left\{ -\frac{1}{2}\Big(\frac{y-g(x)}{\tilde\sigma(x)}\Big)^2 \right\}, \text{ and } \xi(x) := \frac{\beta_1\beta_2}{\beta_1 t(x)^2+\beta_2 s(x)^2}\Big(f(x)-h$$

which lets us write the previous expression for $\langle\!\langle m \rangle\!\rangle$ as

$$\langle\!\langle m \rangle\!\rangle = \mathop{\mathbb{E}}_{x\sim D}\left[\inf_{\mu(Y)} \mathop{\mathbb{E}}_{y\sim\mu} \Big[\log\mu(y)^{\beta_1+\beta_2} - \log\psi(x,y) \Big] + \xi(x) \right]. \qquad (5.9)$$

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

$$\begin{aligned} \psi(x,y)^{\frac{1}{\beta_1+\beta_2}} &= (2\pi)^{-\frac{1}{2}}\, s(x)^{\left(\frac{-\beta_1}{\beta_1+\beta_2}\right)} t(x)^{\left(\frac{-\beta_2}{\beta_1+\beta_2}\right)} \exp\left\{ -\frac{1}{2}\Big(\frac{y-g(x)}{\tilde\sigma(x)}\Big)^2 \right\}^{\frac{1}{\beta_1+\beta_2}} \\ &= \frac{1}{\sqrt{2\pi}\, s(x)^{\hat\beta_1} t(x)^{\hat\beta_2}} \exp\left\{ \frac{-1}{2(\beta_1+\beta_2)}\Big(\frac{y-g(x)}{\tilde\sigma(x)}\Big)^2 \right\}. \end{aligned}$$

Recall that the Gaussian densitity $\mathcal{N}(y|\, g(x),\, \tilde\sigma(x)\sqrt{\beta_1+\beta_2})$ of mean $g(x)$ and variance $\tilde\sigma(x)^2(\beta_1+\beta_2)$ is given by

$$\mathcal{N}\left(y \,\middle|\, g(x),\, \tilde\sigma(x)\sqrt{\beta_1+\beta_2}\right) = \frac{1}{\sqrt{2\pi}\ \tilde\sigma(x)\sqrt{\beta_1+\beta_2}} \exp\left\{ \frac{-1}{2(\beta_1+\beta_2)}\Big(\frac{y-g(x)}{\tilde\sigma(x)}\Big)^2 \right\},$$

which is quite similar, and has an identical dependence on y. To facilitate converting one to the other, we explicitly compute the ratio:

$$\begin{aligned}
\frac{\psi(x,y)^{\frac{1}{\beta_1+\beta_2}}}{\mathcal{N}\left(y \,\middle|\, g(x),\ \tilde{\sigma}(x)\sqrt{\beta_1+\beta_2}\right)} &= \frac{\tilde{\sigma}\sqrt{2\pi(\beta_1+\beta_2)}}{\sqrt{2\pi}\ s(x)^{\hat{\beta}_1} t(x)^{\hat{\beta}_2}} = \frac{\tilde{\sigma}\sqrt{\beta_1+\beta_2}}{s(x)^{\hat{\beta}_1} t(x)^{\hat{\beta}_2}} \\
&= \left(\frac{s(x)t(x)}{\sqrt{\beta_1 t(x)^2 + \beta_2 s(x)^2}} \right) \frac{\sqrt{\beta_1+\beta_2}}{s(x)^{\hat{\beta}_1} t(x)^{\hat{\beta}_2}} && \text{[expand defn of } \tilde{\sigma}(x)\text{]} \\
&= s(x)^{1-\hat{\beta}_1}\, t(x)^{1-\hat{\beta}_2} \sqrt{\frac{\beta_1+\beta_2}{\beta_1\, t(x)^2 + \beta_2\, s(x)^2}} \\
&= s(x)^{1-\hat{\beta}_1}\, t(x)^{1-\hat{\beta}_2} \sqrt{\frac{1}{\hat{\beta}_1\, t(x)^2 + \hat{\beta}_2\, s(x)^2}} && \text{[defn of } \hat{\beta}_1, \hat{\beta}_2\text{]} \\
&= \frac{s(x)^{\hat{\beta}_2}\, t(x)^{\hat{\beta}_1}}{\sqrt{\hat{\beta}_1\, t(x)^2 + \hat{\beta}_2\, s(x)^2}} && \text{[since } \hat{\beta}_1 + \hat{\beta}_2 = 1\text{]}
\end{aligned}$$

Now, picking up from where we left off in (5.9), we have

$$\begin{aligned}
\langle\!\langle m \rangle\!\rangle &= \mathop{\mathbb{E}}_{x\sim D} \left[\inf_{\mu(Y)} \mathop{\mathbb{E}}_{y\sim\mu} \left[\log \mu(y)^{\beta_1+\beta_2} - \log \psi(x,y) \right] + \xi(x) \right] \\
&= \mathop{\mathbb{E}}_{x\sim D} \left[\inf_{\mu(Y)} \mathop{\mathbb{E}}_{y\sim\mu} \left[\log \frac{\mu(y)^{\beta_1+\beta_2}}{\psi(x,y)^{\frac{\beta_1+\beta_2}{\beta_1+\beta_2}}} \right] + \xi(x) \right] \\
&= \mathop{\mathbb{E}}_{x\sim D} \left[\inf_{\mu(Y)} \mathop{\mathbb{E}}_{y\sim\mu} \left[(\beta_1+\beta_2) \log \frac{\mu(y)}{\psi(x,y)^{\frac{1}{\beta_1+\beta_2}}} \right] + \xi(x) \right] \\
&= \mathop{\mathbb{E}}_{x\sim D} \left[\inf_{\mu(Y)} \mathop{\mathbb{E}}_{y\sim\mu} \left[(\beta_1+\beta_2) \log \frac{\mu(y)}{\mathcal{N}\left(y \,\middle|\, g(x),\ \tilde{\sigma}(x)\sqrt{\beta_1+\beta_2}\right) \frac{s(x)^{\hat{\beta}_2}\, t(x)^{\hat{\beta}_1}}{\sqrt{\hat{\beta}_1\, t(x)^2 + \hat{\beta}_2\, s(x)^2}}} \right] + \xi(x) \right] \\
&= \mathop{\mathbb{E}}_{x\sim D} \left[\inf_{\mu(Y)} \mathop{\mathbb{E}}_{y\sim\mu} \left[(\beta_1+\beta_2) \log \frac{\mu(y)}{\mathcal{N}\left(y \,\middle|\, g(x),\ \tilde{\sigma}(x)\sqrt{\beta_1+\beta_2}\right)} \right] + (\beta_1+\beta_2) \log \frac{\sqrt{\hat{\beta}_1\, t(x)^2 + \hat{\beta}_2\, s(x)^2}}{s(x)^{\hat{\beta}_2}\, t(x)^{\hat{\beta}_1}} \right.
\end{aligned}$$

but now the entire left term is the infemum of a KL divergence, 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 infemum on the left is equal to zero.

$$\langle\!\langle m \rangle\!\rangle = \mathop{\mathbb{E}}_{x\sim D} \left[(\beta_1+\beta_2) \log \frac{\sqrt{\hat{\beta}_1\, t(x)^2 + \hat{\beta}_2\, s(x)^2}}{s(x)^{\hat{\beta}_2}\, t(x)^{\hat{\beta}_1}} + \xi(x) \right] \tag{5.10}$$

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{x\sim D}\left[(\beta_1{+}\beta_2)\log\sqrt{\hat\beta_1\, t(x)^2+\hat\beta_2\, s(x)^2}-(\beta_1{+}\beta_2)\log\Big(s(x)^{\hat\beta_2}\, t(x)^{\hat\beta_1}\Big)+\xi(x)\right]\\
&= \mathop{\mathbb{E}}_{x\sim D}\left[(\beta_1{+}\beta_2)\log\sqrt{\hat\beta_1\, t(x)^2+\hat\beta_2\, s(x)^2}\;\begin{array}{l}-\beta_2\log s(x)\\ -\beta_1\log t(x)\end{array}\;+\xi(x)\right]\\
&= \mathop{\mathbb{E}}_{x\sim D}\left[(\beta_1{+}\beta_2)\log\sqrt{\frac{\beta_1\, t(x)^2+\beta_2\, s(x)^2}{\beta_1+\beta_2}}\;\begin{array}{l}-\beta_2\log s(x)\\ -\beta_1\log t(x)\end{array}\;+\frac{1}{2}\frac{\beta_1\beta_2}{\beta_1 t(x)^2+\beta_2 s(x)^2}\Big(f(x)-h(x)\Big)^2\right.
\end{aligned}
\tag{5.11}
$$

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 σ_1 (the random variable) in place of $s(x)$ and σ_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 σ_1 and σ_2 are:

$$
\mathrm{GM}_{\hat\beta}(\sigma_1,\sigma_2)=(\sigma_1)^{\hat\beta_2}(\sigma_2)^{\hat\beta_1}\qquad\text{and}\qquad \mathrm{QM}_{\hat\beta}(\sigma_1,\sigma_2)=\sqrt{\hat\beta_2\sigma_1^2+\hat\beta_1\sigma_2^2}.
$$

So, now we can write $\xi(x)$ as

$$
\begin{aligned}
\frac{1}{2}\frac{\beta_1\beta_2}{\beta_1 t(x)^2+\beta_2 s(x)^2}\Big(f(x)-h(x)\Big)^2 &= \frac{1}{2}\frac{\beta_1\beta_2}{\beta_1+\beta_2}\frac{\beta_1+\beta_2}{\beta_1 t(x)^2+\beta_2 s(x)^2}\Big(f(x)-h(x)\Big)^2\\
&= \frac{1}{2}\frac{\beta_1\beta_2}{\beta_1+\beta_2}\left(\frac{1}{\mathrm{QM}_{\hat\beta}(\sigma_1,\sigma_2)}\right)^2\Big(f(x)-h(x)\Big)^2\\
&= \frac{1}{2}\frac{\beta_1\beta_2}{\beta_1+\beta_2}\left(\frac{\mu_1-\mu_2}{\mathrm{QM}_{\hat\beta}(\sigma_1,\sigma_2)}\right)^2;
\end{aligned}
$$

in the last step, we have replaced $f(x)$ and $g(x)$ with their respective random variables μ_1 and μ_2. As a result, (5.10) can be written as

$$
\langle\!\langle m\rangle\!\rangle = \mathop{\mathbb{E}}_{D}\left[(\beta_1{+}\beta_2)\log\frac{\mathrm{QM}_{\hat\beta}(\sigma_1,\sigma_2)}{\mathrm{GM}_{\hat\beta}(\sigma_1,\sigma_2)}+\frac{1}{2}\frac{\beta_1\beta_2}{\beta_1+\beta_2}\left(\frac{\mu_1-\mu_2}{\mathrm{QM}_{\hat\beta}(\sigma_1,\sigma_2)}\right)^2\right]
$$

...which is perhaps more comprehensible, and proves the first half of our proposition. □

Claim 2. *The sum of two functions that are unshifted parabolas as functions of y (i.e., both functions are of 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

$$g:=\frac{\beta_1\sigma_2^2 f+\beta_2\sigma_1^2 h}{\beta_1\sigma_2^2+\beta_2\sigma_1^2}\quad\textit{and}\quad\tilde{\sigma}:=\left(\frac{\beta_1}{\sigma_1^2}+\frac{\beta_2}{\sigma_2^2}\right)^{-1/2}=\frac{\sigma_1\sigma_2}{\sqrt{\beta_1\sigma_2^2+\beta_2\sigma_1^2}}.$$

Proof. Expand terms and complete the square. Starting from the left hand side, we have

$$\begin{aligned}
&\frac{\beta_1}{\sigma_1^2}(y-f)^2+\frac{\beta_2}{\sigma_2^2}(y-h)^2\\
&=\frac{\beta_1}{\sigma_1^2}(y^2-2yf+f^2)+\frac{\beta_2}{\sigma_2^2}(y^2-2yh+h^2)\\
&=\left(\frac{\beta_1}{\sigma_1^2}+\frac{\beta_2}{\sigma_2^2}\right)y^2-2\left(\frac{\beta_1 f}{\sigma_1^2}+\frac{\beta_2 h}{\sigma_2^2}\right)y+\left(\frac{\beta_1 f^2}{\sigma_1^2}+\frac{\beta_2 h^2}{\sigma_2^2}\right)\\
&=\left(\frac{\beta_1}{\sigma_1^2}+\frac{\beta_2}{\sigma_2^2}\right)y^2-2\left(\frac{\beta_1 f}{\sigma_1^2}+\frac{\beta_2 h}{\sigma_2^2}\right)y+\left(\frac{\beta_1 f^2}{\sigma_1^2}+\frac{\beta_2 h^2}{\sigma_2^2}-\frac{\beta_1\beta_2(f-h)^2}{\beta_1\sigma_2^2+\beta_2\sigma_1^2}\right)+\frac{\beta_1\beta_2(f-h)^2}{\beta_1\sigma_2^2+\beta_2\sigma_1^2}
\end{aligned}\tag{5.12}$$

where in the last step we added and removed the same term (i.e., the completion of the square, although it is probably still unclear why this quantity will do that). The third parenthesized quantity needs the most work. Isolating it and getting a common denominator gives us:

$$\begin{aligned}
&\frac{\beta_1\, f^2}{\sigma_1^2}+\frac{\beta_2\, h^2}{\sigma_2^2}-\frac{\beta_1\beta_2(f-h)^2}{\beta_1\sigma_2^2+\beta_2\sigma_1^2}\\
&=\frac{\beta_1\, f^2(\beta_1\sigma_2^2+\beta_2\sigma_1^2)\sigma_2^2}{\sigma_1^2(\beta_1\sigma_2^2+\beta_2\sigma_1^2)\sigma_2^2}+\frac{\beta_2\, h^2(\beta_1\sigma_2^2+\beta_2\sigma_1^2)\sigma_1^2}{\sigma_2^2(\beta_1\sigma_2^2+\beta_2\sigma_1^2)\sigma_1^2}-\frac{\beta_1\beta_2(f^2-2fh+h^2)(\sigma_1^2\sigma_2^2)}{(\beta_1\sigma_2^2+\beta_2\sigma_1^2)(\sigma_1^2\sigma_2^2)}\\
&=\frac{\beta_1^2\sigma_2^4 f^2+\cancel{\beta_1\beta_2\sigma_2^2\sigma_1^2 f^2}+\cancel{\beta_1\beta_2\sigma_1^2\sigma_2^2 h^2}+\beta_2^2\sigma_1^4 h^2-\cancel{\beta_1\beta_2\sigma_1^2\sigma_2^2 f^2}+2\beta_1\beta_2\sigma_1^2\sigma_2^2 fh-\cancel{\beta_1\beta_2\sigma_1^2\sigma_2^2 h^2}}{(\beta_1\sigma_2^2+\beta_2\sigma_1^2)(\sigma_1^2\sigma_2^2)}
\end{aligned}$$

$$= \frac{\beta_1^2\sigma_2^4 f^2 + \beta_2^2\sigma_1^4 h^2 + 2\beta_1\beta_2\sigma_1^2\sigma_2^2 fh}{(\beta_1\sigma_2^2 + \beta_2\sigma_1^2)(\sigma_1^2\sigma_2^2)}.$$

Substituting this expression into the third term of (5.12), while simultaneously computing common denominators for the first and second terms, yields

$$\left(\frac{\beta_1\sigma_2^2 + \beta_2\sigma_1^2}{\sigma_1^2\sigma_2^2}\right)y^2 - 2\left(\frac{\beta_1\sigma_2^2 f + \beta_2\sigma_1^2 h}{\sigma_1^2\sigma_2^2}\right)y + \frac{\beta_1^2\sigma_2^4 f^2 + \beta_2^2\sigma_1^4 h^2 + 2\beta_1\beta_2\sigma_1^2\sigma_2^2 fh}{(\beta_1\sigma_2^2 + \beta_2\sigma_1^2)(\sigma_1^2\sigma_2^2)} + \frac{\beta_1\beta_2(f-h)^2}{\beta_1\sigma_2^2 + \beta_2\sigma_1^2}. \tag{5.13}$$

On the other hand, using the definitions of g and $\tilde{\sigma}$, we compute:

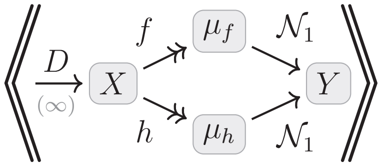
$$\begin{aligned}
\left(\frac{y-g}{\tilde{\sigma}}\right)^2 &= \left(\frac{\beta_1\sigma_2^2 + \beta_2\sigma_1^2}{\sigma_1^2\sigma_2^2}\right)\left(y - \frac{\beta_1\sigma_2^2 f + \beta_2\sigma_1^2 h}{\beta_1\sigma_2^2 + \beta_2\sigma_1^2}\right)^2 \\
&= \left(\frac{\beta_1\sigma_2^2 + \beta_2\sigma_1^2}{\sigma_1^2\sigma_2^2}\right)\left(y^2 - 2y\frac{\beta_1\sigma_2^2 f + \beta_2\sigma_1^2 h}{\beta_1\sigma_2^2 + \beta_2\sigma_1^2} + \frac{\beta_1^2\sigma_2^4 f^2 + \beta_2^2\sigma_1^4 h^2 + 2\beta_1\beta_2\sigma_2^2\sigma_1^2\, fh}{(\beta_1\sigma_2^2 + \beta_2\sigma_1^2)^2}\right) \\
&= \left(\frac{\beta_1\sigma_2^2 + \beta_2\sigma_1^2}{\sigma_1^2\sigma_2^2}\right)y^2 - 2\left(\frac{\beta_1\sigma_2^2 f + \beta_2\sigma_1^2 h}{\sigma_1^2\sigma_2^2}\right)y + \frac{\beta_1^2\sigma_2^4 f^2 + \beta_2^2\sigma_1^4 h^2 + 2\beta_1\beta_2\sigma_2^2\sigma_1^2\, fh}{(\beta_1\sigma_2^2 + \beta_2\sigma_1^2)(\sigma_1^2\sigma_2^2)}
\end{aligned}$$

. . . which is precisely the first 3 terms of (5.13). Putting it all together, we have shown that

$$\frac{\beta_1}{\sigma_1^2}(y-f)^2 + \frac{\beta_2}{\sigma_2^2}(y-h)^2 = \left(\frac{y-g}{\tilde{\sigma}}\right)^2 + \frac{\beta_1\beta_2(f-h)^2}{\beta_1\sigma_2^2 + \beta_2\sigma_1^2}$$

as desired. □

Proposition 5.3.6.



$$= \frac{1}{2}\,\mathbb{E}_D\big|f(X) - h(X)\big|^2$$
$$=: \mathrm{MSE}_D(f,h)\,,$$

where $\mathcal{N}_1(Y\,|\,\mu)$ is a unit Gaussian on Y with mean μ.

Proof. An immediate corolary of Proposition 5.B.1; simply set $s(x) = t(x) = \beta_1 = \beta_2 = 1$ □

Lemma 5.5.2. *The inconsistency $\boldsymbol{D}^{\mathrm{PDG}}_{(r,s)}(p\|q)$ of a PDG comprising $p(X)$ with confidence r and $q(X)$ with confidence s is given in closed form by*

$$\left\langle\!\!\left\langle \xrightarrow[(r)]{p} \boxed{X} \xleftarrow[(s)]{q} \right\rangle\!\!\right\rangle = -(r+s)\log\sum_x \Big(p(x)^r q(x)^s\Big)^{\frac{1}{r+s}} .$$

Proof.

$$\begin{aligned}
\left\langle\!\!\left\langle \xrightarrow[(\beta:r)]{p} \boxed{X} \xleftarrow[(\beta:s)]{q} \right\rangle\!\!\right\rangle &= \inf_{\mu} \mathop{\mathbb{E}}_{\mu} \log \frac{\mu(x)^{r+s}}{p(x)^r q(x)^s} \\
&= (r+s) \inf_{\mu} \mathop{\mathbb{E}}_{\mu} \left[\log \frac{\mu(x)}{(p(x)^r q(x)^s)^{\frac{1}{r+s}}} \cdot \frac{Z}{Z} \right] \\
&= \inf_{\mu} (r+s) \boldsymbol{D}\left(\mu \;\middle\|\; \frac{1}{Z} p^{\frac{r}{r+s}} q^{\frac{s}{r+s}} \right) - (r+s)\log Z
\end{aligned}$$

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 μ. $\boldsymbol{D}(\mu \parallel \cdots)$ achives its minimum of zero when μ is the second distribution, so our formula becomes

$$\begin{aligned}
&= -(r+s)\log Z \\
&= -(r+s)\log\sum_x \Big(p(x)^r q(x)^s\Big)^{\frac{1}{r+s}} \qquad \text{as promised.}
\end{aligned}$$

□

Proposition 5.4.1. *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,*

$$\left\langle\!\!\left\langle \begin{tikzcd} \arrow[dr, "q", "(\beta)"'] & & \\ & \boxed{\Theta} \arrow[r, "p"] & \boxed{Y} \\ \arrow[ur, twoheadrightarrow, "\theta"'] & & D \arrow[u, "(\infty)"'] \end{tikzcd} \right\rangle\!\!\right\rangle = \mathop{\mathbb{E}}_{y\sim D} \log \frac{1}{p(y\,|\,\theta)} + \beta \log \frac{1}{q(\theta)} - \mathrm{H}(D) \tag{5.4}$$

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 θ and D, then, are satisfied by this μ 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:

$$\begin{aligned}
\langle\!\langle \mathcal{m} \rangle\!\rangle &= \mathop{\mathbb{E}}_{\mu}\left[\log\frac{\mu(Y|\Theta)}{p(Y|\Theta)} + \beta\log\frac{\mu(\Theta)}{q(\Theta)}\right] \\
&= \mathop{\mathbb{E}}_{y\sim D}\left[\log\frac{D(y)}{p(y|\theta)} + \beta\log\frac{1}{q(\theta)}\right] \\
&= \mathop{\mathbb{E}}_{y\sim D}\left[\log\frac{1}{p(y|\theta)} + \beta\log\frac{1}{q(\theta)} + \log D(y)\right] \\
&= \mathop{\mathbb{E}}_{y\sim D}\left[\log\frac{1}{p(y|\theta)}\right] + \beta\log\frac{1}{q(\theta)} - \mathrm{H}(D)
\end{aligned}$$

as desired. □

Proposition 5.6.1. *The negative ELBO of x is the inconsistency of the PDG containing p,q, and $X{=}x$, with high confidence in q. That is,*

$$-\mathrm{ELBO}_{p,q}(x) = \left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "q", "(\infty)"'] & \boxed{Z} \arrow[r, "p"] & \boxed{X} & {} \arrow[l, two heads, "x"'] \end{tikzcd} \right\rangle\!\!\right\rangle.$$

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,

$$\left\langle\!\!\left\langle \xrightarrow[(\infty)]{q} \boxed{Z} \overset{p}{\swarrow\searrow} \boxed{X} \overset{x}{\twoheadleftarrow} \right\rangle\!\!\right\rangle = \inf_{\mu} \left[\!\!\left[\xrightarrow[(\infty)]{q} \boxed{Z} \overset{p}{\swarrow\searrow} \boxed{X} \overset{x}{\twoheadleftarrow} \right]\!\!\right](\mu)$$

$$= \left[\!\!\left[\xrightarrow[(\infty)]{q} \boxed{Z} \overset{p}{\swarrow\searrow} \boxed{X} \overset{x}{\twoheadleftarrow} \right]\!\!\right](\delta_x(X)q(Z))$$

$$= \mathop{\mathbb{E}}_{\substack{x'\sim\delta_x \\ z\sim q}} \log \frac{\delta_x(x')q(z)}{p(x',z)} = - \mathop{\mathbb{E}}_{z\sim q} \frac{p(x,z)}{q(z)} = -\mathrm{ELBO}_{p,q}(x).$$

□

We prove both Proposition 5.6.2 and Proposition 5.B.2 at the same time.

Proposition 5.6.2. *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,*

$$-\mathrm{ELBO}_{p,e,d}(x) = \left\langle\!\!\left\langle \xrightarrow{p} \boxed{Z} \underset{\underset{(\infty)}{e}}{\overset{d}{\rightleftarrows}} \boxed{X} \overset{x}{\twoheadleftarrow} \right\rangle\!\!\right\rangle.$$

Proposition 5.B.2. *The following analog of Proposition 5.6.2 for a whole dataset $\mathcal{D}$ holds:*

$$-\mathop{\mathbb{E}}_{\Pr_{\mathcal{D}}} \mathrm{ELBO}_{p,e,d}(X) = \left\langle\!\!\left\langle \xrightarrow{p} \boxed{Z} \underset{\underset{e}{(\infty)}}{\overset{d}{\rightleftarrows}} \boxed{X} \xleftarrow[(\infty)]{\Pr_{\mathcal{D}}} \right\rangle\!\!\right\rangle + \mathrm{H}(\Pr_{\mathcal{D}}).$$

Proof. The two proofs are similar. For Proposition 5.6.2, the optimal distribution must be $\delta_x(X)e(Z \mid X)$, and for Proposition 5.B.2, 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

$$\boldsymbol{D}\Big(\delta_x(X)e(Z \mid X) \,\Big\|\, p(Z)d(X \mid Z)\Big) = \mathop{\mathbb{E}}_{z\sim e|x} \left[\log \frac{p(z)d(x \mid z)}{e(z \mid x)}\right] = \mathrm{ELBO}_{p,e,d}(x)$$

in the first, case, and

$$\boldsymbol{D}\Big(\Pr_{\mathcal{D}}(X)e(Z \mid X) \,\Big\|\, p(Z)d(X \mid Z)\Big) = \frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}} \mathop{\mathbb{E}}_{z\sim e|x}\left[\log\frac{p(z)d(x \mid z)}{e(z \mid x)} + \log\frac{1}{\Pr_{\mathcal{D}}(x)}\right]$$

$$= \mathrm{ELBO}_{p,e,d}(x) - \mathrm{H}(\Pr_{\mathcal{D}})$$

in the second. □

Now, we formally state and prove the more general result for β-VAEs.

Proposition 5.B.3. *The negative β-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 β. That is,*

$$-\beta\text{-ELBO}_{p,e,d}(x) = \left\langle\!\!\!\left\langle \begin{tikzcd} {} \arrow[r, "p", "(\beta)"'] & \boxed{Z} \arrow[r, bend left, "d"] & \boxed{X} \arrow[l, bend left, "e", "(\infty)"'] & {} \arrow[l, two heads, "x"'] \end{tikzcd} \right\rangle\!\!\!\right\rangle$$

Proof.

$$\left\langle\!\!\!\left\langle \begin{tikzcd} {} \arrow[r, "p", "(\beta)"'] & \boxed{Z} \arrow[r, bend left, "d"] & \boxed{X} \arrow[l, bend left, "e", "(\infty)"'] & {} \arrow[l, two heads, "x"'] \end{tikzcd} \right\rangle\!\!\!\right\rangle = \inf_{\mu} \left[\!\!\left[\begin{tikzcd} {} \arrow[r, "p", "(\beta)"'] & \boxed{Z} \arrow[r, bend left, "d"] & \boxed{X} \arrow[l, bend left, "e", "(\infty)"'] & {} \arrow[l, two heads, "x"'] \end{tikzcd} \right]\!\!\right](\mu)$$

$$= \inf_{\mu} \mathop{\mathbb{E}}_{\mu(X,Z)}\left[\beta\log\frac{\mu(Z)}{p(Z)} + \log\frac{\mu(X,Z)}{\mu(Z)d(X \mid Z)}\right]$$

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 $\int_x \delta_x(X)e(Z \mid X)\,\mathrm{d}x = e(Z \mid x)$. Thus, our equation becomes

$$= \mathop{\mathbb{E}}_{\delta_x(X)e(Z|X)}\left[\beta\log\frac{e(Z \mid x)}{p(z)} + \log\frac{\delta_x(X)e(Z \mid X)}{e(Z \mid x)d(x \mid Z)}\right]$$

$$= \mathop{\mathbb{E}}_{e(Z|x)}\left[\beta\log\frac{e(Z \mid x)}{p(Z)} + \log\frac{1}{d(x \mid Z)}\right]$$

$$= \boldsymbol{D}(e(Z|\,x) \parallel p) + \mathrm{Rec}_{e,d}(x)$$

$$= -\beta\text{-ELBO}_{p,e,d}(x).$$

□

Proposition 5.7.1. *For all weighted factor graphs* Ψ*, we have that* $\langle\!\langle \boldsymbol{\mathit{pdg}}(\Psi) \rangle\!\rangle_1 = -\log Z_\Psi$.

Proof. In the main text, we defined $\boldsymbol{\mathit{pdg}}(\Psi)$ to be the PDG with edges $\{\xrightarrow{J}\mathbf{X}_J\}_{\mathcal{J}}$, cpds $p_J(\mathbf{X}_J) \propto \phi_J(\mathbf{X}_J)$, and weights $\alpha_J, \beta_J := \theta_J$. Let $\mathrm{the}(\{x\}) := x$ be a function that extracts the unique element singleton set. It was shown by **?**] (Corolary 4.4.1) that

$$\mathrm{the}\llbracket \mathcal{m}_\Psi \rrbracket_1^* = \Pr_{\Phi,\theta}(\mathbf{x}) = \frac{1}{Z_\Psi} \prod_J \phi_J(\mathbf{x}_J)^{\theta_J}.$$

Recall the statement of Prop 4.6 from **?**]:

$$\llbracket \mathcal{m} \rrbracket_\gamma(\mu) = \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu} \left\{ \sum_{X \xrightarrow{L} Y} \left[\beta_L \log \frac{1}{\mathbb{P}_L(y^{\mathbf{w}}|x^{\mathbf{w}})} + (\gamma\alpha_L - \beta_L) \log \frac{1}{\mu(y^{\mathbf{w}}|x^{\mathbf{w}})} \right] - \gamma \log \frac{1}{\mu(\mathbf{w})} \right\}, \tag{5.14}$$

where $x^{\mathbf{w}}$ and $y^{\mathbf{w}}$ are the respective values of the variables X and Y in the world w. Note that if $\gamma = 1$, and α, β are both equal to θ in $\boldsymbol{\mathit{pdg}}(\Psi)$, the middle term (in purple) is zero. So in our case, since the edges are $\{\xrightarrow{J} \mathbf{X}_J\}$ and $\mathbb{P}_J(\mathbf{X}_J) = \phi_J(\mathbf{X}_J)$, (5.14) reduces to the standard variational free energy

$$\mathit{VFE}_\Psi(\mu) = \mathop{\mathbb{E}}_{\mu} \left[\sum_{J \in \mathcal{J}} \theta_J \log \frac{1}{\phi_J(\mathbf{X}_J)} \right] - \mathrm{H}(\mu) \tag{5.15}$$

$$= \mathop{\mathbb{E}}_{\mu} \big\langle \boldsymbol{\varphi}, \boldsymbol{\theta} \big\rangle_{\mathcal{J}} - \mathrm{H}(\mu), \quad \text{where } \varphi_J(\mathbf{X}_J) := \log \frac{1}{\phi_J(\mathbf{X}_J)}.$$

By construction, $\Pr_\Psi$ uniquely minimizes VFE. The 1-inconsistency, $\langle\!\langle \mathcal{m}_\Psi \rangle\!\rangle$ is the minimum value attained. We calculate:

$$\langle\!\langle \mathcal{m} \rangle\!\rangle_1 = \mathit{VFE}_\Psi(\Pr_\Psi)$$

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} \left\{ \sum_{J\in\mathcal{J}} \left[\theta_J \log \frac{1}{\phi_J(\mathbf{x}_J)} \right] - \log \frac{1}{\Pr_{\Phi,\theta}(\mathbf{x})} \right\} && \Big[\text{by (5.15)}\Big] \\
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} \left\{ \sum_{J\in\mathcal{J}} \left[\theta_J \log \frac{1}{\phi_J(\mathbf{x}_J)} \right] - \log \frac{Z_\Psi}{\prod_{J\in\mathcal{J}} \phi_J(\mathbf{x}_J)^{\theta_j}} \right\} && \Big[\text{definition of } \Pr_{\Psi}\Big] \\
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} \left\{ \sum_{J} \left[\theta_J \log \frac{1}{\phi_J(\mathbf{x}_J)} \right] - \sum_{J\in\mathcal{J}} \left[\theta_J \log \frac{1}{\phi_J(\mathbf{x}_J)} \right] - \log Z_\Psi \right\} \\
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} [-\log Z_\Psi] \\
&= -\log Z_\Psi && \Big[Z_\Psi \text{ is constant in } \mathbf{x} \Big]
\end{aligned}
$$

□

Proposition 5.9.1. $\left\langle\!\!\left\langle \xrightarrow[(\infty)]{p} \boxed{X} \xrightarrow{\hat{c}} \boxed{\mathtt{T}} \overset{\mathtt{t}}{\twoheadleftarrow} \right\rangle\!\!\right\rangle = \mathop{\mathbb{E}}_{x\sim p} c(x).$

Proof. Since p has high confidence, and $\mathtt{T}$ is always equal to $\mathtt{t}$, the only joint distribution on $(X, \mathtt{T})$ with finite score is $\mu(X, \mathtt{T}) = p(X)\delta_{\mathtt{t}}(\mathtt{T})$. We compute its score directly:

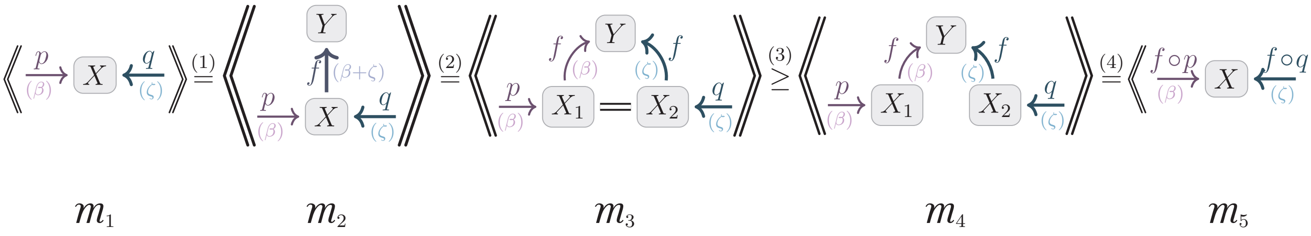
$$
\begin{aligned}
\left\langle\!\!\left\langle \xrightarrow[(\infty)]{p} \boxed{X} \xrightarrow{\hat{c}} \boxed{\mathtt{T}} \overset{\mathtt{t}}{\twoheadleftarrow} \right\rangle\!\!\right\rangle &= \mathop{\mathbb{E}}_{\mu} \log \frac{\mu(X,\mathtt{T})}{\hat{c}(\mathtt{t}\,|X)} = \mathop{\mathbb{E}}_{p} \log \frac{1}{\hat{c}(\mathtt{t}\,|X)} = \mathop{\mathbb{E}}_{p} \log \frac{1}{\exp(-c(X))} \\
&= \mathop{\mathbb{E}}_{p} \log \exp(c(X)) = \mathop{\mathbb{E}}_{p} c(X) = \mathop{\mathbb{E}}_{x\sim p} c(x).
\end{aligned}
$$

□

5.C.1 Additional Proofs for Unnumbered 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 2 of the main text. We repeat it now for convenience, with labeled PDGs ($\mathit{m}_1, \ldots, \mathit{m}_5$) and numbered (in)equalities.



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

$$
\begin{aligned}
\mathit{Inc}_{\mathit{m}_2}(\mu(X,Y)) &= \beta \boldsymbol{D}(\mu(X) \,\|\, p(X)) + \zeta \boldsymbol{D}(\mu(X) \,\|\, q(X)) + (\beta{+}\zeta) \mathop{\mathbb{E}}_{x\sim\mu} \big[\boldsymbol{D}(\mu(Y|x) \,\|\, f(Y|x))\\
&= \mathit{Inc}_{\mathit{m}_1}(\mu(X)) + (\beta{+}\zeta) \mathop{\mathbb{E}}_{x\sim\mu} \boldsymbol{D}(\mu(Y|x) \,\|\, f(Y|x))\\
&= \langle\!\langle \mathit{m}_1 \rangle\!\rangle \qquad \Big[\text{as } \mu(Y|x) = f(Y|x) \text{ wh} \\
&\qquad\qquad\qquad \text{and } \mu(X) \text{ minim}
\end{aligned}
$$

So $\mu(X,Y)$ witnesses the fact that $\langle\!\langle \mathit{m}_2 \rangle\!\rangle \leq \mathit{Inc}_{\mathit{m}_2}(\mu(X,Y)) = \langle\!\langle \mathit{m}_1 \rangle\!\rangle$. 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 $\mathit{Inc}_{\mathit{m}_1}(\nu(X)) \geq \mathit{Inc}_{\mathit{m}_1}(\mu(X)) = \langle\!\langle \mathit{m}_1 \rangle\!\rangle$. And since this is true for all $\nu(X,Y)$, we have that $\langle\!\langle \mathit{m}_2 \rangle\!\rangle \geq \langle\!\langle \mathit{m}_1 \rangle\!\rangle$. So $\langle\!\langle \mathit{m}_2 \rangle\!\rangle = \langle\!\langle \mathit{m}_1 \rangle\!\rangle$.

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 μ 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 $\mathit{Inc}\mathit{m}_3$ must be isomorphic to a distribution $\mu'(X, Y)$.

   Furthermore, it is easy to verify that $\mathit{Inc}_{\mathit{m}_2}(\mu'(X,Y)) = \mathit{Inc}_{\mathit{m}_3}(\mu(X,X,Y))$. More formally, we have:

   $$\langle\!\langle \mathit{m}_3 \rangle\!\rangle = \inf_{\mu(X_1,X_2,Y)} \mathop{\mathbb{E}}_{\mu} \left[\beta \log \frac{\mu(X_1)}{p(X_1)} + \zeta \log \frac{\mu(X_2)}{q(X_2)} + \beta \log \frac{\mu(Y|X_1)}{f(Y|X_1)} + \zeta \log \frac{\mu(Y|X_2)}{f(Y|X_2)} + \log \frac{\mu(X}{eq(X} \right.$$

   but if X_1 always equals X_2 (which we call simply X), as it must for the optimal μ, this becomes

   $$\begin{aligned} &= \inf_{\mu(X_1=X_2=X,Y)} \mathop{\mathbb{E}}_{\mu} \left[\beta \log \frac{\mu(X)}{p(X)} + \zeta \log \frac{\mu(X)}{q(X)} + \beta \log \frac{\mu(Y|X)}{f(Y|X)} + \zeta \log \frac{\mu(Y|X)}{f(Y|X)} \right] \\ &= \inf_{\mu(X,Y)} \mathop{\mathbb{E}}_{\mu} \left[\beta \log \frac{\mu(X)}{p(X)} + \zeta \log \frac{\mu(X)}{q(X)} + (\beta{+}\zeta) \log \frac{\mu(Y|X)}{f(Y|X)} \right] \\ &= \inf_{\mu(X,Y)} \mathit{Inc}_{\mathit{m}_2}(\mu) \\ &= \langle\!\langle \mathit{m}_2 \rangle\!\rangle. \end{aligned}$$

3. Eliminating the edge or edges enforcing the equality $(X_1 = X_2)$ cannot increase inconsistency, by Lemma 5.2.1.

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:

   $$\langle\!\langle \mathit{m}_4 \rangle\!\rangle = \inf_{\mu(X,Z,Y)} \mathop{\mathbb{E}}_{\mu} \left[\beta \log \frac{\mu(X)\,\mu(Y|X)}{p(X)\,f(Y|X)} + \zeta \log \frac{\mu(Z)\,\mu(Y|Z)}{q(Z)\,f(Y|Z)} \right]$$

$$= \inf_{\mu(X,Z,Y)} \mathop{\mathbb{E}}_{\mu} \left[\beta \log \frac{\mu(Y)\,\mu(X|Y)}{p(X)\,f(Y|X)} + \zeta \log \frac{\mu(Y)\,\mu(Z|Y)}{q(Z)\,f(Y|Z)} \right] \quad \text{[apply Bayes Rule in numera}$$

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

$$\langle\!\langle \boldsymbol{m}_4 \rangle\!\rangle = \inf_{\mu(Y)} \inf_{\mu(X|Y)} \inf_{\mu(Z|Y,X)} \mathop{\mathbb{E}}_{y\sim\mu(Y)} \mathop{\mathbb{E}}_{x\sim\mu(X|y)} \mathop{\mathbb{E}}_{z\sim\mu(Z|y,x)} \left[\beta \log \frac{\mu(y)\,\mu(x\,|\,y)}{p(x)\,f(y\,|\,x)} + \zeta \log \frac{\mu(y)\,\mu(z\,|\,y)}{q(z)\,f(y\,|\,z)} \right],$$

where $\mu(Z|Y)$ is the 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 β) does not depend on z, we can take it out of the expectation, so

$$\langle\!\langle \boldsymbol{m}_4 \rangle\!\rangle = \inf_{\mu(Y)} \inf_{\mu(X|Y)} \inf_{\mu(Z|Y,X)} \mathop{\mathbb{E}}_{y\sim\mu(Y)} \mathop{\mathbb{E}}_{x\sim\mu(X|y)} \left[\beta \log \frac{\mu(y)\,\mu(x\,|\,y)}{p(x)\,f(y\,|\,x)} + \mathop{\zeta\ \mathbb{E}}_{z\sim\mu(Z|y,x)} \left[\log \frac{\mu(y)\,\mu(z\,|\,y)}{q(z)\,f(y\,|\,z)} \right] \right];$$

we can split up $\mathbb{E}_{\mu(X|y)}$ by linearity of expectation, to get

$$\langle\!\langle \boldsymbol{m}_4 \rangle\!\rangle = \inf_{\mu(Y)} \inf_{\mu(X|Y)} \inf_{\mu(Z|Y,X)} \mathop{\mathbb{E}}_{y\sim\mu(Y)} \left[\beta \mathop{\mathbb{E}}_{x\sim\mu(X|y)} \left[\log \frac{\mu(y)\,\mu(x\,|\,y)}{p(x)\,f(y\,|\,x)} \right] + \zeta \mathop{\mathbb{E}}_{\substack{x\sim\mu(X|y)\\ z\sim\mu(Z|y,x)}} \left[\log \frac{\mu(y)\,\mu(z\,|\,y)}{q(z)\,f(y\,|\,z)} \right] \right]$$

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 infemum over $\mu(Z|X, Y)$ with one over $\mu(Z|Y)$ yields:

$$\langle\!\langle \boldsymbol{m}_4 \rangle\!\rangle = \inf_{\mu(Y)} \inf_{\mu(X|Y)} \inf_{\mu(Z|Y)} \mathop{\mathbb{E}}_{y\sim\mu(Y)} \left[\beta \mathop{\mathbb{E}}_{x\sim\mu(X|y)} \left[\log \frac{\mu(y)\,\mu(x\,|\,y)}{p(x)\,f(y\,|\,x)} \right] + \zeta \mathop{\mathbb{E}}_{z\sim\mu(Z|y)} \left[\log \frac{\mu(y)\,\mu(z\,|\,y)}{q(z)\,f(y\,|\,z)} \right] \right]$$

Now, a cpd $\mu(X|Y)$ is just[7] a (possibly different) distribution $\nu_y(X)$ for every value of Y. Observe that, inside the expectation over $\mu(Y)$, the cpds $\mu(X|Y)$

[7] modulo measurability concerns that do not affect the infemum; see Section 5.A

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 to take the infemum over $\nu(X)$, separately for each y. Symmetrically, we may as well take the infemum over $\lambda(Z)$ separately for each y, rather than jointly finding the optimal $\mu(Z|Y)$ all at once. Operationallly, this means we can pull the infema 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:

$$\langle\!\langle \mathcal{m}_4 \rangle\!\rangle = \inf_{\mu(Y)} \mathop{\mathbb{E}}_{y\sim\mu(Y)} \left[\inf_{\nu(X)} \beta \mathop{\mathbb{E}}_{\nu(X)} \left[\log \frac{\mu(y)\,\nu(X)}{p(X)\,f(y\,|X)} \right] + \inf_{\lambda(Z)} \zeta \mathop{\mathbb{E}}_{\lambda(Z)} \left[\log \frac{\mu(y)\,\lambda(Z)}{q(Z)\,f(y\,|Z)} \right] \right]$$

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

$$\langle\!\langle \mathcal{m}_4 \rangle\!\rangle = \inf_{\mu(Y)} \mathop{\mathbb{E}}_{y\sim\mu(Y)} \left[\begin{array}{c} \beta \inf_{\nu(X)} \boldsymbol{D}\Big(\nu(X) \,\Big\|\, \frac{1}{C_1(y)} p(X) \frac{f(y|X)}{\mu(y)}\Big) - \beta \log C_1(y) \\ +\zeta \inf_{\lambda(Z)} \boldsymbol{D}\Big(\lambda(Z) \,\Big\|\, \frac{1}{C_2(y)} q(Z) \frac{f(y|Z)}{\mu(y)}\Big) - \zeta \log C_2(y) \end{array} \right],$$

where

$$C_1(y) = \sum_x p(x) \frac{f(y|x)}{\mu(y)} = \frac{1}{\mu(y)} \mathop{\mathbb{E}}_{p(X)} f(y|X) \qquad \text{and} \qquad C_2(y) = \sum_z q(z) \frac{f(y|z)}{\mu(y)} = \frac{1}{\mu(y)} \mathop{\mathbb{E}}_{q(Z)} f$$

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

$$\langle\!\langle \mathcal{m}_4 \rangle\!\rangle = \inf_{\mu(Y)} \mathop{\mathbb{E}}_{y\sim\mu(Y)} \left[\beta \log \frac{1}{C_1(y)} + \zeta \log \frac{1}{C_2(y)} \right]$$

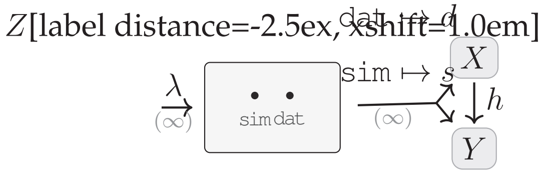
$$= \inf_{\mu(Y)} \mathop{\mathbb{E}}_{y\sim\mu(Y)} \left[\beta \log \frac{\mu(y)}{\mathbb{E}_{p(X)}\, f(y|X)} + \zeta \log \frac{\mu(y)}{\mathbb{E}_{q(Z)}\, f(y|Z)} \right]$$

$$= \inf_{\mu(Y)} \mathop{\mathbb{E}}_{\mu} \left[\beta \boldsymbol{D}(\mu \parallel f \circ p) + \zeta \boldsymbol{D}(\mu \parallel f \circ q) \right]$$

$$= \langle\!\langle \mathcal{m}_5 \rangle\!\rangle.$$

Details for Claims made in Section 5.8

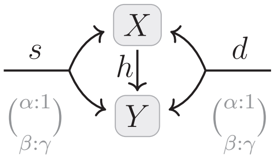
First, the fact that

$$\mathcal{L}_1 = \lambda_{\mathrm{d}} \mathcal{L}_{\mathtt{dat}} + \lambda_{\mathrm{s}} \mathcal{L}_{\mathtt{sim}} = \Bigg\langle\!\!\Bigg\langle$$  $$\Bigg\rangle\!\!\Bigg\rangle,$$

where $\lambda(Z = \mathtt{sim}) = \lambda_{\mathrm{s}}$ and $\lambda(Z = \mathtt{dat}) = \lambda_{\mathrm{d}}$ is immediate. The two cpds with infinite confidence ensure that the only joint distribution with a finite score is $\lambda_{\mathrm{s}} s + \lambda_{\mathrm{d}} d$, and the inconsistency with h is its surprisal, so the inconsistency of this PDG is

$$\mathop{\mathbb{E}}_{\lambda_{\mathrm{s}} s + \lambda_{\mathrm{d}} d} \left[\log \frac{1}{h(Y|X)} \right] = -\lambda_{\mathrm{s}} \mathop{\mathbb{E}}_{s}[\log h(Y|X)] - \lambda_{\mathrm{d}}\, \mathbb{E}\, d[\log h(Y|X)] = \lambda_{\mathrm{d}} \mathcal{L}_{\mathtt{dat}} + \lambda_{\mathrm{s}} \mathcal{L}_{\mathtt{sim}} = \mathcal{L}_1, \quad \text{as pro}$$

The second correspondence is the least straightforward. Let $C = \int sd$ be the normalization constant required to normalize the joint density sd. We claim that, for large fixed γ, we have

$$\mathcal{L}_2 \approx C \Bigg\langle\!\!\Bigg\langle$$  $$\Bigg\rangle\!\!\Bigg\rangle_{\gamma} + \mathit{const},$$

where const does not depend on h. To see this, let $\mathcal{m}_2$ be the PDG above, and compute

$$\langle\!\langle \mathcal{m}_2 \rangle\!\rangle_{\gamma} = \inf_{\mu(X,Y)} \mathop{\mathbb{E}}_{\mu} \left[\overbrace{\gamma \log \frac{\mu(XY)}{s(XY)} \frac{\mu(XY)}{d(XY)} + \log \frac{\mu(Y|X)}{h(Y|X)}}^{\mathit{Inc}(\mu)} + \overbrace{\gamma \log \frac{1}{s(XY)} \frac{1}{d(XY)} - \gamma \log \frac{1}{\mu(XY)}}^{\mathit{IDef}\,(\mu)} \right]$$

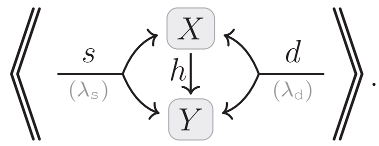
$$
\begin{aligned}
&= \inf_{\mu(X,Y)} \mathop{\mathbb{E}}_{\mu} \left[\gamma \log \frac{\mu(XY)}{s(XY)} \frac{\mu(XY)}{d(XY)} \frac{1}{\mu(XY)} \frac{1}{\mu(XY)} \frac{\mu(XY)}{1} + \log \frac{\mu(Y|X)}{h(Y|X)} \right] \\
&= \inf_{\mu(X,Y)} \mathop{\mathbb{E}}_{\mu} \left[\gamma \log \frac{\mu(XY)}{s(XY)d(XY)} + \log \frac{\mu(Y|X)}{h(Y|X)} \right] \\
&= \inf_{\mu(X,Y)} \mathop{\mathbb{E}}_{\mu} \left[\gamma \log \frac{\mu(XY)C}{s(XY)d(XY)} - \gamma \log C + \log \frac{\mu(Y|X)}{h(Y|X)} \right] \\
&= \inf_{\mu(X,Y)} \gamma \boldsymbol{D}\left(\mu \;\middle\|\; \frac{1}{C} sd \right) + \mathop{\mathbb{E}}_{\mu} \left[\log \frac{\mu(Y|X)}{h(Y|X)} \right] - \gamma \log C
\end{aligned}
$$

$\boldsymbol{D}$ is (γm)-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 γ becomes large, the first term dominates, and the optimizing μ quickly approaches the normalized density $\nu := \frac{1}{C} sd$. Plugging in ν, we find that the value of the infemum approaches

$$
\begin{aligned}
\langle\!\langle \mathcal{m}_2 \rangle\!\rangle &\approx \mathop{\mathbb{E}}_{\nu} \left[\log \frac{1}{h(Y|X)} \right] - H_\nu(Y|X) - \gamma \log C \\
&= \int_{XY} \frac{1}{C} \log \frac{1}{h(Y|X)} s(X,Y) d(X,Y) \quad - H_\nu(Y|X) - \gamma \log C \\
&= \frac{1}{C} \mathop{\mathbb{E}}_{s} \left[d(X,Y) \log \frac{1}{h(Y|X)} \right] - H_\nu(Y|X) - \gamma \log C \\
&= \frac{1}{C} \mathcal{L}_2 - H_\nu(Y|X) - \gamma \log C,
\end{aligned}
$$

and therefore
$$
\begin{aligned}
\mathcal{L}_2 &= C \langle\!\langle \mathcal{m}_2 \rangle\!\rangle + C \,\mathrm{H}_\nu(Y|X) - \gamma \, C \log C \\
&= C \langle\!\langle \mathcal{m}_2 \rangle\!\rangle + \mathit{const}.
\end{aligned}
$$

Finally, we turn to

$$\mathcal{L}_3 :=$$


.

To see the why the optimal distribution $\mu^*(XY)$ is the λ-weighted geometric mean of s and d, let us first consider the same PDG, except without h. From

Lemma 5.5.2, we have this loss without h in closed form, and from the proof of Lemma 5.5.2, we see that the optimizing distribution in this case is the λ-weighted geometric distribution $\mu^* \propto s(XY)^{\lambda_{\mathrm{s}}} d(XY)^{\lambda_{\mathrm{d}}}$. Now (Lemma 5.2.1), including h cannot make the PDG any less inconsistent. In particular, by choosing

$$h^*(Y|X) := \mu^*(Y|X) \propto (Y|X)^{\lambda_{\mathrm{s}}} d(Y|X)^{\lambda_{\mathrm{d}}},$$

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 5.9

Distortion Due to Inconsistency. In the footnote on Page 156, we claimed that if the model confidence β_p were 1 rather than ∞, we would have obtained an incconsistency of $-\log \mathbb{E}_{x\sim p} \exp(-c(x))$, and that the optimal distribution would not have been $p(X)$.

$$\begin{aligned}
\left\langle\!\!\left\langle \xrightarrow{p} \boxed{X} \xrightarrow{\hat{c}} \boxed{\mathtt{T}} \overset{\mathtt{t}}{\twoheadleftarrow} \right\rangle\!\!\right\rangle &= \inf_{\mu(X)} \mathop{\mathbb{E}}_{x\sim\mu} \left[\log \frac{\mu(x)}{p(x)} + \log \frac{\mu(\mathtt{t}\,|\,x)}{\hat{c}(\mathtt{t}\,|\,x)} \right] \\
&= \inf_{\mu(X)} \mathop{\mathbb{E}}_{x\sim\mu} \left[\log \frac{\mu(x)}{p(x)} + \log \frac{1}{\hat{c}(\mathtt{t}\,|\,x)} \right] \\
&= \inf_{\mu(X)} \mathop{\mathbb{E}}_{x\sim\mu} \left[\log \frac{\mu(x)}{p(x)\exp(-c(x))} \cdot \frac{Z}{Z} \right]
\end{aligned}$$

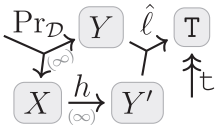
where $Z = \sum_x p(x)\exp(-c(x)) = \mathbb{E}_p \exp(-c(X))$ is the constant required to normalize the distribution

$$
\begin{aligned}
&= \inf_{\mu(X)} \boldsymbol{D}\left(\mu \,\middle\|\, \frac{1}{Z}\, p(X)\exp(-c(X))\right) - \log Z \\
&= -\log Z \\
&= -\log \mathop{\mathbb{E}}_{x\sim p} \exp(-c(x))
\end{aligned}
$$

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. We also claimed without careful proof in Section 5.9 that, if $\alpha_h = \alpha_{\Pr_{\mathcal{D}}} = 1$, then

$$\lim_{\gamma\to\infty} \left\langle\!\!\left\langle$$



$$\right\rangle\!\!\right\rangle_{\gamma} = \mathop{\mathbb{E}}_{\substack{(x,y)\sim\Pr_{\mathcal{D}} \\ y'\sim p(Y'|\,x)}} \big[\ell(y,y')\big]$$

Why is this? For such a setting of α, which intuitively articulates a causal picture where X, Y is generated from $\Pr_{\mathcal{D}}$, and Y' generated by $h(Y'|X)$, the information deficiency $\mathit{IDef}_{\mathcal{S}}(\mu(X,Y,Y'))$ of a distribution μ is

$$
\begin{aligned}
\mathit{IDef}_{\mathcal{S}}(\mu(X,Y,Y')) &= -\,\mathrm{H}_\mu(X,Y,Y') + \mathrm{H}(X,Y) + \mathrm{H}(Y'|X) \\
&= \mathrm{H}_\mu(Y'|X) - \mathrm{H}_\mu(Y'|X,Y) \\
&= \mathrm{I}_\mu(Y;Y'|X).
\end{aligned}
$$

Both equalities of the derivation above standard information theoretic identities [See, for instance, 24], and the final quantity $\mathrm{I}_\mu(Y;Y'|X)$ 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 $\gamma \to \infty$ any distribution that for which Y' and Y are not independent given X will incur infinite cost. Since the confidences in h and $\Pr_{\mathcal{D}}$ 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 5.9.1 applies: all other cpds must be matched, and the inconsistency is the expected incompatibility of $\hat{l}$, which equals

$$\mathop{\mathbb{E}}_{\substack{(x,y)\sim\Pr_{\mathcal{D}}\\ y'\sim p(Y'|x)}} \log \frac{1}{\hat{\ell}(\mathtt{t}\,|y,y')} = \mathop{\mathbb{E}}_{\substack{(x,y)\sim\Pr_{\mathcal{D}}\\ y'\sim p(Y'|x)}} \log \frac{1}{\exp(-\ell(y,y'))} = \mathop{\mathbb{E}}_{\substack{(x,y)\sim\Pr_{\mathcal{D}}\\ y'\sim p(Y'|x)}} \Big[\log \exp(\ell(y,y'))\Big] = \mathop{\mathbb{E}}_{\substack{(x,y)\sim\Pr_{\mathcal{D}}\\ y'\sim p(Y'|x)}} \Big[\ell(y,y')\Big] = \mathcal{L}$$

5.D More Notes

5.D.1 Maximum A Posteriori and Priors

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 Θ. The maximum-likelihood estimate of Θ 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 Θ. 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 θ, 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,

$$\text{log-posterior } \log \Pr(\Theta|X) = \text{log-likelihood } \log \Pr(X|\Theta) + \text{log-prior } \log \Pr(\Theta) - \text{log-evidence } \log$$

The final term does not depend on θ, so it is not relevant for finding the optimal θ by this metric. Swapping the signs so that we are taking a minimum rather than a maximum, the MAP estimate is then given by

$$\theta^{\mathrm{MAP}}(X) := \arg\min_{\theta \in \Theta} \left\{ \log \frac{1}{\Pr(X|\theta)} + \log \frac{1}{\Pr(\theta)} \right\}.$$

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 Θ, to our loss function. It is in this sense that priors classically correspond to regularizers.

CHAPTER 6

THE LOCAL INCONSISTENCY RESOLUTION (LIR) ALGORITHM

Part III

Algorithms and Complexity

CHAPTER 7

INFERENCE FOR PDGS, VIA EXPONENTIAL CONIC PROGRAMMING

CHAPTER 8

EQUIVALENCE BETWEEN

Part IV

Reasoning about PDGs

CHAPTER 9

Part V

Foundations

CHAPTER 10

CONFIDENCE

CHAPTER 11

THE CATEGORY THEORY OF PDGS

A Categorical Definition of a PDG

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

$$\mathbb{P} : \mathcal{A}^* \to \mathbb{S}\mathbf{toch}$$

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 $F : J \to \mathcal{C}$ is often called a *diagram* of $\mathcal{C}$ (of shape J). Therefore, an unweighted PDG is a diagram of the $\mathbb{S}\mathbf{toch}$, of shape generated by its underlying hyprgraph. In addition to probabilities, a PDG also contains confidences $\boldsymbol{\beta} = \{\beta_a\}_{a \in \mathcal{A}}$ about the reliability of those probabilities.

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 $-\hat{\nabla}_\mu \operatorname{H}_\mu(Y|X)$, the gradient of the structural loss corresponding to a hyperarc $X \xrightarrow{a} Y$, is the same as $+\hat{\nabla} \boldsymbol{D}(\mu(X,Y) \parallel \mu(X)\lambda_Y(Y))$, the gradient of the observational loss corresponding to a uniform distribution. Furthermore, the weight β_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[1] as a single conditional measure

[1] However, there is no way to combine β with p that results in a quantity q (independent of μ)

$(1-e^{-\beta})p(Y|X)$ 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 sematically, 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 β_1 with $q(Z|Y)$ with confidence β_2, where both $\beta_1, \beta_2 \in [0, \infty]$. Then the composite

$$\begin{aligned} r(Z|X) &= \int_Y (1-e^{-\beta_2})p(Z|Y)(1-e^{-\beta_1})\mathrm{d}p(Y|X) \\ &= (1-e^{-\beta_1}-e^{-\beta_2}+e^{-\beta_1-\beta_2})\;(q\circ p)(Z|X) \\ &\approx (1-\exp(-\min\{\beta_1,\beta_2\}))\;(q\circ p)(Z|X). \end{aligned}$$

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.[2] As a result, all data in a PDG ($\mathcal{A}$, $\mathcal{N}$, $\boldsymbol{\alpha}$, $\mathcal{V}$, $\mathbb{P}$, $\boldsymbol{\beta}$) may be specified together with a single functor

$$\mathscr{m} : \mathcal{A}^* \to \mathbb{M}\mathbf{eas}_{\underline{\Delta}}. \tag{11.1}$$

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:

that can be plugged directly into the ordinary expression for KL divergence:

$$\beta \log \frac{\mu}{p} = \log \frac{\mu}{q} \qquad \Longrightarrow \qquad q = \mu^{1-\beta} p^{\beta}.$$

[Zhu & Rower] suggest that this is not the appropriate notion of relative entropy, for unsigned measures; instead, one should use $\boldsymbol{D}(\mu \parallel \nu) := \int \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\mathrm{d}\mu + \int \mathrm{d}\mu - \int \mathrm{d}\nu$, but even with this

[2] 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 α should enter the picture, ideally.

1. A PDG $\mathcal{M} : \mathcal{A}^* \to \mathbb{M}\mathrm{eas}_{\underline{\Delta}}$ 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 $\boldsymbol{\alpha}$ combined with $\boldsymbol{\beta}$ 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?

11.0.1 Limits

Definition 11.0.1. Let $\mathcal{M}$ be a PDG, with variables $\mathcal{X}$. The *local marginal polytope*

$$\mathbb{L}(\mathcal{M}) := \Big\{ \{\mu_X \in \Delta\mathcal{V}X\}_{X \in \mathcal{X}} \;\Big|\; \forall S \xrightarrow{a} T \in \mathcal{A}.\ \mu_T = \mathbb{P}_a \circ \mu_S \Big\}. \tag{11.2}$$

consists of all marginals over the variables $\mathcal{N}$ that are locally consistent with all arcs. □

$$\begin{tikzcd}
 & & V \arrow[dl, "\forall"'] \arrow[d] \arrow[dr] \arrow[r, dashed, "\exists!"] & L \arrow[dll] \arrow[dl] \arrow[d] & \cong X_1X_3 \\
\cdots & X_1 \arrow[r, two heads] & X_2 & X_3 \arrow[l] & \cdots
\end{tikzcd}$$

Example 17. 1. Suppose $\mathcal{A} = \{\to X\}$, and its unique arc is

□

Our next results are sensitive to the particulars of the PDG encoding as a functor. Let $\mathcal{M}^{+\mathcal{X}} := \mathcal{M} \cup \{\mathbf{X} \to \mathbf{Y}\}_{\mathbf{Y} \subset \mathbf{Y} \subset \mathcal{X}}$ be the PDG $\mathcal{M}$ 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.[3] Let $\mathcal{M}^{hold}$ be the PDG where $p(\mathbf{Y}|\mathbf{X})$ is really attached to a hyperarc $\mathbf{X} \to \mathbf{X} \cup \mathbf{Y}$, implicitly the identity along $\mathbf{X} \setminus \mathbf{Y}$. Write $\mathcal{M}^{+\mathcal{X},hold}$ for the PDG with both alterations.

Theorem 11.0.1. *Suppose $\mathcal{M}$ is a PDG in which every arc has full confidence. Then,*

1. $\mathrm{Cones}(\mathcal{M}, 1) \cong \mathbb{L}(\mathcal{M})$;
2. $\mathrm{Cones}(\mathcal{M}^{+\mathcal{X},hold}, 1) \cong \{\!\{\mathcal{M}\}\!\}$.
3. $\lim \mathcal{M}^{+\mathcal{X},hold} = \mathrm{Ext}\{\!\{\mathcal{M}\}\!\}$, *where* $\mathrm{Ext}(S)$ *is the set of extreme points of a set* S *(e.g., the vertices of* S*, when* S *is a polytope).*
4. $\lim \mathcal{M} = \mathrm{Ext}\,\mathbb{L}(\mathcal{M})$.

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.

[3] Note that these cohrerence maps do not include hyperarcs on these variables to put them back together, e.g., $\{\{X\},\{Y\}\} \to \{X, Y\}$. Including such a map is appropriate only if one believes X and Y are independent.

Now for part 2. A cone over $\mathcal{M}^{+\mathcal{X},hold}$ 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 $\mu_T(S,T) = \int_S \mathbb{P}_a(S,T|s')\mu_S(s')\,\mathrm{d}s'$, properly overwriting variables in S according to p). In particular, $\mathcal{M}^{+\mathcal{X},hold}$ 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 $\mathcal{M}^{+\mathcal{X}}$ and $\mathcal{M}^{+\mathcal{X},hold}$. Such a distribution (and its induced 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 μ 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 $p(X|X)$,

□

11.0.2 Colimits

The colimits are arguably even more interesting: they summarize the most general thing that is known by all variables.

There's always a co-cone with vertex 1, and there's a unique way to f

11.0.3 Natural Transformations

Suppose $\mathcal{M}_1, \mathcal{M}_2 : \mathcal{A}^* \to \mathbb{M}\mathrm{eas}_{\underline{\Delta}}$ 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 $\{\eta_X : \mathscr{m}_1(X) \to \mathscr{m}_2(X)\}_{X \in \mathcal{N}}$,[4] satisfying the property that, for all $a : S \to T \in \mathcal{A}$,[5] the diagram

$$\begin{tikzcd}
\mathscr{m}_1(X) \arrow[r, "\mathscr{m}_1(a)"] \arrow[d, "\eta_X"] & \mathscr{m}_1(Y) \arrow[d, "\eta_Y"] \\
\mathscr{m}_2(X) \arrow[r, "\mathscr{m}_2(a)"] & \mathscr{m}_2(Y)
\end{tikzcd}$$

(or, in the original notation,

$$\begin{tikzcd}
\mathcal{V}_1 X \arrow[r, "\mathbb{P}_a"] \arrow[d, "\eta_X"'] & \mathcal{V}_1 Y \arrow[d, "\eta_Y"] \\
\mathcal{V}_2 X \arrow[r, "\mathbb{P}^2_a"] & \mathcal{V}_2 Y
\end{tikzcd}$$

)

commutes. This is a diagram in the category $\mathbb{M}\mathbf{eas}_{\underline{\Delta}}$. I immediately have questions:

1. What does the space of natural transformations from $\mathscr{m}_1$ to itself look like?
2. What if $\mathscr{m}_1$ and $\mathscr{m}_2$ differ only in $\boldsymbol{\beta}$? What is the effect of different encodings?
3. In what situations is there a natural transformation from one PDG into the other?
4. What about for certain special PDGs? What are some special PDGs with structure $\mathcal{A}$?

Fix a PDG $\mathscr{m} : \mathcal{A}^* \to \mathbb{M}\mathbf{eas}_\Delta$. For a a measurable space W, let $\boldsymbol{\Delta}_W$ be the functor assigning each $N \in \operatorname{ob} \mathcal{A}^*$ to the measurable space W, and each arc a to the identity map $\mathrm{id}_W : W \to W$. Can we characterize the natural transformations from $\mathscr{I}$ to $\mathscr{m}$?

[4] 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.

[5] 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 $a \in \mathcal{A}$.

11.0.4 Additional Structure Preserved by PDGs

Monoidal Structure

If we allow unions of variables at the qualitative level, this means we are working in a different category $\mathcal{A}^{**}$ whose objects $\mathrm{ob}\,\mathcal{A}^{**}$ are all subsets of $\mathcal{N}$, and equipped with down-projection morphisms, and joining hyperarcs. Every PDG $\mathcal{M} : \mathcal{A}^* \to \mathbb{M}\mathbf{eas}_{\underline{\Delta}}$ can be naturally lifted to a PDG $\mathcal{M}^{+\mathcal{X}} : \mathcal{A}^{**} \to \mathbb{M}\mathbf{eas}_{\underline{\Delta}}$ in the obvious way: taking the additional joint variables to the appropriate product of measurable spaces, and treating their downprojections ($\mathbf{X} \twoheadrightarrow \mathbf{Y}$, for $\mathbf{X} \supseteq \mathbf{Y}$) appropriately. We already saw one consequence of this change: the limits of such a PDG must be internally coherent in a certain way; the represent not the local marginal polytope, but the global marginal polytope.

Variable Union as Monoidal Structure, for $\mathcal{A}^{*,hold}$ Usually people call the a monoidal operation "tensor", but we will now define a monoidal operation that does not line up with the usual tensor product. On objects, which are sets of variables, $\odot$ behaves like union, and on morphisms, it simply multiplies densities.

$$\begin{aligned} A \odot C &= A \cup C; \\ \odot(p(B|A), q(D|C)) &:= p(B|A) \cdot q(D|C) \\ &= (a_0, y, c_0) \mapsto \Big((b_0, z, d_0) \mapsto p(b_0, z|a_0, y) q(d_0, z|c_0, y) \Big). \end{aligned}$$

where z gives the values of the common variables $A \cap C$, so that $a = (a_0, z)$ and $c = (c_0, z)$.

If A and C are disjoint, as are B and D, then $\odot$ coincides with the tensor product $\otimes$. When they share variables, the resulting operation does something different: instead of having two distinct copies of eacn input and output, $p \odot q$ takes in only one copy of each shared sources, and produces a density over only one copy of their shared targets. For example $p(X) \odot p(X)$ is the subprobability $p(X)^2$. It is worth noting that a morphism $p(Y|X)$ is idempotent, in the sense that $p(Y|X) \odot p(Y|X)$ iff it is a deterministic function.

Although $\odot$ is well-defined, it is not functorial in all cases,[6] and hence inadmissible as the basis of a monoidal structure for the full category of stochastic morphisms . However, if we restrict to the subcategory of purely generative morphisms—that is, arcs a satisfying $T_a \supseteq S_a$—then $\odot$ becomes functorial.

CAREFUL! This messes up the types of composition! If $f : X \to Y$ is converted to $f' : X \to XY$ and $g : Y \to Z$ is converted to $g' : Y \to YZ$, then $g' \circ f'$ is not defined (because $XY \neq Y$) and so composition cannot proceed, without first including a forget/downprojection map!

There are also many other properties that we must verify in order to get a symmetric monoidal category; we now verify them.

- **Functoriality.** We need to show that

$$\left(\begin{tikzcd} X_1 \arrow[d, "f_1"'] \arrow[dd, bend left, "p"] \\ X_2 \arrow[d, "f_2"'] \\ X_3 \end{tikzcd} \right) \odot \left(\begin{tikzcd} Y_1 \arrow[d, "g_1"'] \arrow[dd, bend left, "q"] \\ Y_2 \arrow[d, "g_2"'] \\ Y_3 \end{tikzcd} \right) = \begin{tikzcd} X_1 \cup Y_1 \arrow[d, "f_1 \odot g_1"] \\ X_2 \cup Y_2 \arrow[d, "f_2 \odot g_2"] \\ X_3 \cup Y_3 \end{tikzcd} .$$

[6] For example, $p(Y|X) \odot p(Y|X) = p^2(Y|X)$ which is a strict subprobability measure, while $(q \circ p)(Z|X)$ and $(r \circ p)(W|X)$ are both cpds, and $(q \circ p) \odot (r \circ p)$ will also be a cpd on $W \cup Z$, supposing that W and Z are disjoint. So it cannot be the case that $(q \circ p) \odot (q \circ)$

As mentioned above, this is not true in general. But we have assumed that $X_1 \subseteq X_2 \subseteq X_3$ and $Y_1 \subseteq Y_2 \subseteq X_3$. To simplify notation, let's redefine $X_3 \leftarrow X_3 \setminus X_2$ and $X_2 \leftarrow X_2 \setminus X_1$. In this new notation, our goal becomes proving the commutativity of the following diagram:

$$\left(\begin{tikzcd} X_1 \arrow[d, "f_1"'] \\ X_2 \cup X_1 \arrow[d, "f_2"'] \\ X_3 \cup X_2 \cup X_1 \end{tikzcd}\right) \odot \left(\begin{tikzcd} Y_1 \arrow[d, "g_1"'] \\ Y_2 \cup Y_1 \arrow[d, "g_2"'] \\ Y_3 \cup Y_2 \cup Y_1 \end{tikzcd}\right) = \begin{tikzcd} X_1 \cup Y_1 \arrow[d, "f_1 \odot g_1" description] \\ X_2 \cup X_1 \cup Y_2 \cup Y_1 \arrow[d, "f_2 \odot g_2" description] \\ X_3 \cup X_2 \cup X_1 \cup Y_3 \cup Y_2 \cup Y_1 \end{tikzcd} .$$

We now compute

$$\begin{aligned}
(f_2 \circ f_1)(x_1'')(x_1, x_2, x_3) &= \iint_{X_2, X_1} f_1(x_2', x_1' | x_1'') f_2(x_3, x_2, x_1 | x_2', x_1') \mathrm{d}x_1' \mathrm{d}x_2' \\
&= \iint_{X_2, X_1} f_1(x_2' | x_1'') f_2(x_3 | x_2) \delta(x_1) \mathrm{d}x_1' \mathrm{d}x_2'
\end{aligned}$$

•

(Pre)additivity

11.0.5 The Category of PDGs

11.1 Dependency Graphs for Other Monads

The Big Questions:

1. How does the monadic view of composition (bind / multiply), which

describes composition in the underlying Kleisli category, interact with the "scoring function semantics"?

2. Is there an important shared feature among the monads T for which analogues of PDGs work out? To set up the scoring semantics, it seems we need

   (a) a way to quantify "degree of funtional dependence" along an arc $S \to T$, in the limit object $\lim \mathcal{M}$, and

   (b) a way to quantify degree of between $T(X, Y, Z)$ and $X \to T(Y)$.

   If there is an analogue of marginalization, then there is a map $T(X, Y, Z) \to T(X, Y)$, and there is a

3. Since most monads do not construct continuous geometry as nicely as the probability monad, relations are not continuous, there is no obvious analogue of parallel, symmetric, "mixture composition". Even when we give a loss function semantics (which we do below) this does not correspond to an obvious computational picture in the same way.

11.1.1 Relational Dependency Graphs

We will use $\mathcal{P}$ do denote the relational monad.

Scoring Function Semantics. Suppose that instead of mapping $X \xrightarrow{a} Y$ to a conditional probability $\mathbb{P}_a(Y|X)$, we instead map it to a relation $R_a(X, Y)$.

The analogue of a scoring function might operate on a universal reation

$U \subseteq \mathcal{V}\mathcal{X}$, and look something like:

$$Inc(U) = \sum_{X \overset{a}{\to} Y \in \mathcal{A}} \beta_a \Big\| U(X,Y) - R_a(X,Y) \Big\|_1$$
$$\text{where} \quad \Big\| U(X,Y) - R_a(X,Y) \Big\|_1 = \#\Big\{ (x,y) \in \mathcal{V}(X,Y) \;\Big|\; R(x,y) \Leftrightarrow \exists \mathbf{z}.\, U(x,y,\mathbf{z}) \Big\},$$

with each $\beta_a \in \mathbb{N}$.

The analogue of a qualitative arrow $X \overset{a}{\to} Y$, indicating that one attribute determines another, also requires a scoring function. The fact that the argument to $\mathit{IDef}\,\mathcal{A}, \boldsymbol{\alpha}$ is a joint distribution may not be critical, so long as we can find a suitable replacement notion of uncertainty along an arc. One possible analogue of conditional entropy might then be

$$\begin{aligned} \mathrm{H}_U(Y|X) &= \log(\text{maximum \# of possible values of } Y \text{ given } X) \\ &= \max_{x \in \mathcal{V}X} \log \#\{ y \in \mathcal{V}Y \mid \exists \omega \in U.\;\; S_\omega = s,\; T_\omega = t \}. \\ &= \max_{x \in \mathcal{V}X} \log \#\{ y \in \mathcal{V}Y \mid \exists \mathbf{z}.\, U(x,y,\mathbf{z}) \}. \end{aligned} \tag{11.3}$$

This shares an important property with conditional entropy: it is zero iff the value of Y is determined by X in the relation. It is undefined when U is empty, and otherwise non-negative. The other important property of conditional entropy is monotonicity with respect to weakening.[7]

Altogether, the analogue of IDef is

$$\mathit{IDef}\,\mathcal{A}^*(U) := \sum_{a \in \mathcal{A}} \max_{x \in \mathcal{V}X} \log \#\{ y \in \mathcal{V}Y \mid \exists \mathbf{z}.\, U(x,y,\mathbf{z}) \} - \log \#U$$

Here, as in the probabilistic case, there is an interpretation in terms of storage costs. Suppose U is fixed and known. The first term is the number of bits needed

[7] Clearly an extra target to a hyperarc (i.e., extending Y to $Y' = (Y,Z)$) can only make (11.3) larger. Perhaps less obviously: it also has the property that adding an extra source (i.e., extending X to $X' = (X,Z)$) can only reduce the value of (11.3). This is because if the maximum is over joint pairs (x,z), then only the maximum number of $\{y : (x,y,z) \in U\}$ contribute, while all such z are amalgamated and make the number larger in the case of an existential quantifier.

to specify separately each target given the source (knowing that the result is in U), while the second is the number of bits needed to specify an element of U directly. The value is undefined iff $|U| = 0$ because, intuitively, it is impossible to specify a joint setting $\omega \in U$ if U is empty.

Now, for some examples.

Example 18. Suppose $\mathcal{A} = \{\to X, \;\; Y \leftarrow\}$. Then $\mathit{IDef}\,\mathcal{A}^*(U(X,Y)) \geq 0$ with equality iff $U(X,Y) = U_X(X) \bowtie U_Y(Y)$, for some unary relations $U_X(X)$ and $U_Y(Y)$. □

More generally, it can be shown that, for target-partitinal hypergraphs $\mathcal{A}$ without sources, the quantity $\mathit{IDef}\,\mathcal{A}^*(U)$ measures how far a joint relation U is from decomposing independently along the specified arcs.

Proposition 11.1.1. *Let $\{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$ be a partition of $\mathcal{X}$, and $\mathcal{A} = \{\to \mathcal{X}_i\}_{i=1}^n$ be the hypergraph consisting of a single hyperarc pointing to each partition, each without any sources. In this case, $\mathit{IDef}\,\mathcal{A}^*(U) \geq 0$, with equality iff $U = U_1 \bowtie \cdots \bowtie U_n$ is the natural join of n subrelations $U_i \subseteq \mathcal{V}(\mathcal{X}_i)$. (In addition, in that case it must be that each $U_i = \prod_{\mathcal{X}_i}(U)$ is the projection of U onto the variables $\mathcal{X}_i$.))*

Let's turn to some more complicated examples. First, let's start with overlapping targets, and keep everything unconditional. It is easy to see that

$$\mathit{IDef}\,[\to X \leftarrow]^*(U(X)) = \log |U|,$$

which is defined when $U \neq \emptyset$, non-negative on this domain, and equal to zero if and only if $U(X)$ is a singleton. Thus, it captures determinism (almost) exactly the same way as in the probabilistic case. Now, for a more complicated example.

Example 19. In the probabilistic setting, all $\mu(X, Y)$ are compatible with the hypergraph $[\rightarrow X \rightarrow Y]$, intuitively because optimal codes for specifying $(x, y) \sim \mu$ directly take the same amount of information as optimal codes to first specify x, and then use a code dependent on x to specify y.

In this relational setting, the two are not always the same. If $U = \{(x, y) : x \in S, y = f(x)\}$, then $\mathrm{H}_U(X) = \log |S|$ and $\mathrm{H}_U(Y|X) = 0$, since there is always precisely 1 y for which $(x, y) \in U$. Thus $\mathit{IDef}[\rightarrow X \rightarrow Y]^*(U) = 0$. Under the cost-of-storage interpretation: this means when U can be generated by selecting a subset of S and applying a (known) function, then specifying a joint sample (x, y) requires the same number of bits as specifying x.

A generalization of this holds for what might be called "k-multi-functional" relationships, where $k \geq 1$ is some natural number. Suppose $f : X \rightarrow Y^k$ produces k distinct values of Y for each $x \in X$. If $U = \{(x, y) : x \in S, y \in f(x)\}$, then $\mathrm{H}_U(Y|X) = \log k$, but also $|U| = k|S|$. So again $\mathit{IDef}[\rightarrow X \rightarrow Y]^*(U) = 0$. This is becaus, again, it takes the same number of bits to first specify a value of X, and then use that value to specicfy a value of Y, as it does to specify (x, y) together. It can be shown that, as a function of non-empty relations U, the value of $\mathit{IDef}[\rightarrow X \rightarrow Y]^*(U)$ is non-negative and zero precisely if U is of the form described above.

This is beause, when f may produce a variable number of points depending on x, $\mathit{IDef}[\rightarrow X \rightarrow Y]^*(U)$ will be positive overall. So, when x is such that $|f(x)|$ is maximal, then specifying fist x and then the appropriate y, is less efficient than specifying (x, y) together. □

More generally, for directed graphs, we have an analogue of a conditional independence.

Definition 11.1.1. Suppose $A, B, C \subseteq \mathcal{X}$. In a relation $R(\mathcal{X})$, A and B are said to be conditionally independent given C (symbolically, $R \models A \perp\!\!\!\perp B \mid C$) iff

$$\forall (a, b, c) \in \mathcal{V}(A, B, C). \qquad R(a, b, c) \iff \Big(\exists a' \in \mathcal{V}(A).\ R(a', b, c) \land \exists b' \in \mathcal{V}(B).\ R(a, b', c) \Big).$$

We write $A \perp\!\!\!\perp B$ to abbreviate $A \perp\!\!\!\perp B \mid \emptyset$, i.e., the special case where there are no given variables. □

Proposition 11.1.2. $R(A) \models A \perp\!\!\!\perp A$

Proposition 11.1.3. *Suppose A, B, C are sets of attributes. $R \models A \perp\!\!\!\perp B \mid C$ iff $R \models (A \setminus C) \perp\!\!\!\perp (B \setminus C) \mid C$.*

Proof. First, we claim that, for all $a, b, c \in \mathcal{V}(A, B, C)$, we have that $R(a, b, c) \iff R(a[A \setminus C], b[B \setminus C], c)$. If $\{a, b, c\}$ agree on shared values, then $R(a[A \setminus C], b[B \setminus C], c)$ must equal $R(a, b, c)$, by definition. On the other hand, if $\{a, b, c\}$ do not agree on shared values, then $R(a, b, c) = 0$, and there are three possibilities for conflict. If this is because a and c conflict, then it is possible that $R(a[A \setminus C], b, c) = 1$. contradicing our claim!

Let $A' := A \setminus C$ and $B' := B \setminus C$.

Suppose $R \models A \perp\!\!\!\perp B \mid C$, meaning that for all a, b, c,

$$\begin{aligned} R(a, b, c) &\iff \exists a'' \in \mathcal{V}(A).\ R(a'', b, c) \land \exists b'' \in \mathcal{V}(B).\ R(a, b'', c) \\ &\iff \exists a'' \in \mathcal{V}(A).\ R(a''[A \setminus C], b[B \setminus C], c) \land a''[A \cap C] = c[A \cap C] \\ &\qquad \land \exists b'' \in \mathcal{V}(B).\ R(a[A \setminus C], b''[B \setminus C], c) \land b''[B \cap C] = c[B \cap C] \\ &\iff \exists a' \in \mathcal{V}(A').\ R(a', b[B'], c) \land \exists b' \in \mathcal{V}(B').\ R(a[A'], b', c). \end{aligned}$$

incomplete; possibly untrue □

Proposition 11.1.4. $R(\mathcal{X}_1) \bowtie S(\mathcal{X}_2) \models \mathcal{X}_1 \perp\!\!\!\perp \mathcal{X}_2 \mid \mathcal{X}_1 \cap \mathcal{X}_2$.

Proof. □

Proposition 11.1.5. *If G is a directed acyclic graph, then IDef $\mathcal{A}_G{}^*(U) \geq 0$, with*

Example 20. Now, consider the 2-cycle □

The relationship between relations and probabilities. There is a map $\mathrm{Supp}_X : \Delta X \to 2^X$ that takes a probability measure to its support set. In fact, it is a natural transformation

$$\begin{tikzcd} \mathbf{FinSet} \arrow[r, bend left=50, "\Delta", ""{name=U, below}] \arrow[r, bend right=50, "2^{(-)}"', ""{name=D}] & \mathbf{Set} \arrow[Rightarrow, from=U, to=D, "\mathrm{Supp}"] \end{tikzcd} ,$$

since the diagram

$$\begin{tikzcd} \Delta X \arrow[r, "\delta f"] \arrow[d, "\mathrm{Supp}"'] & \Delta Y \arrow[d, "\mathrm{Supp}"] \\ 2^X \arrow[r, "\bar{f}"] & 2^Y \end{tikzcd}$$

commutes for all $f : X \to Y$,[8] where $\bar{f}(S) = \{f(x) : x \in S\}$, often simply written as just 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 $\mathrm{H}_\mu(Y|X)$ vs $\mathrm{H}_{\mathrm{Supp}\,\mu}(Y|X)$?))

[8] *Proof:* $y \in \mathrm{Supp}(\delta f(\mu)) \iff y \in f^{-1}(\mathrm{Supp}(\mu)) \iff y \in \bar{f}(\mathrm{Supp}(\mu))$.

CHAPTER 12

RELATIVE ENTROPY SOUP

Part VI

Conclusions

BIBLIOGRAPHY

[1] 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.

[2] A. Balke and J. Pearl. Probabilistic evaluation of counterfactual queries. In *Proc. Eleventh National Conference on Artificial Intelligence (AAAI '94)*, pages 200–207, 1994.

[3] Sander Beckers, Joseph Y. Halpern, and Christopher Hitchcock. Causal models with constraints, 2023.

[4] David Maxwell Chickering. A transformational characterization of equivalent bayesian network structures. *CoRR*, abs/1302.4938, 2013. URL http://arxiv.org/abs/1302.4938.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[6] Karl Friston. The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, 13(7):293–301, 2009.

[7] 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)90045-P. URL https://www.sciencedirect.com/science/article/pii/0166218X9390045P.

[8] 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.

[9] 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.

[10] Joseph Y Halpern. Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research*, 12:317–337, 2000.

[11] Joseph Y Halpern. *Reasoning about uncertainty*. MIT press, 2017.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[18] 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.

[19] D. Koller and N. Friedman. *Probabilistic Graphical Models*. MIT Press, Cambridge, MA, 2009.

[20] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[21] 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.

[22] 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.

[23] leonbloy (https://math.stackexchange.com/users/312/leonbloy). conditioning reduces mutual information. Mathematics Stack Ex-

change, 2015. URL https://math.stackexchange.com/q/1219753. URL:https://math.stackexchange.com/q/1219753 (version: 2015-04-04).

[24] David J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[25] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[26] Pavel Naumov and Brittany Nicholls. R.e. axiomatization of conditional independence, 2013.

[27] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, 1988.

[28] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000.

[29] J Pearl and A Paz. Graphoids: A graphbased logic for reasoning about relevance relations. advances in artificial intelligence, vol. ii, 1987.

[30] Judea Pearl. *Causality*. Cambridge university press, 2009.

[31] Oliver E Richardson. Loss as the inconsistency of a probabilistic dependency graph: Choose your model, not your loss function. *AISTATS '22*, 151, 2022.

[32] Oliver E Richardson and Joseph Y Halpern. Probabilistic dependency graphs. In *Proc. Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, pages 12174–12181, 2021.

[33] Oliver E Richardson, Joseph Y Halpern, and Christopher De Sa. Inference for probabilistic dependency graphs. In *Uncertainty in Artificial Intelligence*, pages 1741–1751. PMLR, 2023.

[34] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[35] 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.

[36] 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.

[37] Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information, 2010.