CONDENSATION: A THEORY OF CONCEPTS

SAM EISENSTAT

Abstract. We understand agents as creating concepts in order to organize their understanding of the world, and as often sharing concepts and so being able to work out how to make sense of the world together, forming language communities. Here, we look at how probability distributions can be organized by introducing appropriate latent variables, which we aim to use as a model of these phenomena. Our main result shows that under certain information-theoretic hypotheses, different systems of latent variables stand in a kind of correspondence.

Contents

1. Introduction

> "Entification begins at arm's length; the points of condensation in the primordial conceptual scheme are things glimpsed, not glimpses."
>
> W. V. Quine [Qui60]

We are concerned here with understanding, as may be possessed by an agent, shared in a language community, or presented in a scientific theory. In particular, we will examine a mathematical model intended to show some aspects of this phenomenon which we otherwise might see less clearly.

Our motivating idea of meaning goes as follows. We possess and create bodies of meaning. We may regard the traditional logical elements—propositions, entities, relations—as among their conceptual constituents. We introduce such concepts—say, entities like trees, numbers, or people—and they help us organize the world. We do not always introduce new elements by definition. For example, one cannot define physical objects in terms of one's visual perceptions, though one can make inferences from perception.

Instead, the physical objects have a reality that surpasses their connection to one's visual perception. We expect the objects to also be perceived by other senses, and by other people. If one changes somewhat one's understanding of the connection between objects and perception, we expect that roughly the same objects will still be comprehended, that there will be a correspondence for the most part between old and new. Even without a source in shared culture, as with babies or upon making new contact with people with a different way of life, we expect to have enough meaning in common to be able to get started in working out a greater shared understanding over time.

Here, we will try to say these things in the language of probability theory, to the extent possible. Something postulated that goes beyond one's observations can be understood as a parameter of a statistical models, or a latent variable. (Reflecting a Bayesian view, we won't try to distinguish these.) But we want to be more precise about the role of latent variables here. One motive behind introducing a latent variable is as a "mere postulate"—as we conceive it, its only role is to predict the observable variables more accurately. However, we can point out other reasons, following the above discussion. We might like latent variables to be *intersubjective*, in that other agents represent the world in terms of similar variables. We also might like latent variables to be more deeply integrated into a body of meaning, and thus to be things that we can look at many ways, as it suits us, rather than acting like "black boxes", which we only use to predict something else. To these reasons we can add another. We may have certain concepts which one has reason to use even if one only wants to predict, but which we want to also use to understand our values, such as concepts dealing with the experience of others.

With this motivation, we give conditions under which certain latent variables of different probabilistic models will admit a kind of correspondence. We can take this to give some indication of how and why different agents might be expected to use corresponding concepts to understand the world, even before they have been inducted into a shared language community. This is carried out in Sections 4 and 5. Section 4 uses rather restricted hypotheses, in order to illustrate the idea, in Theorem 4.15, and Section 5 generalizes the idea using quantitative hypotheses with more interesting examples in Theorem 5.8, which demonstrates a kind of approximate correspondence.

We can related these ideas to others in statistical theory. Many traditional parametric models introduce variables whose meaningfulness is created by human understanding of the subject matter. In other contexts, such as latent causal discovery with structural causal models [SGS01; PJS17], hidden Markov models, independent component analysis, factor analysis[Bis06; Mac02], sparse autoencoders[Cun+23], and factored space models[Gar+24], we introduce particular latent variables as part of a statistical method, and we often hope that these variables will make sense to us, and will fit into our bodies of meaning. We aim for the theory introduced here to be able to clarify how these ideas work, and what we are asking for when we ask that they discover meaningful variables. We do not, though, initiate such an analysis here.

The particular form of latent variable models introduced here most parallels structural causal models and factored space models. One point of contrast is that these theories are organized around determination and conditional independence, whereas here we will express things using inequalities, in terms of entropy and

mutual information. We should note that information-theoretic methods are widely used in work on structural causal models though.

The aim towards intersubjectivity here can be seen in the context of the work of de Finetti [Fin29] on exchangability and Wentworth and Lorell [WL24] on natural latents. The principal difference in approach is that we seek here to work with many latent variables, which together form a kind of system of meaning—a model of the world structured in terms of a set of diverse conceptual parts.

2. Background and notation

First, we review some basic points about probability theory in order to establish some conventions for our discussion.

Definition 2.1. A *random variable* is a measurable function $X\colon \Omega \to R$ between measurable spaces. We will say that X is a random variable *on* Ω and *valued* in R, or with *range* R.

However, we will consider many measurable functions here, but we will only call some of them random variables. When we call a measurable function a random variable, we indicate that we intend to use the following forms of expression. First, and most importantly, we may talk about a random variable valued in R as if it is an element of R. For example, if X is a random variable valued in R and $f\colon R \to S$ is a measurable function, we write $f(X)$ to mean $f \circ X$, and given a pair of random variables $X\colon \Omega \to R$ and $Y\colon \Omega \to S$, we write (X, Y) for the random variable $\Omega \to R \times S$ defined as

$$\omega \mapsto (X(\omega), Y(\omega)). \tag{2.1}$$

We may treat random variables as elements of their ranges in other such ways if we feel the meaning to be clear. This idea is the main reason for the random-variable concept, but we will also establish some other conventions involving random variables.

Second of all, if X is a random variable on Ω and $\pi\colon \Lambda \to \Omega$ is a measurable map, then we call the random variable $X \circ \pi$, which is defined on Λ and has the same range as X, the *pullback* of X by π, and we denote it symbolically as $\pi^* X$. However, when we feel that the meaning is clear, we will just write the pullback as X. We can get away with this because probabilistic concepts are preserved here. For example, if Ω and Λ have the structure of probability spaces, $\pi\colon \Lambda \to \Omega$ is measure preserving, and X and Y are random variables on Ω, then the mutual information of X and Y satisfies

$$I(X;Y) = I(\pi^* X; \pi^* Y). \tag{2.2}$$

Following a similar idea, we can write expressions like $I(X;Y \mid Z)$, where Z is a random variable on Λ. This can only mean $I(\pi^* X; \pi^* Y \mid Z)$, since the pullback lets us make sense of X and Y as random variables on Λ, but we don't have a convention for making sense of Z as a random variable on Ω.

Third, if X is a random variable, we may use the notation $\mathcal{R}X$ to denote the codomain of X. Fourth, if $X\colon \Omega \to R$ is a random variable and the domain Ω is given the structure of a probability space (i.e. it is equipped with a probability measure $\mathbf{P}$), then we call the probability measure $X_* \mathbf{P}$ is the *distribution* of X. Fifth, again given a random variable $X\colon \Omega \to R$ on a probability space $(\Omega, \mathbf{P})$, if the range R is given the structure of a subset of a vector space V, then we call the

integral of X with respect to $\mathbf{P}$ the *expectation* of X. Fifth, if X and Y are random variables, we may say that Y *is a function of* X, with the meaning that there is a measurable function $f\colon \mathcal{R}X \to \mathcal{R}Y$ such that $Y = f(X)$. Analogously, we may say that Y *is a function of* X *almost everywhere*.

Next, we'll fix some notations.

Definition 2.2. The expression $\mathcal{P}^+S$ will denote the *nonempty power set* of a set S, that is, the set of all nonempty subsets of S.

Definition 2.3. We use the standard notations $H(X)$ and $H(X \mid Y)$ for the *entropy* of a random variable X and the *conditional entropy* of a random variable X given a random variable Y. We also use $I(X;Y \mid Z)$ to denote the *mutual information* of random variables X and Y given a random variable Z. Sometimes we write multiple random variables in such an expression using commas, such as in $H(X,Y)$; this means the entropy of the product random variable (X,Y)—this quantity is known as the *joint entropy* of X and Y. Other such expressions have corresponding meanings.

We will also have need for the *interaction information*,

$$I(X;Y;Z) = I(X;Y) - I(X;Y \mid Z). \tag{2.3}$$

Note that this quantity is invariant under permutation of its arguments.

While many of our arguments readily generalize to more continuous settings, where Halmos [Hal59] defines information-theoretic quantities in a more general sense, we will generally assume here for simplicity that our probability spaces are countable and discrete, and have finite entropy.

It will simplify a few things to equip the space of probability measures on a measurable space with the structure of a measurable space. This idea has been carried much further, especially by Giry [Gir82] and Fritz [Fri20].

Definition 2.4. The space of probability measures on a measurable space Ω can itself be equipped with the structure of a measurable space, using the smallest σ-algebra such that for each measurable set $E \subseteq \Omega$, the map

$$\mathbf{P} \mapsto \mathbf{P}(E) \tag{2.4}$$

is measurable. This space will be denoted $G(\Omega)$.

Proposition 2.5. *Let X and Y be random variables with finite entropy and countable discrete range on a countable discrete probability space $(\Omega, \mathbf{P})$, and suppose that $H(Y \mid X) = 0$. Then, then there is a measurable function $f\colon \mathcal{R}X \to \mathcal{R}Y$ such that $Y = f(X)$ almost everywhere.*

Proof. Let $A \subseteq \mathcal{R}X$ be the set of those x such that the singleton $\{x\}$ has positive measure under the pushforward $X_*\mathbf{P}$. Since $\mathcal{R}X$ is countable, A has full measure. Everywhere on the preimage $X^{-1}A$, we can define the conditional probability distribution $\mathbf{P}(\cdot \mid X = x)$. From the definition of conditional entropy, each such probability distribution is a Dirac measure in $G(\Omega)$ for almost every $x \in \mathcal{R}X$. This gives us a function $\widetilde{f}\colon A \to \mathcal{R}Y$ such that

$$Y = \widetilde{f}(X) \tag{2.5}$$

almost everywhere. Since $\mathcal{R}X$ is discrete, we can extend this to a function $f\colon \mathcal{R}X \to \mathcal{R}Y$, which is automatically measurable. □

3. Set-up

In this section, we will introduce the central concepts of latent variable models and latent string models. We intend for latent variable models to organize the structure of random variable models by positing additional latent variables, which cannot necessarily be defined from the given random variables. However, our definition of latent variable models will be rather weak. We ask that the given variables can be recovered from the latent variable, but we don't ask that this serve any organizing role, we don't ask that we attain an enlightening perspective on the given variables. So, we supplement this using various scoring functions. A latent variable model that gets a good score may more likely help us understand the underlying random variable model.

Now, we can proceed with our objects of study.

Definition 3.1. A *random variable model* is a countable discrete probability space Ω with finite entropy, together with a finite family of random variables $X_i \colon \Omega \to R_i$, each of which has countable and discrete range.

Our aim here is to understand random variable models by means of auxiliary random variables, which we'll call latent variables. In particular, in a random variable model $\left(\Omega, (X_i)_{i \in I}\right)$, we want the random variables X_i to be functions of certain latent variables. We won't necessarily define these latent variables on Ω; instead we might need an extension of the probability space. This leads to a definition.

Definition 3.2. A *latent variable model* for a random variable model $\left(\Omega, (X_i)_{i \in I}\right)$ is an ordered pair consisting of a random variable model $\left(\Lambda, (Y_A)_{A \in \mathcal{P}^+ I}\right)$ with its random variables indexed by the power set of I, together with a probability-preserving map $\pi \colon \Lambda \to \Omega$ such that for each random variable X_i the pullback $\pi^* X_i$ is almost everywhere a function of the random variables Y_A such that $A \subseteq I$ and $A \ni i$. In other words, $\pi^* X_i$ is almost everywhere equal to $f_i \left(Y_A \colon A \subseteq I, A \ni i\right)$ for some measurable function f_i from the product of the ranges of the random variables $(Y_A)_{A \subseteq I, A \ni i}$ to the range of X_i. We call the variables $(Y_A)_{A \in \mathcal{P}^+ I}$ *latent variables*, and if $i \in A$, we say that index A *contributes* to i.

We can compare different latent variable models using scoring functions. Lower scores should be "better"; we'll see how later.

Definition 3.3. Let $\mathcal{M}$ be a random variable model, with random variables $(X_i)_{i \in I}$, and $\mathcal{L}$ an associated latent variable model with latent variables $(Y_A)_{A \in \mathcal{P}^+ I}$. The *simple score* of $\mathcal{L}$ at $A \subseteq I$ is

$$\sigma_{\mathcal{L}}(A) = \sum_{\substack{B \in \mathcal{P}^+ I \\ B \cap A \neq \emptyset}} H(Y_B), \tag{3.1}$$

and correspondingly, the *conditioned score* of $\mathcal{L}$ at A is

$$\chi_{\mathcal{L}}(A) = \sum_{B \cap A \neq \emptyset} H\left(Y_B \mid (Y_C)_{C \supsetneq B}\right). \tag{3.2}$$

Finally, the *reconstruction score* of $\mathcal{L}$ at A is

$$\varrho_{\mathcal{L}}(A) = H\left((Y_B)_{B \supseteq A} \mid (X_i)_{i \in A}\right). \tag{3.3}$$

All these quantities are finite, since all the random variables of the form X_i and Y_A have finite entropy. These scores are in some sense local, measuring the complexity of the latent model as it pertains to the product of the random variables $(X_i)_{i\in A}$. To score the latent model $\mathcal{L}$ as a whole, we could perform an aggregation of these scores. For example, for a latent model $\mathcal{L}$ with latent variables $(Y_A)_{A\in\mathcal{P}^+I}$,

$$\begin{aligned}\sigma^\lambda_{\mathcal{L}} &= \sum_{A\in\mathcal{P}^+I} \lambda_A \sigma_{\mathcal{L}}(A) \\ &= \sum_{A\in\mathcal{P}^+I} \sum_{B\cap A\neq\emptyset} \lambda_A H(Y_B) \\ &= \sum_{B\subseteq I} \left(\sum_{A\cap B\neq\emptyset} \lambda_A \right) H(Y_B)\end{aligned} \tag{3.4}$$

and the analogous statements holds for the conditioned score as well. We will not pursue that route further here.

In order to work with the latent variable models, we define some notation for random variables.

Definition 3.4. Let $\mathcal{M}$ be a random variable model with variables $(X_i)_{i\in I}$ and $\mathcal{L}$ an associated latent variable model with variables $(Y_A)_{A\in\mathcal{P}^+I}$. We will write X and Y with certain subscripts other than elements, respectively, of I and $\mathcal{P}^+I$ to denote certain products of the random variables in these families. If $A\subseteq I$, we write X_A to denote the *joint random variable* at A, which is the product random variable $(X_i)_{i\in A}$. Similarly, for any $\mathcal{F}\subseteq\mathcal{P}^+I$, we write $Y_{\mathcal{F}}$ to denote the joint random variable $(Y_A)_{A\in\mathcal{F}}$. We also define the following notations:

$$Y_{\cap A} = \left(Y_B \colon B\in\mathcal{P}^+I, B\cap A\neq\emptyset\right) \tag{3.5}$$

$$Y_{\supseteq A} = \left(Y_B \colon B\in\mathcal{P}^+I, A\subseteq B\right) \tag{3.6}$$

$$Y_{\supsetneq A} = \left(Y_B \colon B\in\mathcal{P}^+I, A\subsetneq B\right) \tag{3.7}$$

$$Y_{\ni i} = \left(Y_B \colon B\in\mathcal{P}^+I, i\in B\right). \tag{3.8}$$

In particular, note that for any $i\in I$,

$$Y_{\ni i} = Y_{\cap\{i\}} = Y_{\supseteq\{i\}}. \tag{3.9}$$

In words, we call the random variables $Y_{\cap A}$ and $Y_{\ni i}$ the latents that *contribute* to A or i, respectively.

3.1. **Morphisms.** We will also define morphisms of random variable models. We'll give a bit of the idea first. A probability-preserving map of probability spaces $\pi\colon\Omega\to\Lambda$ forgets distinctions. We can think of Ω as an extension of Λ. In other words, we can say that Ω has all the measurable sets $\pi^{-1}E$ corresponding to measurable sets E in Λ, but it can also have other measurable sets. So, a morphism of random variable models will be similar, but we also account for the random variables named by our index set. In particular, we correspondingly let the random variables of the source model make more distinctions than those of the target model. Now, we'll say all this more precisely.

Definition 3.5. A *morphism* of random variable models has the form

$$\left(\pi, \iota, (f_j)_{j\in J}\right) \colon \left(\Omega, (X_i)_{i\in I}\right) \to \left(\Lambda, (Y_j)_{j\in J}\right), \tag{3.10}$$

where $\pi\colon \Omega \to \Lambda$ is a probability-preserving map, ι is a function $J \to I$, and f_j is a function from the range of $X_{\iota(j)}$ to the range of Y_j. We require, for all $j \in J$, that $Y_j = f_j\left(X_{\iota(j)}\right)$ almost everywhere on Ω. Note that the function f_j is automatically measurable, since these random variables have countable and discrete range, and from the same premises that the condition $Y_j = f_j\left(X_{\iota(j)}\right)$ defines a measurable set.

Making the pullback explicit, we can write this as $\pi^* Y_j = f_j\left(X_{\iota(j)}\right)$. Also, note that this is equivalent to the condition that on a set of full measure in Ω, we have $Y_j = f_j\left(X_{\iota(j)}\right)$ for all $j \in J$.

(We may observe that the map π in a latent variable model $((\Lambda, Y), \pi)$ is not necessarily a morphism of random variable models here, since each random variable X_i may depend nontrivially on multiple latent variables.)

Definition 3.6. Given two morphisms

$$\left(\pi, \iota, (f_j)_{j\in J}\right)\colon \left(\Omega, (X_i)_{i\in I}\right) \to \left(\Lambda, (Y_j)_{j\in J}\right) \tag{3.11}$$

$$\left(\rho, \nu, (g_k)_{k\in K}\right)\colon \left(\Lambda, (Y_j)_{j\in J}\right) \to \left(\Pi, (Z_k)_{k\in K}\right), \tag{3.12}$$

their *composite* is the morphism

$$\left(\rho\circ\pi, \iota\circ\nu, \left(g_k \circ f_{\nu(k)}\right)_{k\in K}\right)\colon \left(\Omega, (X_i)_{i\in I}\right) \to \left(\Pi, (Z_k)_{k\in K}\right). \tag{3.13}$$

We can confirm that this is well-defined, checking in particular that $Z_k = g_k \circ f_{\nu(k)}\left(X_{\iota\circ\nu(k)}\right)$ for all $k \in K$ almost everywhere. We know that, for all $k \in K$,

$$\rho^* Z_k = g_k\left(Y_{\nu(k)}\right) \tag{3.14}$$

almost everywhere, so since π is probability-preserving, we have almost everywhere that

$$\begin{aligned}(\rho\circ\pi)^* Z_k = \pi^*\rho^* Z_k = \pi^* g_k\left(Y_{\nu(k)}\right) = g_k\left(\pi^* Y_{\nu(k)}\right)\\ = g_k \circ f_{\nu(k)}\left(X_{\iota\circ\nu(k)}\right).\end{aligned} \tag{3.15}$$

as desired.

Proposition 3.7. *Random variable models and morphism form a category.*

Proof. On a random variable model $\left(\Omega, (X_i)_{i\in I}\right)$, we have the morphism

$$\left(\mathrm{id}_\Omega, \mathrm{id}_I, (\mathrm{id}_{\mathcal{R}\, X_i})_{i\in I}\right), \tag{3.16}$$

which we can see serves as an identity morphism. Further, we can see that composition is associative by checking associativity for any three morphisms

$$\left(\Omega_1, (W_i)_{i\in I}\right) \xrightarrow{\left(\pi, \iota, (f_j)_{j\in J}\right)} \left(\Omega_2, (X_j)_{j\in J}\right) \xrightarrow{\left(\rho, \nu, (g_k)_{k\in K}\right)} \left(\Omega_3, (Y_k)_{k\in K}\right) \xrightarrow{\left(\sigma, o, (h_\ell)_{\ell\in L}\right)} \left(\Omega_4, (Z_\ell)_{\ell\in L}\right). \tag{3.17}$$

The composite of the first two morphisms is

$$\left(\rho\circ\pi, \iota\circ\nu, \left(g_k \circ f_{\nu(k)}\right)_{k\in K}\right) \tag{3.18}$$

and that of the last two is

$$\left(\sigma\circ\rho, \nu\circ o, \left(h_\ell \circ g_{o(\ell)}\right)_{\ell\in L}\right), \tag{3.19}$$

so the triple composite, interpreted in either order, is

$$\left(\sigma \circ \rho \circ \pi, \iota \circ \nu \circ o, \left(h_\ell \circ g_{o(\ell)} \circ f_{\nu(o(\ell))}\right)_{\ell \in L}\right). \tag{3.20}$$

□

Because of the phenomenon of equality almost everywhere, there are a few different notions of equivalence that we may use in discussing this category.

Proposition 3.8. *A morphism of random variable models*

$$\left(\pi, \iota, (f_j)_{j \in J}\right) : \left(\Omega, (X_i)_{i \in I}\right) \to \left(\Lambda, (Y_j)_{j \in J}\right) \tag{3.21}$$

is an isomorphism if and only if π is an isomorphism of measurable spaces; ι is a bijection; and for every $j \in J$, the map f_j is an isomorphism of measurable spaces.

Proof. It is clear that isomorphisms have these properties. Conversely, if a morphism $\left(\pi, \iota, (f_j)_{j \in J}\right)$ has these properties, consider the triple $\left(\pi^{-1}, \iota^{-1}, \left(f^{-1}_{\iota^{-1}(i)}\right)_{i \in I}\right)$. Since π is probability-preserving and is an isomorphism of measurable spaces, π^{-1} is also probability-preserving. For all $i \in I$, we have

$$Y_{\iota^{-1}(i)} = f_{\iota^{-1}(i)}\left(X_i\right) \tag{3.22}$$

almost everywhere, so

$$f^{-1}_{\iota^{-1}(i)}\left(Y_{\iota^{-1}(i)}\right) = X_i, \tag{3.23}$$

almost everywhere, and we see that our triple is in fact a morphism. It is immediate that it is an inverse to $\left(\pi, \iota, (f_j)_{j \in J}\right)$, so that map is an isomorphism. □

Definition 3.9. Morphisms $\left(\pi, \iota, (f_j)_{j \in J}\right)$ and $\left(\rho, \nu, (g_j)_{j \in J}\right)$ from (Ω, X) to (Λ, Y) are *equal almost everywhere* if

1. the maps π and ρ are equal almost everywhere as measurable functions, and
2. the functions ι and ν are equal.

Note that we put no further condition on f and g. It is possible that they are unequal; by the definition of morphisms, we have for all j that

$$f_j\left(X_{\iota(j)}\right) = Y_j = g_j\left(X_{\nu(j)}\right) = g_j\left(X_{\iota(j)}\right) \tag{3.24}$$

almost everywhere (on Ω), but not necessarily everywhere. Even if $f_j\left(X_{\iota(j)}\right) = g_j\left(X_{\iota(j)}\right)$ everywhere, we may have $f_j \neq g_j$ since $X_{\iota(j)}$, considered as a measurable function, may not be surjective.

Definition 3.10. Two random variable models $\mathcal{M}$ and $\mathcal{N}$ are *equivalent* if there are morphisms

$$\boldsymbol{\pi} = \left(\pi, \iota, (f_j)_{j \in J}\right) : \mathcal{M} \to \mathcal{N} \tag{3.25}$$

$$\boldsymbol{\rho} = \left(\rho, \nu, (g_i)_{i \in I}\right) : \mathcal{N} \to \mathcal{M} \tag{3.26}$$

such that $\boldsymbol{\rho} \circ \boldsymbol{\pi}$ and $\boldsymbol{\pi} \circ \boldsymbol{\rho}$ are, respectively, equal almost everywhere to the identity morphisms on $\mathcal{M}$ and $\mathcal{N}$. In this case, we also say that the pair $(\boldsymbol{\pi}, \boldsymbol{\rho})$ is an *equivalence*.

We can say informally that an equivalence is an isomorphism almost everywhere. To express this another way, suppose that $\left(\Omega, (X_i)_{i\in I}\right)$ and $\left(\Lambda, (Y_i)_{i\in I}\right)$ are random variable models, $\pi\colon \Omega \to \Lambda$ and $\rho\colon \Lambda \to \Omega$ are probability-preserving maps, and $f_i\colon \mathcal{R}X_i \to \mathcal{R}Y_i$ and $g_i\colon \mathcal{R}Y_i \to \mathcal{R}X_i$ are measurable maps for every $i \in I$. We'd like to know when the obvious triples we can make from these are a pair of morphisms constituting an equivalence. Laying out the definitions, we see that this holds if and only if

(1) $\rho \circ \pi$ and $\pi \circ \rho$ are respectively equal almost everywhere to id_Ω and id_Λ, and
(2) $f_i(X_i) = \pi^* Y_i$ and $g_i(Y_i) = \rho^* X_i$ almost everywhere for all $i \in I$.

We are taking a little more care by making the pullbacks explicit here, to avoid the potential for ambiguity.

We can say a few things to establish that these notions behave as we expect. In categorical language, the next proposition amounts to saying that random variable models, morphisms of random variable models, and equality almost everywhere together form a (strict) 2-category. We won't use the language of 2-categories further here though.

Proposition 3.11. *Equality almost everywhere of morphisms of random variable models is a congruence with respect to composition. That is,*

(1) *equality almost everywhere of morphisms of random variable models is an equivalence relation, and*
(2) *given random variable models $\mathcal{L}$, $\mathcal{M}$, and $\mathcal{N}$, and morphisms*

$$\boldsymbol{\pi}, \boldsymbol{\rho}\colon \mathcal{L} \to \mathcal{M}, \qquad \boldsymbol{\sigma}, \boldsymbol{\tau}\colon \mathcal{M} \to \mathcal{N} \tag{3.27}$$

such that $\boldsymbol{\pi}$ is equal almost everywhere to $\boldsymbol{\rho}$, and $\boldsymbol{\sigma}$ is to $\boldsymbol{\tau}$, the composite $\boldsymbol{\sigma} \circ \boldsymbol{\pi}$ is equal almost everywhere to $\boldsymbol{\tau} \circ \boldsymbol{\rho}$.

Proof. (1) is immediate. To confirm (2), we will verify that the underlying measurable maps of the morphisms $\boldsymbol{\sigma} \circ \boldsymbol{\pi}$ and $\boldsymbol{\tau} \circ \boldsymbol{\rho}$ are equal. Let's use lightface symbols to denote the underlying measurable maps of morphisms denoted with corresponding boldface symbols. Then, π and ρ agree on a set $E \subseteq \Omega$ of full measure, and σ and τ similarly agree on such a set $F \subseteq \Lambda$. The set $E \cap \pi^{-1}(F)$ has full measure, and for all ω in this set

$$\sigma \circ \pi(\omega) = \tau \circ \pi(\omega) = \tau \circ \rho(\omega), \tag{3.28}$$

as desired. □

Since we have a 2-category, it follows that equivalence of random variable models is also an equivalence relation. We'll spell this out a bit more.

Proposition 3.12. *Equivalence of random variable models is an equivalence relation.*

Proof. Reflexivity and symmetry are immediate. Suppose $(\boldsymbol{\pi}, \boldsymbol{\rho})$ is an equivalence between $\mathcal{L}$ and $\mathcal{M}$, and $(\boldsymbol{\sigma}, \boldsymbol{\tau})$ is an equivalence between $\mathcal{M}$ and $\mathcal{N}$. Then, $(\boldsymbol{\sigma} \circ \boldsymbol{\pi}, \boldsymbol{\rho} \circ \boldsymbol{\tau})$ is an equivalence between $\mathcal{L}$ and $\mathcal{N}$, since

$$(\boldsymbol{\rho} \circ \boldsymbol{\tau}) \circ (\boldsymbol{\sigma} \circ \boldsymbol{\pi}) = \boldsymbol{\rho} \circ \boldsymbol{\pi} = \mathrm{id}_{\mathcal{L}} \tag{3.29}$$

and the opposite composite is similarly $\mathrm{id}_{\mathcal{N}}$, so we see that equivalence is also transitive. □

4. Perfect condensation

In order to understand our scoring functions, we will ask some questions broadly following two directions of inquiry. First, what is a "good" score? When is a score good enough that we should be interested in a latent variable model that attains that score? Second, what can we conclude about the structure of a latent variable model that gets a good score? We will start with latent variable models rather than latent string models, though the ideas of this section apply to both. Ultimately, latent string models have more interesting applications, but latent variable models are more mathematically convenient. First, we note that we do indeed have uninteresting latent variable models with bad scores.

Example 4.1. Let $\mathcal{M} = \left(\Omega, (X_i)_{i\in I}\right)$ be a random variable model. Consider the latent variable models $\mathcal{L}_1$ and $\mathcal{L}_2$ associated with $\mathcal{M}$, defined as follows. First, $\mathcal{L}_1$ and $\mathcal{L}_2$ have the same underlying probability space as $\mathcal{M}$, that is,

$$\mathcal{L}_1 = \left(\left(\Omega, (Y_A)_{A\in\mathcal{P}^+I}\right), \mathrm{id}_\Omega\right) \qquad \mathcal{L}_2 = \left(\left(\Omega, (Z_A)_{A\in\mathcal{P}^+I}\right), \mathrm{id}_\Omega\right) \tag{4.1}$$

for some families Y and Z of random variables. We will set $Y_{\{i\}} = X_i$ for $i \in I$. For $A \in \mathcal{P}^+I$ with $|A| \neq 1$, let Y_A be constant. Next, let $Z_I = X_I$, and let Z_A be constant for $A \subsetneq I$. If all the following quantities are defined, we have

$$\sigma_{\mathcal{L}_1}(A) = \sum_{B\cap A\neq\emptyset} H(Y_B) = \sum_{i\in A} H(X_i) \tag{4.2}$$

$$\chi_{\mathcal{L}_1}(A) = \sum_{B\cap A\neq\emptyset} H(Y_B \mid Y_{\supsetneq B}) = \sum_{i\in A} H(X_i) \tag{4.3}$$

$$\sigma_{\mathcal{L}_2}(A) = \sum_{B\cap A\neq\emptyset} H(Z_B) = H(Z_I) = H(X_I) \tag{4.4}$$

$$\chi_{\mathcal{L}_2}(A) = \sum_{B\cap A\neq\emptyset} H(Z_B \mid Z_{\supsetneq B}) = H(X_I). \tag{4.5}$$

Since we didn't use anything about the structure of $\mathcal{M}$ to produce these latent variable models, we expect that they don't tell us much about $\mathcal{M}$, at least in the typical case. So, these should usually be "bad" scores. If we want to produce even worse scores, we could add more entropy to the latent variables in a way that is irrelevant to determining the variables X_i.

Now, we can establish some easy lower bounds on the simple and conditioned scores.

Proposition 4.2. *Let $\left(\Omega, (X_i)_{i\in I}\right)$ be a random variable model and $\mathcal{L}$ an associated latent variable model with latent variables $(Y_A)_{A\in\mathcal{P}^+I}$. Then, for any $A \subseteq I$, we have*

$$\sigma_{\mathcal{L}}(A) \geq \chi_{\mathcal{L}}(A) \geq H(Y_{\cap A}) \geq H(X_A). \tag{4.6}$$

Proof. It is immediate that $\sigma_{\mathcal{L}}(A) \geq \chi_{\mathcal{L}}(A)$. To see that $\chi_{\mathcal{L}}(A) \geq H(Y_{\cap A})$, we proceed as follows. The set

$$\left\{B \mid B \in \mathcal{P}^+I, B\cap A \neq \emptyset\right\} \tag{4.7}$$

is partially ordered by the inclusion relation $B_1 \supseteq B_2$. This extends to a total order, i.e. there exists a total order $\preceq$ on this set such that whenever $B_1 \supseteq B_2$, we

have $B_1 \preceq B_2$. Thus, using the nonnegativity of mutual information, we have

$$H\left(Y_B \mid (Y_C)_{C \supsetneq B}\right) \geq H\left(Y_B \mid (Y_C)_{C \prec B}\right). \tag{4.8}$$

Now we can establish the next inequality by a calculation:

$$\begin{aligned} \chi_{\mathcal{L}}(A) &= \sum_{B \cap A \neq \emptyset} H\left(Y_B \mid (Y_C)_{C \supsetneq B}\right) \\ &\geq \sum_{B \cap A \neq \emptyset} H\left(Y_B \mid (Y_C)_{C \prec B}\right) \\ &= H\left(Y_{\cap A}\right). \end{aligned} \tag{4.9}$$

Finally, the random variable X_A is a function of $Y_{\cap A}$ almost everywhere by definition, so $H\left(Y_{\cap A}\right) \geq H\left(X_A\right)$. □

This motivates a definition of perfect condensation.

Definition 4.3. A latent variable model $\mathcal{L}$ *perfectly condenses* a random variable model $\mathcal{M} = \left(\Omega, (X_i)_{i \in I}\right)$ if $\chi_{\mathcal{L}}(A) = H\left(X_A\right)$ for all $A \subseteq I$. Further, $\mathcal{L}$ *simply-perfectly condenses* $\mathcal{M}$ if $\sigma_{\mathcal{L}}(A) = H\left(X_A\right)$ for all $A \subseteq I$.

Example 4.4. Let I be an index set, and consider any random variable model $\mathcal{L} = \left(\Omega, (Y_A)_{A \in \mathcal{P}^+ I}\right)$, indexed by the nonempty power set of I, such that the variables Y_A are jointly independent. We will construct a random variable model $\mathcal{M} = \left(\Omega, (X_i)_{i \in I}\right)$ such that $\mathcal{M}$ is perfectly condensed by $\mathcal{L}$, as related to $\mathcal{M}$ via the identity map id_Ω. For each $i \in I$, we define X_i to be the product random variable

$$X_i = Y_{\ni i} = \left(Y_A \colon i \in A \subseteq I\right). \tag{4.10}$$

Now, for any set $A \subseteq I$, we have

$$\begin{aligned} H\left(X_A\right) = H\left(X_i \colon i \in A\right) &= H\left(Y_B \colon B \cap A \neq \emptyset\right) \\ &= \sum_{B \cap A \neq \emptyset} H\left(Y_B\right), \end{aligned} \tag{4.11}$$

using the independence assumption in the last step. This is just the simple score, so $\mathcal{L}$ simply-perfectly condenses $\mathcal{M}$.

We can say quite a lot about a latent variable model if we know that it is a perfect condensation or a simple-perfect condensation.

Lemma 4.5. *Let $\mathcal{M}$ be a random variable model with random variables $(X_i)_{i \in I}$, let $\mathcal{L}$ be an associated latent variable model with latent variables $(Y_A)_{A \in \mathcal{P}^+ I}$. Then, the following are equivalent.*

1. *For all $A \in \mathcal{P}^+ I$, we have $H\left(Y_{\cap A}\right) = H\left(X_A\right)$.*
2. *For all $i \in I$ and $A \in \mathcal{P} I$ such that $i \in A$, there is some measurable function $f_A^i \colon \mathcal{R} X_i \to \mathcal{R} Y_A$ such that $Y_A = f_A^i\left(X_i\right)$ almost everywhere.*

Proof. ($\Longrightarrow$) For each $i \in A$, we have

$$H\left(X_i\right) = H\left(Y_{\ni i}\right), \tag{4.12}$$

so since X_i is a function of $Y_{\ni i}$ almost everywhere,

$$H\left(Y_{\ni i} \mid X_i\right) = H\left(Y_{\ni i}\right) - H\left(X_i\right) = 0. \tag{4.13}$$

Hence, using Corollary 2.5 $Y_{\ni i}$, and a fortiori Y_A for any A containing i, is almost everywhere equal to a function of X_i.

($\Longleftarrow$) For any $A \in \mathcal{P}^+ I$, we have $H(Y_{\cap A}) \geq H(X_A)$ by Proposition 4.2. Further, picking any $i \in A$, the random variable $Y_{\cap A}$ is almost everywhere equal to a measurable function of X_i, and therefore to a measurable function of X_A, so

$$H(Y_{\cap A} \mid X_A) = 0 \tag{4.14}$$

$$H(Y_{\cap A}) \leq H(Y_{\cap A}, X_A) = H(X_A). \tag{4.15}$$

□

Corollary 4.6. *Let $\mathcal{M}$ be a random variable model with random variables $(X_i)_{i \in I}$, let $\mathcal{L}$ be an associated latent variable model with latent variables $(Y_A)_{A \in \mathcal{P}^+ I}$ that perfectly condenses $\mathcal{M}$. Then, whenever we have $i \in A \in \mathcal{P} I$, there is some measurable function $f_A^i \colon \mathcal{R} X_i \to \mathcal{R} Y_A$ such that $Y_A = f_A^i(X_i)$ almost everywhere.*

Proof. Using Proposition 4.2, this follows immediately. □

We can also express the conclusion of this corollary in terms of an equivalence.

Proposition 4.7. *Let $\mathcal{M} = \left(\Omega, (X_i)_{i \in I}\right)$ be a random variable model and $\mathcal{L} = \left(\left(\Lambda, (Y_A)_{A \in \mathcal{P}^+ I}\right), \pi\right)$ an associated latent variable model. Then, the following are equivalent.*

1. *For all $i \in I$ and $A \in \mathcal{P}^+ I$ such that $i \in A$, there is some measurable function $f_A^i \colon \mathcal{R} X_i \to \mathcal{R} Y_A$ such that $Y_A = f_A^i(X_i)$ almost everywhere.*
2. *$\left(\Lambda, (X_i)_{i \in I}\right)$ and $\left(\Lambda, \left(Y_{\cap \{i\}}\right)_{i \in I}\right)$ are equivalent as random variable models, via an equivalence of the form $\left(\mathrm{id}_\Lambda, \mathrm{id}_I, (g_i)_{i \in I}\right)$ and $\left(\mathrm{id}_\Lambda, \mathrm{id}_I, (h_i)_{i \in I}\right)$ for some families of functions g and h.*

Proof. ($\Longrightarrow$) By hypothesis, for each $i \in I$ and each $A \subseteq I$ satisfying $i \in A$, we have $Y_A = f_A^i(X_i)$ almost everywhere. Define $g_i \colon \mathcal{R} X_i \to \mathcal{R} Y_{\ni i}$ to be the product

$$g_i(x) = \left(f_A^i(x) : A \subseteq I, i \in A\right); \tag{4.16}$$

we can see that $\left(\mathrm{id}_\Lambda, \mathrm{id}_I, (g_i)_{i \in I}\right)$ is a morphism. Also, by the definition of latent variable model, we have functions $h_i \colon \mathcal{R} Y_{\ni i} \to \mathcal{R} X_i$ such that

$$X_i = h_i(Y_{\ni i}) \tag{4.17}$$

almost everywhere, so $\left(\mathrm{id}_\Lambda, \mathrm{id}_I, (h_i)_{i \in I}\right)$ is also a morphism. Now, it is immediate that we have an equivalence.

($\Longleftarrow$) Take any i and A satisfying $i \in A \subseteq I$. Since $\left(\mathrm{id}_\Lambda, \mathrm{id}_I, (g_j)_{j \in I}\right)$ is a morphism, we have

$$Y_{\ni i} = g_i(X_i) \tag{4.18}$$

almost everywhere. Let p_A^i be the coordinate projection

$$\mathcal{R} Y_{\ni i} = \prod_{B \ni i} \mathcal{R} Y_B \to \mathcal{R} Y_A. \tag{4.19}$$

Then,

$$Y_A = p_A^i(Y_{\ni i}) = p_A^i(g_i(X_i)), \tag{4.20}$$

so $p_A^i \circ g_i$ has the desired property. □

Corollary 4.6 tells us something about perfect, and hence simply-perfect, condensations. By imposing further conditions, we can define stronger properties, which will give us equivalences. First, we give a definition about probabilistic independence, which can be seen as a form of the Markov condition from the theory of structural causal models [SGS01].

Definition 4.8. Let I be a finite set, and suppose that $(Y_A)_{A\in\mathcal{P}^+I}$ are random variables on some probability space. The family Y satisfies the *ordered Markov condition* if the following statement holds.

- For any $A \in \mathcal{P}^+I$, let $\mathcal{F} \subseteq \mathcal{P}^+I$ be the collection of all $B \in \mathcal{P}^+I$ such that B is incomparable in the inclusion order to A, i.e. B is neither a subset nor a superset of A. Then, the random variables Y_A and $Y_{\mathcal{F}}$ are independent conditional on $Y_{\supsetneq A}$.

Theorem 4.9. *Let $\mathcal{M} = \left(\Omega, (X_i)_{i\in I}\right)$ be a random variable model and $\mathcal{L} = \left(\left(\Lambda, (Y_A)_{A\in\mathcal{P}^+I}\right), \pi\right)$ an associated latent variable model. The following are equivalent.*

(A1) *$\mathcal{L}$ is a simple-perfect condensation of $\mathcal{M}$.*
(A2) *For all $i \in I$ and $A \in \mathcal{P}^+I$ such that $i \in A$, the latent variable Y_A is a function of X_i almost everywhere. Further, the latent variables $(Y_A)_{A\in\mathcal{P}^+I}$ are jointly independent.*
(A3) *$\mathcal{L}$ is a perfect condensation of $\mathcal{M}$ and the latent variables $(Y_A)_{A\in\mathcal{P}^+I}$ are jointly independent.*

Further, the following are also equivalent:

(B1) *$\mathcal{L}$ is a perfect condensation of $\mathcal{M}$.*
(B2) *For all $i \in I$ and $A \in \mathcal{P}^+I$ such that $i \in A$, the latent variable Y_A is a function of X_i. Further, the latent variables obey the ordered Markov condition.*

Proof. (A1 $\implies$ A3) Since $\mathcal{L}$ is a simple-perfect condensation, it follows from Proposition 4.2 that for each $A \subseteq I$,

$$\sum_{B\cap A\neq\emptyset} H\left(Y_B\right) = \sigma_{\mathcal{L}}\left(A\right) \geq \chi_{\mathcal{L}}\left(A\right) \geq H\left(Y_{\cap A}\right) = \sigma_{\mathcal{L}}\left(A\right) \tag{4.21}$$

so all these quantities are equal. In particular,

$$\chi_{\mathcal{L}}\left(A\right) = H\left(Y_{\cap A}\right), \tag{4.22}$$

so $\mathcal{L}$ is a perfect condensation of $\mathcal{M}$. Further, it follows from

$$\sum_{B\in\mathcal{P}^+I} H\left(Y_B\right) = \sigma_{\mathcal{L}}\left(I\right) = H\left(Y_{\cap I}\right) \tag{4.23}$$

that the latent variables $(Y_A)_{A\in\mathcal{P}^+I}$ are jointly independent. To spell this out a bit more, consider any two disjoint families $\mathcal{F}, \mathcal{G} \subseteq \mathcal{P}^+I$, and let

$$\mathcal{H} = \mathcal{P}^+I - \mathcal{F} - \mathcal{G}. \tag{4.24}$$

Then, we can calculate

$$\begin{aligned} H\left(Y_{\cap I}\right) &\leq H\left(Y_{\mathcal{F}\cup\mathcal{G}}\right) + H\left(Y_{\mathcal{H}}\right) \\ &\leq H\left(Y_{\mathcal{F}}\right) + H\left(Y_{\mathcal{G}}\right) + H\left(Y_{\mathcal{H}}\right) \\ &\leq \sum_{B\in\mathcal{P}^+ I} H\left(Y_B\right) = H\left(Y_{\cap I}\right), \end{aligned} \tag{4.25}$$

so all these are equal, and so

$$I\left(Y_{\mathcal{F}}; Y_{\mathcal{G}}\right) = H\left(Y_{\mathcal{F}}\right) + H\left(Y_{\mathcal{G}}\right) - H\left(Y_{\mathcal{F}\cup\mathcal{G}}\right) = 0. \tag{4.26}$$

(A3 $\implies$ A2) This is immediate from Corollary 4.6.

(A2 $\implies$ A1) Each latent variable Y_A for $A \in \mathcal{P}^+ I$ is almost everywhere a function of X_i for any $i \in A$, and therefore is almost everywhere a function of X_B for any $B \subseteq I$ with $B \cap A \neq \emptyset$. Taking the product random variable over all such A for a fixed B, we see that $Y_{\cap B}$ is a function of X_B almost everywhere, and so

$$H\left(Y_{\cap B}\right) \leq H\left(X_B\right). \tag{4.27}$$

Now, using also the independence hypothesis,

$$\sigma_{\mathcal{L}}\left(B\right) \geq H\left(X_B\right) \geq H\left(Y_{\cap B}\right) = \sigma_{\mathcal{L}}\left(B\right), \tag{4.28}$$

so $\mathcal{L}$ is a simple-perfect condensation of $\mathcal{M}$.

(B1 $\implies$ B2) The first part of this is simply the statement of Lemma 4.5. Next, as in the proof of Proposition 4.2, we will choose a linear order $\preceq$ on $\mathcal{P}^+ I$ such that whenever $B \supseteq C$, we have $B \preceq C$. In this case, we want to choose $\preceq$ so that every element of $\mathcal{F}$ precedes A. We can do this starting with the partial order $\preceq_p$ defined so that $B \preceq_p C$ if and only if either (1) $B \supseteq C$ or (2) B is incomparable to A in the inclusion order and $A \supseteq C$. It is straightforward to see that $\preceq_p$ is indeed a partial order, and any extension of $\preceq_p$ to a linear order gives an order $\preceq$ with the desired property.

Using the perfect condensation hypothesis, we have

$$\begin{aligned} H\left(Y_{\mathcal{P}^+ I}\right) &= \sum_{B\in\mathcal{P}^+ I} H\left(Y_B \mid Y_C \colon C \prec B\right) \\ &\leq \sum_{B\in\mathcal{P}^+ I} H\left(Y_B \mid Y_{\supsetneq B}\right) \\ &= \chi_{\mathcal{L}}\left(I\right) = H\left(Y_{\mathcal{P}^+ I}\right), \end{aligned} \tag{4.29}$$

and in particular corresponding elements of the sums here are equal. Looking at the terms in the sums corresponding to $B = A$, we have

$$H\left(Y_A \mid Y_C \colon C \prec A\right) = H\left(Y_A \mid Y_{\supsetneq A}\right), \tag{4.30}$$

and since

$$\left\{C \mid C \prec A\right\} \supseteq \mathcal{F} \cup \left\{D \mid D \supsetneq A\right\}, \tag{4.31}$$

it follows that

$$H\left(Y_A \mid Y_{\mathcal{F}}, Y_{\supsetneq A}\right) = H\left(Y_A \mid Y_{\supsetneq A}\right). \tag{4.32}$$

This is equivalent to the desired independence statement.

(B2 $\implies$ B1) We want to show that $\chi_{\mathcal{L}}\left(A\right) = H\left(X_A\right)$ for all $A \in \mathcal{P}^+ I$. Take any such A. First, whenever $B \in \mathcal{P}^+ I$ with $B \cap A \neq \emptyset$, the latent variable Y_B is a

function of X_A almost everywhere, so the variable $Y_{\cap A}$ is a function of X_A almost everywhere, and so we have

$$H\left(Y_{\cap A}\right) = H\left(X_A\right). \tag{4.33}$$

Next, let $\preceq$ be a linear order on the set of those $B \in \mathcal{P}^+ I$ which intersect A, such that $B \preceq C$ whenever $C \supseteq B$. Writing our usual sum, we now have

$$H\left(X_A\right) = H\left(Y_{\cap A}\right) = \sum_{B \cap A \neq \emptyset} H\left(Y_B \mid Y_C \colon C \prec B\right) \tag{4.34}$$

For each B in this sum, the set $\{C \mid C \prec B\}$ contains all sets $C \in \mathcal{P}^+ I$ which are strict supersets of B. Further, all its elements that are not strict supersets of B are inclusion-incomparable to B. So, by our Markov-condition-style hypothesis,

$$H\left(Y_B \mid Y_C \colon C \prec B\right) = H\left(Y_B \mid Y_D \colon D \supsetneq B\right) \tag{4.35}$$

for each such B. Hence,

$$H\left(X_A\right) = \sum_{B \cap A \neq \emptyset} H\left(Y_B \mid Y_D \colon D \supsetneq B\right) = \chi_{\mathcal{L}}\left(A\right) \tag{4.36}$$

as desired. □

The ordered Markov condition can be stated in another form, which lets us make another statement equivalent to (B1) and (B2).

Proposition 4.10. *Let I be a finite set, and suppose that $(Y_A)_{A \in \mathcal{P}^+ I}$ are random variables with finite entropy on some probability space. Then, the following are equivalent.*

1. For any $A \in \mathcal{P}^+ I$, if $\mathcal{F} \subseteq \mathcal{P}^+ I$ is the collection of all $B \in \mathcal{P}^+ I$ that are incomparable to A, then the random variables Y_A and $Y_{\mathcal{F}}$ are independent conditional on $Y_{\supsetneq A}$.
2. For any two upward-closed sets $\mathcal{F}, \mathcal{G} \subseteq \mathcal{P}^+ I$, the random variables $Y_{\mathcal{F}}$ and $Y_{\mathcal{G}}$ are independent conditional on $Y_{\mathcal{F} \cap \mathcal{G}}$.

Proof. (1 $\implies$ 2) Let $\mathcal{F}, \mathcal{G} \subseteq \mathcal{P}^+ I$ be two upward-closed sets. We want to show that

$$H\left(Y_{\mathcal{G}} \mid Y_{\mathcal{F}}\right) = H\left(Y_{\mathcal{G}} \mid Y_{\mathcal{F} \cap \mathcal{G}}\right). \tag{4.37}$$

Let $\preceq$ be a total order on $\mathcal{G} - \mathcal{F}$ such that whenever $A \supseteq B$, we have $A \preceq B$. We can expand

$$\begin{aligned} H\left(Y_{\mathcal{G}} \mid Y_{\mathcal{F}}\right) &= \sum_{A \in \mathcal{G} - \mathcal{F}} H\left(Y_A \mid Y_{\mathcal{F}}, \left(Y_B \colon B \in \mathcal{G} - \mathcal{F}, B \prec A\right)\right) \\ H\left(Y_{\mathcal{G}} \mid Y_{\mathcal{F} \cap \mathcal{G}}\right) &= \sum_{A \in \mathcal{G} - \mathcal{F}} H\left(Y_A \mid Y_{\mathcal{F} \cap \mathcal{G}}, \left(Y_B \colon B \in \mathcal{G} - \mathcal{F}, B \prec A\right)\right), \end{aligned} \tag{4.38}$$

so it would suffice to show that the corresponding terms are equal. Now, for each $A \in \mathcal{G} - \mathcal{F}$, let $\mathcal{I}_A \subseteq \mathcal{P}^+ I$ be the set of all such sets incomparable with A, and $\mathcal{S}_A \subseteq \mathcal{P}^+ I$ be the set of strict supersets of A; we know by hypothesis that A is conditionally independent of $Y_{\mathcal{I}_A}$ given $Y_{\mathcal{S}_A} = Y_{\supsetneq A}$. By construction,

$$\begin{aligned} \mathcal{S}_A &\subseteq \mathcal{F} \cup \{B \in \mathcal{G} - \mathcal{F} \mid B \prec A\} \subseteq \mathcal{I}_A \\ \mathcal{S}_A &\subseteq (\mathcal{F} \cap \mathcal{G}) \cup \{B \in \mathcal{G} - \mathcal{F} \mid B \prec A\} \subseteq \mathcal{I}_A, \end{aligned} \tag{4.39}$$

so

$$H(Y_A \mid Y_{\mathcal{F}}, (Y_B \colon B \prec A)) = H(Y_A \mid Y_{\supsetneq A}) = H(Y_A \mid Y_{\mathcal{F}\cap\mathcal{G}}, (Y_B \colon B \prec A)) \tag{4.40}$$

as desired.

($2 \implies 1$) Let $A \in \mathcal{P}^+ I$, let $\mathcal{F} \subseteq \mathcal{P}^+ I$ be the collection of all such sets incomparable to A, and let $\mathcal{S}$ be the collection of strict supersets of A. Statement (2) tells us that $\mathcal{F} \cup \mathcal{S}$ is conditionally independent of $\mathcal{S} \cup \{A\}$ given $\mathcal{S}$, from which the conclusion follows. □

Theorem 4.9 looks like an analogue of Lemma 4.5, strengthening the condition that $H(Y_{\cap A}) = H(X_A)$. This condition is fairly different from perfect condensation in other ways though. Recall the latent variable model $\mathcal{L}_1$ from Example 4.1, in which $Y_{\{i\}} = X_i$ and Y_A is constant for all other A. Here, the condition $H(Y_{\cap A}) = H(X_A)$ is satisfied for every set A. We were able to construct such a latent variable model for any given random variable model—we could for example construct a random variable model $\mathcal{M}$ with random variables X and an associated perfect condensation with many nontrivial latents Y, using Theorem 4.9, and then $\mathcal{M}$ would admit a very different random variable model as in Example 4.1 with latents Z, and both these latent variable models would satisfy the same condition:

$$H(Y_{\cap A}) = H(Z_{\cap A}) = H(X_A) \tag{4.41}$$

for all subsets A of the index set.

By contrast, the condition of perfect condensation is much more rigid. Given a random variable model $\mathcal{M}$ and associated latent variable models $\mathcal{L}_1$ and $\mathcal{L}_2$, we want to say that $\mathcal{L}_1$ and $\mathcal{L}_2$ are essentially the same. It would be straightforward to express this by asserting the existence of an equivalence between $\mathcal{L}_1$ and $\mathcal{L}_2$ satisfying certain properties. Unfortunately, the condition of an equivalence $\mathcal{L}_1 \simeq \mathcal{L}_2$ would be too strong. It may be that the underlying measure spaces of our two latent variable models—call them Λ_1 and Λ_2—differ in a way that does not interact with the random variables of interest. Maybe different points of Λ_1 can always be distinguished by some latent variable, but Λ_2 is the product of Λ_1 by the unit interval equipped with Lebesgue measure, for example. In order to regard such a difference as inessential, we should be willing to extend our latent variable models by arbitrary morphisms. That is, we should be satisfied with studying latent variable models $\widetilde{\mathcal{L}_1}$ and $\widetilde{\mathcal{L}_2}$, together with morphisms $\widetilde{\mathcal{L}_k} \to \mathcal{L}_k$ for each k, and an equivalence between $\widetilde{\mathcal{L}_1}$ and $\widetilde{\mathcal{L}_2}$.

Definition 4.11. Let Ω, Λ_1, and Λ_2 be probability spaces, and $\pi_k \colon \Lambda_k \to \Omega$ probability preserving maps for $k \in \{1, 2\}$. An *amalgamation* of the diagram

$$\begin{array}{ccc} & & \Lambda_1 \\ & & \downarrow{\scriptstyle \pi_1} \\ \Lambda_2 & \xrightarrow{\pi_2} & \Omega \end{array} \tag{4.42}$$

is a countable discrete probability space Λ_0 together with probability-preserving maps $\rho_k \colon \Lambda_0 \to \Lambda_k$ such that the diagram

$$\begin{array}{ccc} \Lambda_0 & \xrightarrow{\rho_1} & \Lambda_1 \\ \downarrow{\scriptstyle \rho_2} & & \downarrow{\scriptstyle \pi_1} \\ \Lambda_2 & \xrightarrow{\pi_2} & \Omega \end{array} \tag{4.43}$$

of probability-preserving maps commutes.

Definition 4.12. Let $\mathcal{M} = \left(\Omega, (X_i)_{i \in I}\right)$ be a random variable model, and let $\mathcal{L}_1 = \left(\left(\Lambda_1, (Y_A)_{A \in \mathcal{P}^+ I}\right), \pi_1\right)$ and $\mathcal{L}_2 = \left(\left(\Lambda_2, (Z_A)_{A \in \mathcal{P}^+ I}\right), \pi_2\right)$ be latent variable models associated with $\mathcal{M}$. An *amalgamation* of $\mathcal{L}_1$ and $\mathcal{L}_2$ consists of

(1) a countable discrete probability space Λ_0;
(2) two latent variable models $\widetilde{\mathcal{L}_1}$ and $\widetilde{\mathcal{L}_2}$, both of which have underlying probability space Λ_0 and associated random variable model $\mathcal{M}$; and
(3) two morphisms, $\rho_k \colon \widetilde{\mathcal{L}_k} \to \mathcal{L}_k$ for $k \in \{1, 2\}$, which each act as the identity on the respective index sets of the latent variable models and on the respective ranges of all the latent variables, that is, they have the forms

$$\boldsymbol{\rho}_1 = \left(\rho_1, \mathrm{id}_{\mathcal{P}^+ I}, (\mathrm{id}_{\mathcal{R} Y_A})_{A \in \mathcal{P}^+ I}\right) \colon \widetilde{\mathcal{L}_1} \to \mathcal{L}_1 \tag{4.44}$$

$$\boldsymbol{\rho}_2 = \left(\rho_2, \mathrm{id}_{\mathcal{P}^+ I}, (\mathrm{id}_{\mathcal{R} Z_A})_{A \in \mathcal{P}^+ I}\right) \colon \widetilde{\mathcal{L}_2} \to \mathcal{L}_2. \tag{4.45}$$

Lemma 4.13. *Let $\mathcal{M} = \left(\Omega, (X_i)_{i \in I}\right)$ be a random variable model, and let $\mathcal{L}_1 = \left(\left(\Lambda_1, (Y_A)_{A \in \mathcal{P}^+ I}\right), \pi_1\right)$ and $\mathcal{L}_2 = \left(\left(\Lambda_2, (Z_A)_{A \in \mathcal{P}^+ I}\right), \pi_2\right)$ be latent variable models associated with $\mathcal{M}$. Then, there is a probability space Λ_0 with maps $\rho_k \colon \Lambda_0 \to \Lambda_k$ for $k \in \{1, 2\}$, which is an amalgamation of the diagram made by π_1 and π_2. Further, there is an amalgamation of $\mathcal{L}_1$ and $\mathcal{L}_2$, consisting of Λ_0 together with the objects*

(4.46)
$$\widetilde{\mathcal{L}_1} = \left(\left(\Lambda_0, (\rho_1^* Y_A)_{A \in \mathcal{P}^+ I}\right), \pi_1 \circ \rho_1\right) \qquad \widetilde{\mathcal{L}_2} = \left(\left(\Lambda_0, (\rho_2^* Z_A)_{A \in \mathcal{P}^+ I}\right), \pi_2 \circ \rho_2\right)$$

and

$$\boldsymbol{\rho}_1 = \left(\rho_1, \mathrm{id}_{\mathcal{P}^+ I}, (\mathrm{id}_{\mathcal{R} Y_A})_{A \in \mathcal{P}^+ I}\right) \tag{4.47}$$

$$\boldsymbol{\rho}_2 = \left(\rho_2, \mathrm{id}_{\mathcal{P}^+ I}, (\mathrm{id}_{\mathcal{R} Z_A})_{A \in \mathcal{P}^+ I}\right). \tag{4.48}$$

Proof. First, we construct Λ_0. The plan is to construct a measurable space for Λ_0 as a pullback in the category of measurable spaces, and then to construct a probability measure using a conditional-independence idea. Consider the set

$$S = \{(\lambda_1, \lambda_2) \mid \lambda_1 \in \Lambda_1, \lambda_2 \in \Lambda_2, \pi_1(\lambda_1) = \pi_2(\lambda_2)\}. \tag{4.49}$$

This is a subset of the product measurable space $\Lambda_1 \times \Lambda_2$, so we can view it as a countable discrete measurable space. We define the measurable maps $\rho_k \colon S \to \Lambda_k$ by

$$\rho_k(\lambda_1, \lambda_2) = \lambda_k, \tag{4.50}$$

and we define $\pi_0 \colon S \to \Omega$ by

$$\pi_0 = \pi_1 \circ \rho_1 = \pi_2 \circ \rho_2. \tag{4.51}$$

Now, let $\mathbf{P}_\Omega$ be the probability measure on Ω, and let $\mathbf{P}_1$ and $\mathbf{P}_2$ be those on Λ_1 and Λ_2, respectively. For any $\omega \in \Omega$ with $\mathbf{P}_\Omega(\{\omega\}) > 0$, since

$$\mathbf{P}_\Omega(\{\omega\}) = \mathbf{P}_1(\pi_1 = \omega) = \mathbf{P}_2(\pi_2 = \omega), \tag{4.52}$$

we can define the conditional probabilities $\omega \mapsto \mathbf{P}_1(\cdot \mid \pi_1 = \omega)$ and $\omega \mapsto \mathbf{P}_2(\cdot \mid \pi_1 = \omega)$, taking values in the spaces of measures $G(\Lambda_1)$ and $G(\Lambda_2)$, for almost all $\omega \in \Omega$. Hence, we can define the integral

$$\widetilde{\mathbf{P}_0}(E) = \int_\Omega \left[\mathbf{P}_1(\cdot \mid \pi_1 = \omega)\right] \times \left[\mathbf{P}_2(\cdot \mid \pi_2 = \omega)\right](E)\, \mathrm{d}\mathbf{P}_\Omega \tag{4.53}$$

for every set $E \subseteq \Lambda_1 \times \Lambda_2$. The integrand

$$\left[\mathbf{P}_1(\cdot \mid \pi_1 = \omega)\right] \times \left[\mathbf{P}_2(\cdot \mid \pi_2 = \omega)\right] \tag{4.54}$$

denotes a probability measure for almost all ω, so the function $\widetilde{\mathbf{P}_0}$ is nonnegative and satisfies $\widetilde{\mathbf{P}_0}(\Lambda_1 \times \Lambda_2) = 1$. We also have countable additivity; we can apply Tonelli's theorem since the summand is nonnegative.

$$\begin{aligned}
\widetilde{\mathbf{P}_0}\left(\bigcup_{k=1}^\infty E_k\right) &= \int_\Omega \sum_{k=1}^\infty \left[\mathbf{P}_1(\cdot \mid \pi_1 = \omega)\right] \times \left[\mathbf{P}_2(\cdot \mid \pi_2 = \omega)\right](E_k)\, \mathrm{d}\mathbf{P}_\Omega \\
&= \sum_{k=1}^\infty \int_\Omega \left[\mathbf{P}_1(\cdot \mid \pi_1 = \omega)\right] \times \left[\mathbf{P}_2(\cdot \mid \pi_2 = \omega)\right](E_k)\, \mathrm{d}\mathbf{P}_\Omega \\
&= \widetilde{\mathbf{P}_0}\left(\bigcup_{k=1}^\infty E_k\right).
\end{aligned} \tag{4.55}$$

Hence, $\widetilde{\mathbf{P}_0}$ is a probability measure on $\Lambda_1 \times \Lambda_2$. Further,

$$\begin{aligned}
\widetilde{\mathbf{P}_0}(S) &= \int_\Omega \left[\mathbf{P}_1(\cdot \mid \pi_1 = \omega) \times \mathbf{P}_2(\cdot \mid \pi_2 = \omega)\right]\left(S \cap \pi_0^{-1}(\omega)\right)\, \mathrm{d}\mathbf{P}_\Omega \\
&= \int_\Omega \mathbf{P}_1\left(\pi_1^{-1}(\omega) \mid \pi_1 = \omega\right) \mathbf{P}_2\left(\pi_2^{-1}(\omega) \mid \pi_2 = \omega\right)\, \mathrm{d}\mathbf{P}_\Omega \\
&= \int_\Omega 1\, \mathrm{d}\mathbf{P}_\Omega = 1,
\end{aligned} \tag{4.56}$$

so, letting $\mathbf{P}_0(E) = \widetilde{\mathbf{P}_0}(E \cap S)$, we see that $\mathbf{P}_0$ is a probability measure on S. We can now define Λ_0 to be the countable discrete probability space $(S, \mathbf{P}_0)$.

Next, we confirm that the maps ρ_k, interpreted as maps $\Lambda_0 \to \Lambda_k$ of probability spaces, are probability-preserving. For any $E \subseteq \Lambda_1$, and any $\omega \in \Omega$,

$$\begin{aligned}
\rho_1^{-1}(E) \cap \pi_0^{-1}(\omega) &= \{(\lambda_1, \lambda_2) \mid \lambda_1 \in E, \pi_1(\lambda_1) = \pi_2(\lambda_2) = \omega\} \\
&= \left(E \cap \pi_1^{-1}(\omega)\right) \times \pi_2^{-1}(\omega),
\end{aligned}$$

which is a rectangle, so

$$\begin{aligned}
\mathbf{P}_0\left(\rho_1^{-1}(E)\right) &= \int_\Omega \left[\mathbf{P}_1(\cdot \mid \pi_1 = \omega) \times \mathbf{P}_2(\cdot \mid \pi_2 = \omega)\right]\left(\rho_1^{-1}(E) \cap \pi_0^{-1}(\omega)\right)\, \mathrm{d}\left(\pi_1\right)_* \mathbf{P}_1 \\
&= \int_\Omega \mathbf{P}_1\left(E \cap \pi_1^{-1}(\omega) \mid \pi_1 = \omega\right) \cdot 1\, \mathrm{d}\left(\pi_1\right)_* \mathbf{P}_1 \\
&= \mathbf{P}_1(E),
\end{aligned} \tag{4.57}$$

and so ρ_1 is probability-preserving. The same reasoning tells us that ρ_2 is probability-preserving as well. It is immediate from definitions that the diagram (4.43) commutes.

Having defined Λ_0 and the maps ρ_k, we have also defined $\widetilde{\mathcal{L}_k}$ for $k \in \{1,2\}$. It is immediate that they are indeed latent variable models for $\mathcal{M}$, it is immediate that $\boldsymbol{\rho}_1$ and $\boldsymbol{\rho}_2$ are morphisms, and the other conditions in the definition of an amalgamation in the sense of latent variable models are also immediate. □

Now we will be able to state a theorem on the relation between different perfect condensations of the same random variable model. In order to prove that theorem, though, we first introduce a lemma.

Lemma 4.14. *Let X, Y_1, Y_2, and C be random variables on some countable discrete probability space $(\Omega, \mathbf{P})$, and suppose that C has discrete range. Suppose further that Y_1 and Y_2 are conditionally independent given C, that X is almost everywhere a function of (C, Y_1), and that X is also almost everywhere a function of (C, Y_2). Then, X is almost everywhere a function of C.*

Proof. Fix any $c \in C$ such that $\mathbf{P}(C = c)$ is positive; working on the measurable subspace

$$\Omega_c = \{\omega \in \Omega \mid C = c\} \tag{4.58}$$

equipped with the probability measure $\mathbf{P}_c = \mathbf{P}(\cdot \mid C = c)$, we will see that X is almost everywhere constant.

Let $A \subseteq \Omega_c$ be the set of points with positive mass. By hypothesis, there are functions $f_i \colon \mathcal{R}Y_i \to \mathcal{R}X$ for $i \in \{1,2\}$ such that

$$X = f_1(Y_1) = f_2(Y_2) \tag{4.59}$$

almost everywhere, and therefore in particular everywhere on A. For any two points $\omega_0, \omega \in A$, if there is some $\upsilon \in A$ with

$$Y_1(\upsilon) = Y_1(\omega_0) \qquad Y_2(\upsilon) = Y_2(\omega), \tag{4.60}$$

then

$$\begin{aligned} X(\omega) &= f_2(Y_2(\omega)) = f_2(Y_2(\upsilon)) \\ &= f_1(Y_1(\upsilon)) = f_1(Y_1(\omega_0)) \\ &= X(\omega_0). \end{aligned} \tag{4.61}$$

We know that Y_1 and Y_2 are conditionally independent given C, so they are in particular independent on Ω_c. Thus,

$$\begin{aligned} \mathbf{P}_c(Y_1 = Y_1(\omega_0) \wedge Y_2 = Y_2(\omega)) &= \mathbf{P}_c(Y_1 = Y_1(\omega_0)) \cdot \mathbf{P}_c Y_2 = Y_2(\omega) \\ &\geq \mathbf{P}_c(\omega_0) \cdot \mathbf{P}_c(\omega) > 0, \end{aligned}$$

so we can indeed always pick such a point v. Hence, X is constant on A, and so almost everywhere on Ω_c.

Since this holds for almost every $c \in \mathcal{R}C$, we now know that there is a function $g \colon \mathcal{R}C \to \mathcal{R}X$ such that $X = g(C)$ almost everywhere. Since $\mathcal{R}C$ is discrete, this function is automatically measurable, as desired. □

Theorem 4.15 (Comparison of perfect condensations)**.** *Let $\mathcal{M}$ be a random variable model with random variables $(X_i)_{i \in I}$, and suppose that $\mathcal{L}_1$ and $\mathcal{L}_2$ are both perfect condensations of $\mathcal{M}$. Then, we can put the latent variables of $\mathcal{L}_1$ and $\mathcal{L}_2$ into correspondence in the following sense. Let $\widetilde{\mathcal{L}_1} = \left(\left(\Lambda_0, (Y_A)_{A \in \mathcal{P}^+ I}\right), \widetilde{\pi}_1\right)$ and $\widetilde{\mathcal{L}_2} = \left(\left(\Lambda_0, (Z_A)_{A \in \mathcal{P}^+ I}\right), \widetilde{\pi}_2\right)$ be the latent variable models in an amalgamation of*

$\mathcal{L}_1$ and $\mathcal{L}_2$. Then, the random variable Y_A is a function of $Z_{\supseteq A}$ almost everywhere, and reciprocally Z_A is a function of $Y_{\supseteq A}$ almost everywhere.

Proof. We know that such an amalgamation exists by Lemma 4.13. We want to deduce that Y_A is a function of $Z_{\supseteq A}$ almost everywhere; using the symmetry of the situation to interchange Y and Z, the result would then follow.

Consider any $i \in A$. By Theorem 4.9, Y_A is a function of X_i almost everywhere, and by the definition of latent variable model, X_i is a function of $Z_{\ni i}$ almost everywhere, so Y_A is a function of $Z_{\ni i}$ almost everywhere.

From here, we will apply Lemma 4.14 repeatedly, using induction. Consider any two upward-closed sets $\mathcal{F}, \mathcal{G} \subseteq \mathcal{P}^+ I$. That lemma tells us that if Y_A is a function of $Z_{\mathcal{F}}$ almost everywhere and is a function of $Z_{\mathcal{G}}$ almost everywhere, and if $Z_{\mathcal{F}}$ is conditionally independent of $Z_{\mathcal{G}}$ given $Z_{\mathcal{F}\cap\mathcal{G}}$, then Y_A is a function of $Z_{\mathcal{F}\cap\mathcal{G}}$ almost everywhere. The conditional independence condition follows from the hypothesis that $\mathcal{L}_2$ is a perfect condensation, using Proposition 4.10. Since

$$\bigcap_{i\in A} \mathcal{F}_i = \{B \colon B \supseteq A\}\,, \tag{4.62}$$

we can conclude that Y_A is a function of $Z_{\supseteq A}$ almost everywhere, as desired. □

5. Comparison of latent variable models

5.1. **Suggestive examples.** We can generalize the ideas of the comparison theorem, Theorem 4.15, beyond the hypothesis of perfect condensation. As we have seen in results like Theorem 4.9, perfect condensation is a significant constraint on the structure of a latent variable model. However, we will see in Theorem 5.8 that an analogue of Theorem 4.15 in a more general setting does exist, if we are willing to exchange a few of the objects in its statement for appropriate approximations. To begin to suggest an idea, consider the following examples.

Example 5.1. Let L be a random variable with range $[0,1]$, and let $(X_i)_{i=1}^n$ be conditionally independent coins with *bias* L. That is, the X_i are **2**-valued random variables, which are conditionally independent given L, and which, conditional on L, take the value 1 with probability L and 0 with probability $1-L$. This determines a random variable model with random variables $(X_i)_{i=1}^n$. We would like to consider an associated latent variable model with latent variables

$$\begin{aligned} \widetilde{Y}_{\{1,\dots,n\}} &= L \\ \widetilde{Y}_i &= X_i \qquad (i = 1 \text{ to } n), \end{aligned} \tag{5.1}$$

but L does not have a countable range, so we can instead consider

$$\begin{aligned} Y_{\{1,\dots,n\}} &= b\,(L) \\ Y_i &= X_i \end{aligned}$$

for some *bucketing function* b. That is, we pick a finite set of disjoint intervals with union $[0,1]$, and define b to assign to each number the unique interval containing it.

Without yet posing a definite sense, one might suspect that the latent variable model constructed here is approximately the only "reasonable" latent variable model associated with the given random variable model, up to some notion of approximation. Indeed, we have constructed a number of different latent variable models,

depending on our choice of bucketing function b, which we can regard as approximating each other, as long as the intervals are sufficiently small. A precise form of the ideas that such models are approximately unique will be realized in a generalization of Theorem 4.15. We will also mention some more diverse examples before continuing.

Example 5.2. Suppose that we have some coins with different unknown but independent biases. We cannot observe flips of the coins directly. Instead, two coins are chosen at a time—we know which two—and we are told the number of heads, which may be zero, one, or two.

Formally, let $(L_j)_{j\in J}$ be a family of independent random variables with range $[0,1]$; let $c_1, c_2 \colon I \to J$ be (deterministic) functions; for $i \in I$, let C_1^i, C_2^i be Bernoulli random variables $C_k^i \sim \mathrm{Bern}\left(L_{c_k(i)}\right)$, conditionally independent given L; and define $X_i = C_1^i + C_2^i$.

We can construct an associated latent variable model with latent variables $(Y_A)_{A\in\mathcal{P}^+I}$ as follows. For each A with $|A| > 1$, let $S \subseteq J$ be the set of all $j \in J$ such that $c_k(i) = j$ for some i in A and $k \in \{1,2\}$—informally, this is the set of $j \in J$ that contribute to X_A. Then, let

$$Y_A = \left(b\left(L_j\right) : j \in S\right), \tag{5.2}$$

for some bucketing function b, and let

$$Y_{\{i\}} = X_i. \tag{5.3}$$

We will see that this is in some sense approximately the only reasonable latent variable model when the buckets are sufficiently small and the sets of observations X_i to which each coin L_j contributes are sufficiently large and sufficiently different as j varies.

Example 5.3. Here we use the language of structural causal models [SGS01; PJS17]. Given a structural causal model, we can produce a corresponding latent variable model in our sense as follows. Let G be a causal graph with vertices $\{X_j\}_{j\in J}$, and with a subset of those vertices, corresponding to indices $I \subseteq J$, designated as observed. We can view the joint distribution as a random variable model $\mathcal{M}_J = \left((\Omega, \mathbf{P}), (X_j)_{j\in J}\right)$ such that G and $\mathbf{P}$ satisfy the causal Markov condition, and we can also consider the random variable model $\mathcal{M}_I = \left((\Omega, \mathbf{P}), (X_i)_{i\in I}\right)$ on only the observed variables. The problem of latent causal discovery is concerned with recovering information about $\mathcal{M}_J$ and G from $\mathcal{M}_I$, generally under reasonable further hypotheses, or with similar questions involving more general sorts of graphical structures. In our language, we can represent $\mathcal{M}_J$ by a latent variable model $\mathcal{L}$ with latent variables $(Y_A)_{A\in\mathcal{P}^+I}$ as follows. For all $j \in J$ and $i \in I$, we denote by $j \blacktriangleleft i$ the relation that there is a directed path from X_j to X_i in the graph G. Then, let

$$Y_A = \left(X_j \colon \exists i \in A.\, j \blacktriangleleft i\right). \tag{5.4}$$

It is common in the theory of latent causal discovery to have failures of identifiability, wherein the desired information about such a pair $(\mathcal{M}_J, G)$ cannot be recovered from $\mathcal{M}_I$. Thus, we cannot expect an analogue of Theorem 4.15 to apply to a latent variable model like $\mathcal{L}$ without further assumptions. But when such a theorem does apply, we can hope to use that fact to derive an identifiability result for structural causal models.

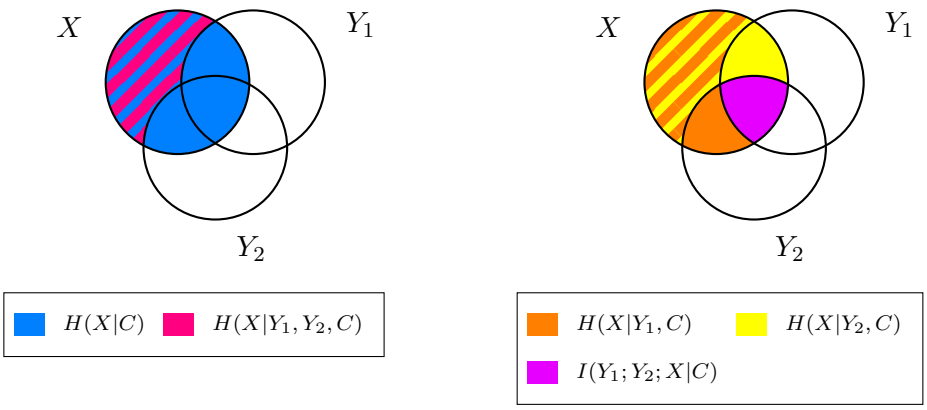

FIGURE 5.1. Information diagrams for Lemma 5.4

5.2. **Comparison of latent variable models.** We now proceed toward an analogue of Theorem 4.15, starting with a quantitative variant of Lemma 4.14.

Lemma 5.4. *Let X, Y_1, Y_2, and C be random variables on some probability space, each of which has finite entropy. Then,*

$$H(X \mid C) \leq H(X \mid Y_1, C) + H(X \mid Y_2, C) + I(Y_1; Y_2 \mid C), \tag{5.5}$$

and further, we can make the exact statement

$$\begin{aligned} H(X \mid C) = {} & H(X \mid Y_1, C) + H(X \mid Y_2, C) - H(X \mid Y_1, Y_2, C) \\ & + I(Y_1; Y_2; X \mid C). \end{aligned} \tag{5.6}$$

Proof. We can verify (5.6) with a straightforward if unenlightening calculation. This can be clarified to a certain extent pictorially, as in Figure 5.1.

$$\begin{aligned} & H(X \mid Y_1, C) + H(X \mid Y_2, C) - H(X \mid Y_1, Y_2, C) + I(Y_1; Y_2; X \mid C) \\ & \quad = H(X \mid Y_1, C) + I(X; Y_1 \mid Y_2, C) + I(X; Y_1; Y_2 \mid C) \\ & \quad = H(X \mid Y_1, C) + I(X; Y_1 \mid C) \\ & \quad = H(X \mid C). \end{aligned} \tag{5.7}$$

To deduce the inequality form, we use the nonnegativity of entropy and mutual information.

$$\begin{aligned} H(X \mid C) &= H(X \mid Y_1, C) + H(X \mid Y_2, C) - H(X \mid Y_1, Y_2, C) \\ &\qquad + I(Y_1; Y_2; X \mid C) \\ &\leq H(X \mid Y_1, C) + H(X \mid Y_2, C) \\ &\qquad + I(Y_1; Y_2 \mid C) - I(Y_1; Y_2 \mid X, C) \\ &\leq H(X \mid Y_1, C) + H(X \mid Y_2, C) + I(Y_1; Y_2 \mid C). \end{aligned} \tag{5.8}$$

□

Definition 5.5. The *polar* of a subset $\mathcal{F}$ of some nonempty power set $\mathcal{P}^+ I$ is the collection

$$\mathcal{F}^\circ = \left\{ B \in \mathcal{P}^+ I : \forall A \in \mathcal{F}.\, A \cap B \neq \emptyset \right\}. \tag{5.9}$$

In order to state the general comparison theorem, we will need to inductively take intersections of certain sets. We can organize this induction with the concept of an intersection tree as follows.

Definition 5.6. An *intersection tree* T on an intersection-closed collection of sets M is a triple (V, E, ℓ) of *vertices*, *edges*, and *labels*, satisfying the following conditions.

(1) (V, E) is a *directed binary tree*; every vertex either has no parents (and so is a *leaf*) or exactly two parents (and so is *internal*), and there is one vertex, the *root*, to which every vertex has a unique directed path.
(2) ℓ is a function from V to M.
(3) For each internal vertex u, with parents v and w, we have

$$\ell(u) = \ell(v) \cap \ell(w). \tag{5.10}$$

For each internal vertex v of T, suppose that a_v and b_v are the labels of its parents. Then, we call the family

$$v \mapsto (\{a_v, b_v\}, \ell(v)), \tag{5.11}$$

ranging over internal vertices v of T, the *family of intersections* of T, and we correspondingly call each element $(\{a_v, b_v\}, \ell(v))$ an *intersection* of T. Be warned that the same intersection may appear more than once in the family of intersections, assigned to different internal vertices.

Proposition 5.7. *Let (V, E) be a directed binary tree, let M be an intersection-closed collection of sets, and let $\tilde{\ell}$ be a function from the set of leaves of V to M. Then, there is a unique extension of $\tilde{\ell}$ to a function $\ell\colon V \to M$ such that (V, E, ℓ) is an intersection tree. That function ℓ assigns to each internal vertex the meet of all the labels of the leaves which are its ancestors.*

Proof. Induction.

□

Now, we are ready for the general comparison theorem. This will involve introducing various sets of indices, and then using them to state an inequality, (5.13). We can compare this inequality to Theorem 4.15 to better understand what it is claiming. To analogize an inequality to an exact statement, we can think of it as saying that if each of the terms on the right-hand side is small, then the left-hand side is small as well. Starting on the right-hand side, we will have two kinds of terms. The terms of the form $H(Y_{\supseteq A} \mid X_B)$ being small is an approximate form of $Y_{\cap B}$ being a function of X_B almost everywhere, assuming that $A \cap B \neq \emptyset$, and that would be a consequence of perfect condensation. So, the analogy is strongest when $A \cap B \neq \emptyset$ for every $B \in \mathcal{F}$, that is, when $A \in \mathcal{G}$. Next, the term $I\left(Z_{\mathcal{L}(v)}; Z_{\mathcal{R}(v)} \mid Z_{\mathcal{I}(v)}\right)$ being small is an approximate form of the statement that $Z_{\mathcal{L}(v)}$ and $Z_{\mathcal{R}(v)}$ are independent given $Z_{\mathcal{I}(v)}$, which follows from Proposition 4.10. On the left-hand side, we conclude that $H(Y_{\supseteq A} \mid Z_{\mathcal{G}})$ is small—here we have replace $Z_{\supseteq A}$ with $Z_{\mathcal{G}}$ relative to Theorem 4.15. Again, consider the case $A \in \mathcal{G}$. Then, every set containing A is also in $\mathcal{G}$, so when we say that $H(Y_{\supseteq A} \mid Z_{\mathcal{G}})$ is small, we are saying that the information in $Y_{\supseteq A}$ is not necessarily in $Z_{\supseteq A}$, but it is mostly in the larger $Z_{\mathcal{G}}$. We can think of $\mathcal{G}$ as a penumbra around $\{C\colon C \supseteq A\}$. For example, if $\mathcal{F}$ consists of all n-element subsets of A, then we can see that $\mathcal{G}$ consists of all sets that contain at least all but $n-1$ elements of A. We can make $\mathcal{G}$ better approximate $\{C\colon C \supseteq A\}$

by picking a larger $\mathcal{F}$, though this comes at a cost in the form of extra terms on the right-hand side.

We also have the exact statement (5.14). This is in some sense stronger, but it has a less clear interpretation, without the analogy to Theorem 4.15. The merit of (5.13) is that it controls a quantity relating Y and Z using terms that each depend only on one of the two specified latent variable models. Thus, this theorem establishes that if each of those two latent variable models has a certain property, then a certain relation between them follows. In contrast, (5.14) has both Y and Z in each of its terms, so it merely reasons from some properties relating Y and Z to other such.

Theorem 5.8 (Comparison of latent variable models)**.** *Let $(X_i)_{i \in I}$ be the random variables of some random variable model, and let $(Y_A)_{A \in \mathcal{P}^+ I}$ and $(Z_A)_{A \in \mathcal{P}^+ I}$ be the latent variable of two associated latent variable models. Form an amalgamation of those latent variable models with some underlying probability space Λ_0; in the sequel, when we write random variables X, Y, or Z, we will mean their pullbacks to Λ_0 under the appropriate maps. Next, consider any set $A \in \mathcal{P}^+ I$ and any collection $\mathcal{F} \subseteq \mathcal{P}^+ I$; let $\mathcal{G} = \mathcal{F}^\circ$ be the polar*

$$\mathcal{G} = \left\{ C \in \mathcal{P}^+ I \colon \forall B \in \mathcal{F}.\, B \cap C \neq \emptyset \right\}; \tag{5.12}$$

let $T = (V, E, \mathcal{I})$ be an intersection tree on the lattice of upward-closed subsets of $\mathcal{P}^+ I$ such that $\mathcal{I}$ restricts to a bijection between the leaves of T and the set of sets $\{C \in \mathcal{P}^+ I \colon B \cap C \neq \emptyset\}$ ranging over $B \in \mathcal{F}$; and write the set of leaves of T as L, its set of internal vertices as N, and its family of intersections as $(\{\mathcal{L}(v), \mathcal{R}(v)\}, \mathcal{I}(v))_{v \in N}$. Then, we have

$$H\left(Y_{\supseteq A} \mid Z_{\mathcal{G}}\right) \leq \left[\sum_{B \in \mathcal{F}} H\left(Y_{\supseteq A} \mid X_B\right) \right] + \left[\sum_{v \in N} I\left(Z_{\mathcal{L}(v)}; Z_{\mathcal{R}(v)} \mid Z_{\mathcal{I}(v)}\right) \right]. \tag{5.13}$$

Further, we can make the exact statement

$$\begin{aligned} H\left(Y_{\supseteq A} \mid Z_{\mathcal{G}}\right) = {} & \left[\sum_{v \in L} H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(v)}\right) \right] - \left[\sum_{v \in N} H\left(Y_{\supseteq A} \mid Z_{\mathcal{L}(v) \cup \mathcal{R}(v)}\right) \right] \\ & + \left[\sum_{v \in N} I\left(Z_{\mathcal{L}(v)}; Z_{\mathcal{R}(v)}; Y_{\supseteq A} \mid Z_{\mathcal{I}(v)}\right) \right]. \end{aligned} \tag{5.14}$$

Proof. We will first prove (5.14), which we can do following the inductive idea of Theorem 4.15, but now repeatedly applying Lemma 5.4. For any vertex v of T, let T_v be the subgraph of T which contains those vertices which are ancestors of v. Then, T_v is itself an intersection tree, which has root v. Write the set of leaves of T_v as $L(v)$ and its set of internal vertices as $N(v)$.

For each $v \in T$, we will establish an analogue of (5.14) for T_v, which will be

$$\begin{aligned} H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(v)}\right) = {} & \left[\sum_{w \in L(v)} H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(w)}\right) \right] - \left[\sum_{w \in N(v)} H\left(Y_{\supseteq A} \mid Z_{\mathcal{L}(w) \cup \mathcal{R}(w)}\right) \right] \\ & + \left[\sum_{w \in N(v)} I\left(Z_{\mathcal{L}(w)}; Z_{\mathcal{R}(w)}; Y_{\supseteq A} \mid Z_{\mathcal{I}(w)}\right) \right]. \end{aligned} \tag{5.15}$$

If v is a leaf, then both sides of this equation are equal to $H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(v)}\right)$. If v is internal, we can use Lemma 5.4—let s and t be the parents of v.

(5.16)

$$\begin{aligned}
H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(v)}\right) &= H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(s)}\right) + H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(t)}\right) - H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(s)\cup\mathcal{I}(t)}\right) \\
&\qquad + I\left(Z_{\mathcal{I}(s)}; Z_{\mathcal{I}(t)}; Y_{\supseteq A} \mid Z_{\mathcal{I}(v)}\right) \\
&= \left[\sum_{w\in L(v)} H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(w)}\right)\right] - \left[\sum_{w\in N(s)\cup N(t)} H\left(Y_{\supseteq A} \mid Z_{\mathcal{L}(w)\cup\mathcal{R}(w)}\right)\right] \\
&\qquad + \left[\sum_{w\in N(s)\cup N(t)} I\left(Z_{\mathcal{L}(w)}; Z_{\mathcal{R}(w)}; Y_{\supseteq A} \mid Z_{\mathcal{I}(w)}\right)\right] \\
&\qquad - H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(s)\cup\mathcal{I}(t)}\right) + I\left(Z_{\mathcal{I}(s)}; Z_{\mathcal{I}(t)}; Y_{\supseteq A} \mid Z_{\mathcal{I}(v)}\right) \\
&= \left[\sum_{w\in L(v)} H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(w)}\right)\right] - \left[\sum_{w\in N(v)} H\left(Y_{\supseteq A} \mid Z_{\mathcal{L}(w)\cup\mathcal{R}(w)}\right)\right] \\
&\qquad + \left[\sum_{w\in N(v)} I\left(Z_{\mathcal{L}(w)}; Z_{\mathcal{R}(w)}; Y_{\supseteq A} \mid Z_{\mathcal{I}(w)}\right)\right].
\end{aligned}$$

Specializing this equation to the root, we establish (5.14).

Equation (5.14) follows by a term-by-term comparison. We have

$$\sum_{v\in L} H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(v)}\right) = \sum_{B\in\mathcal{F}} H\left(Y_{\supseteq A} \mid Z_{\cap B}\right), \tag{5.17}$$

and for all $B \in \mathcal{F}$,

$$H\left(Y_{\supseteq A} \mid Z_{\cap B}\right) \le H\left(Y_{\supseteq A} \mid X_B\right) \tag{5.18}$$

since X_B is a function of $Z_{\cap B}$ almost everywhere. For all $v \in N$,

$$-H\left(Y_{\supseteq A} \mid Z_{\mathcal{L}_v\cup\mathcal{R}_v}\right) \le 0 \tag{5.19}$$

and

$$\begin{aligned}
&I\left(Z_{\mathcal{L}(v)}; Z_{\mathcal{R}(v)}; Y_{\supseteq A} \mid Z_{\mathcal{I}(v)}\right) \\
&\qquad = I\left(Z_{\mathcal{L}(v)}; Z_{\mathcal{R}(v)} \mid Z_{\mathcal{I}(v)}\right) - I\left(Z_{\mathcal{L}(v)}; Z_{\mathcal{R}(v)} \mid Y_{\supseteq A}, Z_{\mathcal{I}(v)}\right) \\
&\qquad \le I\left(Z_{\mathcal{L}(v)}; Z_{\mathcal{R}(v)} \mid Z_{\mathcal{I}(v)}\right).
\end{aligned} \tag{5.20}$$

□

5.3. **Variation.** To better understand what the comparison theorem is saying, and to develop its consequences, we will look at a few variants of it.

First, we can consider the following simplified forms of Theorem 5.8.

Corollary 5.9. *Let $(X_i)_{i\in I}$ be the random variables of a random variable model, let $(Y_A)_{A\in\mathcal{P}^+I}$ and $(Z_A)_{A\in\mathcal{P}^+I}$ be latent variables of associated latent variable models, and form an amalgamation of the latent variable models. Take $A \in \mathcal{P}^+I$ and $\mathcal{F} \subseteq \mathcal{P}^+I$, and suppose that every set $B \in \mathcal{F}$ is a subset of A.*

Then, if we let $\mathcal{G}$ be the polar of $\mathcal{F}$ and we pick an intersection tree with leaves L labeled by $\mathcal{F}$ and intersections $\left(\{\mathcal{L}(v), \mathcal{R}(v)\}, \mathcal{I}(v)\right)_{v \in N}$, we have

$$H\left(Y_{\supseteq A} \mid Z_{\mathcal{G}}\right) \leq \left[\sum_{B \in \mathcal{F}} \varrho_Y(B)\right] + \left[\sum_{v \in N} I\left(Z_{\mathcal{L}(v)}; Z_{\mathcal{R}(v)} \mid Z_{\mathcal{I}(v)}\right)\right]. \tag{5.21}$$

In particular, if Z satisfies the ordered Markov condition, then we have simply

$$H\left(Y_{\supseteq A} \mid Z_{\mathcal{G}}\right) \leq \sum_{B \in \mathcal{F}} \varrho_Y(B). \tag{5.22}$$

Proof. The first form follows from 5.8 since, whenever $B \subseteq A$,

$$H\left(Y_{\supseteq A} \mid X_B\right) \leq H\left(Y_{\supseteq B} \mid X_B\right) = \varrho_Y(B). \tag{5.23}$$

The second form follows from the definition of the ordered Markov condition. □

We can also consider an example of how the polar of a family serves to approximate the upward cone of a set.

Corollary 5.10. *Let $(X_i)_{i \in I}$ be the random variables of a random variable model, let $(Y_A)_{A \in \mathcal{P}^+ I}$ and $(Z_A)_{A \in \mathcal{P}^+ I}$ be latent variables of associated latent variable models, and form an amalgamation of the latent variable models. Suppose that $\varrho_Y(C) \leq \alpha$ for all $C \in \mathcal{P}^+ I$ with cardinality k. Now, let A be an element of $\mathcal{P}^+ I$ and $k \in \mathbf{N}$, and define $\mathcal{F}$ to be the collection of all those $C \subseteq A$ with cardinality k. Then, if we let $\mathcal{G}$ be the polar of $\mathcal{F}$ and we pick an intersection tree with leaves L labeled by $\mathcal{F}$ and intersections $\left(\{\mathcal{L}(v), \mathcal{R}(v)\}, \mathcal{I}(v)\right)_{v \in N}$, we have*

$$\mathcal{G} = \left\{C \subseteq I \colon C \text{ contains at least all but } n-1 \text{ elements of } A\right\}, \tag{5.24}$$

and

$$H\left(Y_{\supseteq A} \mid Z_{\mathcal{G}}\right) \leq \binom{|A|}{k} \cdot \alpha + \left[\sum_{v \in N} I\left(Z_{\mathcal{L}(v)}; Z_{\mathcal{R}(v)} \mid Z_{\mathcal{I}(v)}\right)\right]. \tag{5.25}$$

In general, the structure of a latent variable model may lead us to apply Theorem 5.8 to whichever pairs $(A, \mathcal{F})$ we find appropriate, but the form chosen in Corollary 5.10 gives us a simple characterization of the structure of the polar.

References

[Bis06] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[Cun+23] Hoagy Cunningham et al. *Sparse Autoencoders Find Highly Interpretable Features in Language Models*. Oct. 4, 2023. arXiv: 2309.08600[cs]. URL: http://arxiv.org/abs/2309.08600.

[Den91] Daniel C. Dennett. "Real Patterns". In: *Journal of Philosophy* 88.1 (1991). Publisher: Journal of Philosophy Inc, pp. 27–51.

[Fin29] Bruno de Finetti. "Funzione caratteristica di un fenomeno aleatorio". In: *Atti del Congresso Internazionale dei Matematici: Bologna del 3 al 10 de settembre di 1928*. 1929, pp. 179–190.

[Fri20] Tobias Fritz. "A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics". In: *Advances in Mathematics* 370 (Aug. 26, 2020), p. 107239.

[Gar+24] Scott Garrabrant et al. *Factored space models: Towards causality between levels of abstraction*. Dec. 20, 2024. arXiv: 2412.02579[cs].

[Gir82] Michèle Giry. "A categorical approach to probability theory". In: *Categorical Aspects of Topology and Analysis*. 1982, pp. 68–85.

[Hal59] Paul Richard Halmos. "Entropy in ergodic theory". In: University of Chicago, 1959.

[Mac02] David J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. USA: Cambridge University Press, 2002.

[PJS17] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Accepted: 2019-01-20 23:42:51. The MIT Press, 2017.

[Qui60] Willard Van Orman Quine. *Word and Object*. The MIT Press, 1960.

[SGS01] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. The MIT Press, Jan. 29, 2001.

[WL24] John Wentworth and David Lorell. "Natural Latents: The Concepts". In: (Mar. 20, 2024). URL: https://www.alignmentforum.org/posts/mMEbfooQzMwJERAJJ/natural-latents-the-concepts.