CONDENSATION: A THEORY OF CONCEPTS

SAM EISENSTAT

Abstract. We understand agents as creating concepts in order to organize their understanding of the world. Often agents share concepts, and so are able to work out how to make sense of the world together, forming language communities. Here, we aim to model these phenomena, looking at how probability distributions can be organized by introducing appropriate latent variables. 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 understanding goes as follows. We possess and create bodies of meaning. We may regard the traditional logical elements—propositions, entities, relations—as among the conceptual constituents of an agent's understanding. 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 model, 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 to describe the conceptual *contribution* of a latent variable to a body of meaning, rather thinking of it as a "black box", which we only know improves our accuracy on a particular set of predictions. A better understanding of the contribution of a latent variable can give an agent more freedom to manipulate a body of meaning, for example by constructing a new probabilistic model. 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.

Of these motivations—understanding intersubjectivity, conceptual contributions, and value—the mathematical ideas here will most directly aim at modeling intersubjectivity. In particular, we will 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. In §3, we will see more about what kind of probabilistic models and what kind of correspondence we will investigate here.

We can relate 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 Bayesian networks or 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 [Fin37; Kal05] on exchangeability and Wentworth and Lorell [WL23; WL24] on natural latents. The principal difference in approach is that we seek here to work with many latent variables rather than one. These should together form a kind of system of meaning—a model of the world structured in terms of a set of diverse conceptual parts. Our use of information-theoretic quantities to characterize the latent variables that we are interested in has a lot in common with Wentworth and Lorell's approach. By contrast, de Finetti's notion of exchangeability is defined in terms of symmetries of the set of variables rather than information.

In §2, we establish notational conventions and state some basic definitions and facts about probability and information. In §3, we discuss the meaning, context, and motivation of the main results. §4 defines some of the principal concepts that we will use, including random variable models and latent variable models, and begins to investigate how they behave. §5 discusses a version of the main theorem that gives an exact correspondence between different latent variable models under a strong hypothesis, which we call *perfect condensation*. In §6, these ideas are generalized to give an analogous theorem on the correspondence between general latent variable models, which takes a quantitatively approximate form, using an inequality.

2. Background and notation

In order to prepare our later discussion, we will review some points of probability theory and information theory, and establish some definitions and conventions that will be used in the sequel.

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.

In particular, we will often use a "tuple-builder" notation in order to denote products of many random variables, as follows. If $(X_i)_{i \in I}$ is a family of random variables on a space Ω and ϕ is a predicate, then we will use the expression $(X_i : i \in I, \phi(i))$, or more succinctly $(X_i : \phi(i))$, to denote the *product random variable* of those variables X_i such that $\phi(i)$. The range of this random variable is the product space of all the ranges of those X_i such that $\phi(i)$ holds, and accordingly its value at any $\omega \in \Omega$ is the tuple of all those values $X_i(\omega)$.

The idea of treating random variables as elements of their ranges 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 the composition $X \circ \pi$ is a random variable defined on Λ, which has the same range as X. We call this random variable the *pullback* of X by π, and we denote it symbolically as $\pi^* X$. (The word "pullback" is used in greater generality elsewhere, but this context will not be necessary for an understanding of what is done here.) 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 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 can define a probability measure on $\mathcal{R}X$, which we call the *distribution* of X. We will sometimes denote the distribution of X by $X_* \mathbf{P}$, since it is the (measure-theoretic) pushforward of $\mathbf{P}$ by X. The distribution of X can be defined explicitly by

$$(X_* \mathbf{P})(A) = \mathbf{P}\left(X^{-1}(A)\right) \tag{2.3}$$

for any measurable subset $A \subseteq \mathcal{R}X$. 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. Sixth, 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 set of all nonempty subsets of a set S. Symbolically, $\mathcal{P}^+ S = \mathcal{P} S - \{\emptyset\}$.

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, so for example we will use $H(X,Y)$ to mean the entropy of the product random variable (X,Y). That 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*,

$$I(X;Y;Z) = I(X;Y) - I(X;Y \mid Z). \tag{2.4}$$

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.

Proposition 2.4. *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 probability one. For all $x \in A$, the conditional probability distribution $\mathbf{P}(\cdot \mid X = x)$ is well-defined:

$$\mathbf{P}(B \mid X = x) = \frac{\mathbf{P}\left(B \cap X^{-1}(x)\right)}{\mathbf{P}\left(X^{-1}(x)\right)}. \tag{2.5}$$

From the definition of conditional entropy, for each such probability distribution there is some $y \in \mathcal{R}Y$ such that $\mathbf{P}(Y = y \mid X = x) = 1$, which lets us define a function $\widetilde{f}\colon A \to \mathcal{R}Y$ such that

$$Y = \widetilde{f}(X) \tag{2.6}$$

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

3. Ideas

In order to model intersubjectivity, we will consider ways of describing probability distributions in terms of latent variables. The central ideas here will *random variable models* and *latent variable models*, which will provide a basic language for discussing latent variables, and Theorems 5.15 and 6.8, which will give conditions under which there must be an appropriate sort of correspondence between different latent variable models. In this section, we will discuss these ideas in an informal way, providing context and motivation, but not yet precise statements or proofs.

3.1. Latent variable models.

A random variable model will express the idea of a joint distribution on a family of random variables. Some nonrigorous definitions here will anticipate the details to come in §4.

Definition 3.1 (Informal statement of 4.1)**.** A *random variable model* is a probability space Ω, together with a finite family of random variables $X_i\colon \Omega \to R_i$, satisfying certain analytic conditions.

We might for example think of the random variables $(X_i)_{i\in I}$ of a random variable model as the observable variables predicted by some statistical model or some agent. We want to interpret such predictions as well-described by means of some latent variables. We start with a very general notion of latent variable model, which encompasses a wide variety of statistical models.

Definition 3.2 (Informal statement of 4.2)**.** A *latent variable model* for a random variable model $\left(\Omega, (X_i)_{i\in I}\right)$ is an ordered triple consisting of a random variable model $\left(\Lambda, (Y_j)_{j\in J}\right)$, a probability-preserving map $\pi\colon \Lambda \to \Omega$, and a relation $\blacktriangleright \subseteq J \times I$ of *contribution*, satisfying the following condition. For each random variable X_i, the pullback $\pi^* X_i$ must be almost everywhere a function of the random variables Y_j such that $j \blacktriangleright i$. In other words, $\pi^* X_i$ is almost everywhere equal to $f_i\left(Y_j : j \in J, j \blacktriangleright i\right)$ for some measurable function f_i from the product of the ranges of the random variables $(Y_j)_{j\in J, j\blacktriangleright i}$ to the range of X_i. We call the variables $(Y_j)_{j\in J}$ *latent variables*, and if $j \blacktriangleright i$, we say that index j *contributes* to i.

As reflects usual statistical practice, a latent variable might represent something that an agent cannot observe, but which helps account for their predictions. We'd like latent variables models to give mathematical flesh to our discussion of intersubjectivity above. When we interpret some probabilistic models as latent variable models in this sense, we in some cases suspect that these models are somehow "correct" or "canonical"—that we want to use these latent variables rather than any others. The following examples elaborate on this.

3.2. Examples and related work.

Example 3.3. Let I be a finite set, let Y_I be a random variable valued in $[0,1]$, and let $(X_i)_{i\in I}$ be a family of coins, conditionally independent given Y_I, with bias Y_I. Then, $(X_i)_{i\in I}$ constitute the random variables of a random variable model. If the appropriate analytic conditions hold, we can construct an associated latent variable model by letting $J = I \cup \{I\}$, defining $Y_i = X_i$, and then taking random variables $(Y_j)_{j\in J}$ and defining a contribution relation $\blacktriangleright$ so that $I \blacktriangleright i$ and $i \blacktriangleright i$ for all $i \in I$, and no other contribution relations hold.

There is something appealing about this random variable Y_I, and we want to capture this. The main results of this paper will say something about this, but for this example, previously known methods are available.

De Finetti [Fin37] has shown that, in a setting like this one but with an infinite sequence of variables $(X_i)_{i\in I}$, introducing a variable equivalent to Y_I is the only way to write the joint distribution on $(X_i)_{i\in I}$ as a mixture of distributions in which the variables $(X_i)_{i\in I}$ are independent and identically distributed. (We will be a little loose with this talk of "equivalence" at first.) This uniqueness gives a sense in which the choice to understand $(X_i)_{i\in I}$ in terms of the latent variable Y_I is "objectively" natural. Of course, uniqueness alone isn't sufficient reason to believe that an agent who wants to understand $(X_i)_{i\in I}$ should think about Y_I, but it does give some argument in favor of this choice, and it shows that different agents may make the same choices and end up using the same latent variables. The theory of exchangeability has been extended from infinite to finite sets of variables, such as in Kerns and Székely [KS06], though the corresponding kind of uniqueness then becomes more subtle.

Incidentally, this example also demonstrates the role of the expanded probability space Λ and the measure-preserving map $\pi\colon \Lambda \to \Omega$ in the definition of latent variable models. The bias of a coin is not equal to any function of a finite number of coin flips. While the bias can be estimated very well by a large number of coin flips, it is not determined, so it order to represent it as a random variable, we must work on a larger probability space. (The map π then connects variables defined on Λ to those defined on Ω.)

Here, we want ideas that can apply to settings more heterogeneous and diverse than those to which de Finetti's notion of exchangeability is suited. Exchangeability is closely tied to the idea of random variables being independent and identically distributed—it is suitable for modeling situations in which there is a particular kind of repetition, as in our example of repeated coin flips. For a contrasting example, we can modify the above construction so that instead of each coin having the same bias Y_I, the bias of each coin X_i is a different function $f_i(Y_I)$ of the shared latent variable, for some $f_i\colon [0,1] \to [0,1]$. Note that the functions $(f_i)_{i\in I}$ themselves are here treated as known rather than random.

Given such an example, we cannot apply de Finetti-style notions of exchangeability, since the distribution no longer has these symmetries. The work of Wentworth and Lorell [WL23] is relevant. In particular, the variables $(X_i)_{i\in I}$ are conditionally independent given Y_I, which is one of the hypotheses of their fundamental theorem of natural latents. Suppose that Y_I and Z_I are two random variables such that (1—*mediation*) the variables $(X_i)_{i\in I}$ are conditionally independent given Y_I, and also conditionally independent given Z_I and (2—*redundancy*) informally, for each $i_0 \in I$, the variable X_i is approximately conditionally independent of Y_I given the joint variable $(X_i : i \in I - \{i_0\})$. Then, the fundamental theorem of natural latents tells us that each of Y_I and Z_I is approximately conditionally independent of $(X_i)_{i\in I}$ given the other.

If the hypotheses are satisfied, then we conclude that Y_I and Z_I are "equivalent" with regard to the information that they carry about X, giving a sense in which it positing Y_I is again a "natural" way to understand $(X_i)_{i\in I}$. We can observe that condition (2) should hold in many interesting cases—often, we can approximately recover Y_I from $(X_i : i \in I - \{i_0\})$, just as we can estimate the bias of a coin given a large number of flips.

The ideas studied here will have similar information-theoretic basis to those of Wentworth and Lorell, but they will develop methods suitable for more general configurations of latent variables, in line with our notion of contribution relations. Here, we want to see in what sense there is something "objective" about a structure of latent variables taken as a whole. Many kinds of probabilistic models would serve as examples here. Many of those can be represented by *Bayesian networks*; they satisfy the causal Markov condition for some directed acyclic graph [SGS01].

Definition 3.4. Let $(V_j)_{j\in J}$ be a finite set of random variables on some probability space, and let G be a directed acyclic graph with vertices $\{V_j : j \in J\}$. A vertex V_j is a *parent* of V_i if there is a directed edge from V_j to V_i, and V_j is a *descendant* of V_i if there is a directed path from V_i to V_j. Then, $(V_j)_{j\in J}$ and G satisfy the *causal Markov condition* if for all $j \in J$, the variable V_j is conditionally independent of its nondescendants (taken jointly), conditional on its parents. In some contexts, a subset of vertices may be designated as *observed* or *measured*, with the other vertices being called *latent* or *unmeasured*.

Example 3.5. Given a random variable model, we can posit a latent variable model constructed from a Bayesian network. Calling the vertices of the Bayesian network $(V_j)_{j\in J}$, the random variables of our random variable model would correspond to a subset $(V_i)_{i\in I}$ where $I \subseteq J$—the observed variables. Then, the whole family $(V_j)_{j\in J}$—not just the latent variables in the sense of Bayesian networks—will be the latent variables of the associated latent variable model, and the contribution relation $\blacktriangleright$ would be defined so that $j \blacktriangleright i$ iff i is a descendant of j. (The definition of the descendant relation permits trivial directed paths, so in particular we have $i \blacktriangleright i$ for all $i \in I$.)

As before, we expect that some Bayesian networks are "objectively" appropriate to a given random variable model. The theory of Bayesian networks offers certain limiting facts here. For example, any probability distribution on a finite set of variables with finite ranges satisfies the Markov condition for some trivial graphs, such as for any fully connected directed acyclic graph on that set, and for many more graphs if we allow latent variables. However, in some cases this can be overcome by making further assumptions about the Bayesian networks. Peters, Janzing, and Schölkopf [PJS17] discuss several versions of this in Chapters 4, 7, and 9, with Chapter 9 discussing the problem of discovering probabilistic models with latent variables, *latent causal discovery*. (They work in terms of *structural causal models*, which are slightly different from the Bayesian networks discussed here.)

This example is intended to contextualize the use of latent variable models: the methods developed here will give conditions under which a whole structure of latent variables, like what we see in the example of a Bayesian network, is in a certain sense a uniquely appropriate description of a random variable model.

In particular, these methods will apply to the case of Bayesian networks, using the construction of Example 3.5 to translate Bayesian networks into latent variable models. Then, we will be able to put some constraints on possible Bayesian network structures.

3.3. **Correspondence theorems.** Our main theorem, Theorem 6.8, will state that two latent variable models $\mathcal{L}_1$ and $\mathcal{L}_2$ associated with the same random variable model $\mathcal{M}$ stand in a kind of correspondence—to be explained—assuming that they satisfy certain conditions. The statements and motivation of the components of this theorem will take some further work here to make clear.

The key idea that will make this possible is that different latent variables in $\mathcal{L}_1$ will be distinguished because they contribute to different variables in $\mathcal{M}$. This will let us line them up with those of $\mathcal{L}_2$. In more detail, suppose that $\mathcal{M} = \left(\Omega, (X_i)_{i\in I}\right)$, that $\mathcal{L}_1 = \left(\left(\Lambda_1, (Y_j)_{j\in J}\right), \pi_1, \blacktriangleright\right)$, and that $\mathcal{L}_2 = \left(\left(\Lambda_2, (Z_k)_{k\in K}\right), \pi_2, \triangleright\right)$. This suggests that we'd like to conclude that some pairs of corresponding variables Y_j and Z_k reciprocally determine each other—that each is almost everywhere a deterministic function of the other. Sometimes we will be able to conclude this, but in general we will say something weaker. One of the problems with this statement will be suggested if we consider the possibility that two latent variables indices j and j' contribute to exactly the same indices of $\mathcal{M}$: we won't be able to distinguish them.

Instead, we will compare $\mathcal{L}_1$ and $\mathcal{L}_2$ using the index set I from $\mathcal{M}$. If A is a subset of the index set I, we can define

$$Y_A = (Y_j : \forall i \in I.\, i \in A \leftrightarrow j \blacktriangleright i)$$
$$Y_{\supseteq A} = (Y_j : \forall i \in I.\, i \in A \rightarrow j \blacktriangleright i)$$

(and we can make the analogous definitions $Z_A, Z_{\supseteq A}$). This uses "tuple-builder" notation to express these random variables. To be more explicit, the objects being defined—Y_A and $Y_{\supseteq A}$ on the left-hand sides—are random variables on Λ_1, and the ranges $\mathcal{R}Y_A$ and $\mathcal{R}Y_{\supseteq A}$ are each a product space of the ranges $\mathcal{R}Y_j$ of the random variables Y_j, for j as indicated. In the case of Y_A, this is the set of all those j that contribute to exactly these i such that $i \in A$—neither more nor fewer. We can see how this relates to our problem above—if j and j' contribute to exactly the same indices of $\mathcal{M}$, then there is some set A such that both Y_j and $Y_{j'}$ are among the variables that are combined to make Y_A.

(In the formal definition of latent variable models (Definition 4.2), the latent variables will be indexed by such subsets of I directly, instead of using a different index set J as we are doing here. This choice is just a convenience.)

Exact correspondence of latent variable models. Now, using these definitions, we can start to approach the correspondence theorem. Theorem 5.15 is a special case of the main theorem, Theorem 6.8, deriving an exact correspondence from stronger assumptions, rather than an approximate correspondence from weaker ones. Leaving out some details, it will state the following.

Theorem 3.6 (Informal statement of 5.15)**.** *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, 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.*

Two minor points will be noted here. First, the families Y and Z of random variables are defined on different probability spaces, so this statement does not quite make sense. This will be addressed using standard ideas, which we adapt to our setting using the notion of *amalgamation* (Definition 5.12). Second, we have not yet mentioned perfect condensation—this will be defined in terms of entropy at the beginning of §5.

We can regard the conclusion that Y_A is a function of $Z_{\supseteq A}$ almost everywhere and Z_A is a function of $Y_{\supseteq A}$ almost everywhere as a kind of structural correspondence between $\mathcal{L}_1$ and $\mathcal{L}_2$. This is a weaker notion than the property that Y_A is a function of Z_A almost everywhere, and vice versa, would be. For reassurance that this notion of correspondence is reasonably strong and natural, consider the following. We can observe—from the definitions of latent variable models, the contribution relation, and the variable $Z_{\supseteq A}$—that, in the following sense, whenever Z_A is "available", in fact all of $Z_{\supseteq A}$ is "available". In particular, for any $i \in I$, whenever $k \triangleright i$ for any of those $k \in K$ that was used in constructing Z_A, we in fact have $\widetilde{k} \triangleright i$ for every $\widetilde{k} \in K$ used in constructing $Z_{\supseteq A}$. So, we can say that from the perspective of X, we have no need to consider Z_A on its own, but only in the context of $Z_{\supseteq A}$.

For another related way to think about the conclusion of Theorem 5.15, notice that it is equivalent to the statement that each of $Y_{\supseteq A}$ and $Z_{\supseteq A}$ is a function of the other almost everywhere. This gives us an equivalent formulation which is more

apparently something that we might describe as an equivalence. This is expressed in Corollary 5.16.

Approximate correspondence of latent variable models. We will see that the assumptions of the exact correspondence theorem are too strong to apply to most interesting phenomena. In Theorem 6.8, we will weaken its conclusion in two ways to get an approximate correspondence theorem. First, where we previously said that Y_A is a function of $Z_{\supseteq A}$ almost everywhere, we will instead bound the conditional entropy of $Y_{\supseteq A}$ given $Z_{\supseteq A}$ (and a fortiori the conditional entropy of Y_A). We can think of this as saying that Y_A is "approximately a function of" $Z_{\supseteq A}$. Second, instead of using $Z_{\supseteq A}$ to "approximately determine" Y_A, we will use a set of latent variables from $\mathcal{L}_2$ which may be larger, which in the context of the theorem statement we will call $Z_{\mathcal{G}}$. Recall that $Z_{\supseteq A}$ is the family of those latent variables from $\mathcal{L}_2$ that contribute to all those i such that $i \in A$ (as well as, possibly, other $i \notin A$). Very roughly, we can think of $Z_{\mathcal{G}}$ as family of those latent variables of $\mathcal{L}_2$ that contribute to "most" of those i such that $i \in A$, though we will see the exact definition below. Together, we can say that instead of Y_A being exactly determined by those latent variables of $\mathcal{L}_2$ that contribute to all of A, Theorem 6.8 will state that Y_A is approximately determined by the latent variables of $\mathcal{L}_2$ that contribute to (at least) "most" of A.

The full statement of the approximate correspondence theorem is somewhat involved, and will require a few more definitions before it make sense. In particular, we need to introduce some combinatorial objects, the *polar* (6.5) and *intersection trees* (6.6), in order to state the inequality. Instead, we'll look at an informal statement, which will try to illustrate the basic structure without yet going into this detail.

Theorem 3.7 (Informal statement of 6.8)**.** *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 variables of two associated latent variable models. 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 *of $\mathcal{F}$, and let $\mathcal{L}$ and $\mathcal{R}$ be certain functions (also depending on $\mathcal{F}$) from a certain set N to the set $\mathcal{P}(\mathcal{P}^+I)$ of all subsets of $\mathcal{P}^+I$. Then,*

$$H\left(Y_{\supseteq A} \mid Z_{\mathcal{G}}\right) \leq \left[\sum_{B\in\mathcal{F}} H\left(Y_{\supseteq A} \mid X_B\right)\right] + \left[\sum_{v\in N} I\left(Z_{\mathcal{L}(v)}; Z_{\mathcal{R}(v)} \mid Z_{\mathcal{L}(v)\cap\mathcal{R}(v)}\right)\right]. \tag{3.1}$$

Let's consider various aspects of this statement. We can observe that where Theorem 5.15 states an implication—*if* two latent variable models are perfect condensations, *then* certain latent variables are functions of other latent variables almost everywhere—we instead have an inequality. This inequality generalizes an exact implication—*if* the right-hand side is zero, *then* the left-hand side is zero—to something more quantitatively forgiving. Thus, in particular, the right-hand side here should be seen as the analogue of the hypothesis of perfect condensation, and the left-hand side should be seen as the analogue of $Y_{\supseteq A}$ being a function of $Z_{\supseteq A}$ almost everywhere.

In particular, before we spoke about $Y_{\supseteq A}$ and $Z_{\supseteq A}$ each being a function of the other almost everywhere. Here, $Y_{\supseteq A}$ is "approximately a function of" $Z_{\mathcal{G}}$. To conclude that $Z_{\supseteq A}$ is approximately a function of $Y_{\mathcal{G}}$, we'd need to interchange the roles of Y and Z on the right-hand side of (3.1). In each particular case, this

may or may not give a bound of comparable magnitude; sometimes, this theorem will allow us to conclude that $Y_{\supseteq A}$ is approximately determined by $Z_{\mathcal{G}}$, but not vice versa. This exhibits a phenomenon already observed by Wentworth and Lorell [WL23]. They define a natural latent to satisfy the conjunction of two (quantitative) conditions, which they call *mediation* and *redundancy*. Two natural latents approximately determine each other, but more generally they show that a redund is approximately determined by a mediator. Thus, like here, the symmetric relation of two models being approximately equivalent arises as the conjunction of two asymmetric statements of determination.

The polar in the approximate correspondence theorem. There is more to say about the sets $\mathcal{F}$ and $\mathcal{G}$. These cannot be chosen freely, but instead $\mathcal{G}$ is defined in terms of $\mathcal{F}$, as its polar. The polar of a collection is the set of all sets (in $\mathcal{P}^+ I$) that intersect every element of that collection:

$$\mathcal{G} = \mathcal{F}^\circ = \left\{B \in \mathcal{P}^+ I : \forall A \in \mathcal{F}.\, A \cap B \neq \emptyset\right\}. \tag{3.2}$$

This represents a tradeoff—making $\mathcal{G}$ smaller comes at the cost of making $\mathcal{F}$ larger. A larger set $\mathcal{F}$ makes the right-hand expression larger, since the sum will have more terms, so it weakens the control that we have over how much $Z_{\mathcal{G}}$ determines $Y_{\supseteq A}$. However, when $\mathcal{G}$ is larger, the variable $Z_{\mathcal{G}}$ gets "farther" from $Z_{\supseteq A}$. Then, $Y_{\supseteq A}$ is not being determined by just $Z_{\supseteq A}$, but by a penumbra consisting of $Z_{\supseteq A}$ together with other Z latents.

We can say a bit more about how this works, though this should be clearest when balanced with the details in the proof of the theorem, and with examples. Since Theorem 6.8 holds for every collection $\mathcal{F}$, all these statements are true. That is, $Y_{\supseteq A}$ is approximately determined by $Z_{\mathcal{F}^\circ}$ for each set $\mathcal{F}$, to differing quantitative extents. The tradeoff is one of interestingness. We want $\mathcal{F}$ to be small enough that the bound we derive from the right-hand side is meaningful. It ought to at least improve on the trivial bound $H\left(Y_{\supseteq A} \mid Z_{\mathcal{G}}\right) \leq H\left(Y_{\supseteq A}\right)$! If $Z_{\mathcal{G}}$ is close to $Z_{\supseteq A}$, on the other hand, then the latent variables better preserve their identities between the different latent variable models, rather than having $Y_{\supseteq A}$ indiscriminately determined by all the Z latents taken together.

In particular, we can consider what happens if many of the Z latents are constants. This is a natural case to consider, since there are a priori $2^{|I|} - 1$ latent variables of the form Z_A, and many interesting latent variable models make use of far less than this. Then, it may be that $Z_{\mathcal{G}} = Z_{\mathcal{F}^\circ}$ is equivalent to $Z_{\supseteq A}$ for many meaningfully different choices of $\mathcal{F}$. In this case, we can adapt $\mathcal{F}$ to quantitatively improve the right-hand side, while still controlling the same quantity,

$$H\left(Y_{\supseteq A} \mid Z_{\mathcal{F}^\circ}\right) = H\left(Y_{\supseteq A} \mid Z_{\supseteq A}\right). \tag{3.3}$$

The terms in the inequality of the approximate correspondence theorem. The terms of the form $H\left(Y_{\supseteq A} \mid X_B\right)$ measure the extent to which the latent variables $Y_{\supseteq A}$ can be recovered from the given X variables. If we think of the X variables as observations, then this is an ordinary statistical problem of inferring parameters from data. In other words, these terms are small if, given the joint distribution on the random variables X and Y, the values of the variables $Y_{\supseteq A}$ can be estimated up to a small amount of entropy given the values of X_B.

Terms of the form $I\left(Z_{\mathcal{L}(v)}; Z_{\mathcal{R}(v)} \mid Z_{\mathcal{L}(v)\cap\mathcal{R}(v)}\right)$ measure mutual information; they are large when different Z latent variables share the same information. However, this is conditional mutual information, so repeating information doesn't contribute to such a term if that information is also conditionalized on. We can exhibit this in Example 3.3, where we had a latent variable for the bias of a coin, and an additional latent variable for each flip of the coin. Here, there is mutual information between flips of the coin, but there in no mutual information conditional on the bias—the individual flips are conditionally independent.

In fact, when we define intersection trees, it will turn out $\mathcal{L}(v)$ and $\mathcal{R}(v)$ are defined in such a way that all such terms are zero for Example 3.3. This will also hold of many other interesting latent variable models, such as those constructed from Bayesian networks as in Example 3.5. In general, such terms may be large when we repeat information in latent variables that contribute to different sets of given variables without also stating that information in a latent variable that contributes to many given variables.

In comparison to Wentworth and Lorell [WL23], the role played by the terms $H\left(Y_{\supseteq A} \mid X_B\right)$ is analogous to the role of redundancy, and the terms

$$I\left(Z_{\mathcal{L}(v)}; Z_{\mathcal{R}(v)} \mid Z_{\mathcal{L}(v)\cap\mathcal{R}(v)}\right) \tag{3.4}$$

are analogous to mediation.

4. Set-up

In this section, we will introduce the central concept of latent variable 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. In this definition we ask that the given variables can be recovered from the latent variables, but we don't yet do anything further in order to ask that this serve any organizing role; we don't yet ask that we attain an enlightening perspective on the given variables. As discussed in §3, we want to establish correspondences between different latent variable models associated with the same random variable model, which could be used to understand intersubjectivity. So, we will supplement the definition of latent variable models using 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 4.1. 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 4.2. A *latent variable model for*, or *associated with*, 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 collection $\mathcal{P}^+I$, together with a probability-preserving map $\pi\colon \Lambda \to \Omega$ such that for each random variable X_i the pullback π^*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, π^*X_i is almost everywhere equal to $f_i\left(Y_A : A \subseteq I, A \ni i\right)$ 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 then say that f_i determines the random variable X_i from those Y_A such that A contributes to i.

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

The more important of the scoring functions will be the *conditioned score*. The *simple score* will serve as a simpler analogue, with broadly similar qualitative behavior. These scores represent two desiderata. We would prefer that latent variables compress the random variables of a random variable model, which we can compare to the idea that a good theory should help us compress observations. However, we would also prefer that the latent variables condense information into separate parts that contribute to different given random variables, rather than for example putting all the information in a single latent variable, even at the cost of some degree of compression.

We don't make any strong claim that these scores represent "the best way" to account for these desiderata, or admit a characterization in these terms. Instead, these (or other) scores should be judged by their ability to pick out interesting latent variable models, and by consequences such as the correspondence theorems, Theorem 5.15 and, less directly, Theorem 6.8.

Definition 4.3. 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

$$\sigma_{\mathcal{L}}(A) = \sum_{\substack{B\in\mathcal{P}^+I \\ B\cap A\neq\emptyset}} H(Y_B), \tag{4.1}$$

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

$$\chi_{\mathcal{L}}(A) = \sum_{B\cap A\neq\emptyset} H\left(Y_B \mid (Y_C)_{C\supsetneq B}\right). \tag{4.2}$$

We can see how these quantities relate to our desiderata. Based on the idea of compression, these scores penalize the entropy of the latent variables. By adding together entropies of different variables instead of using the joint entropy, we can penalize redundancy of information between different latent variables.

Because of the dependence of the scores on a set $A \subseteq I$, the scores are in some sense local, measuring the complexity of the latent model as it pertains to the product of the random variables $(X_i)_{i\in A}$. Varying A lets us look at only those latent variables that contribute to a subset of the phenomena to be explained. As we will see in more detail later, in Example 5.1, if we only considered $\sigma_{\mathcal{L}}(I)$ or $\chi_{\mathcal{L}}(I)$ rather than letting $A \subseteq I$ vary, then we would give the best possible score to latent variable models that put all of the information in a single latent variable, rather than a multiplicity of variables serving as distinct conceptual parts. 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 $(Y_A)_{A\in\mathcal{P}^+I}$, and nonnegative real-valued weights $(\lambda_A)_{A\in\mathcal{P}^+I}$, we could define

$$\begin{aligned}\sigma^{\lambda}_{\mathcal{L}} &= \sum_{A\in\mathcal{P}^+I} \lambda_A \sigma_{\mathcal{L}}(A) \\ &= \sum_{A\in\mathcal{P}^+I} \sum_{B\cap A\neq\emptyset} \lambda_A H(Y_B) \\ &= \sum_{B\subseteq I} \left(\sum_{A\cap B\neq\emptyset} \lambda_A \right) H(Y_B)\end{aligned} \tag{4.3}$$

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

We can observe that all the scores are finite, since all the random variables of the form X_i and Y_A have finite entropy.

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

Definition 4.4. 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 $(Y_A)_{A\in\mathcal{F}}$. The product random variable of the latents that contribute to $i\in I$ is defined as

$$Y_{\ni i} = \left(Y_B : B\in\mathcal{P}^+I, i\in B\right), \tag{4.4}$$

using "tuple-builder" notation to express a product random variable. Similarly, for the latents that *contribute* to a set $A\subseteq I$,

$$Y_{\cap A} = \left(Y_B : B\in\mathcal{P}^+I, B\cap A\neq\emptyset\right). \tag{4.5}$$

Given a set $A\subseteq I$, the product of those latent variables that *contribute* to all the latents of A *in common* is

$$Y_{\supseteq A} = \left(Y_B : B\in\mathcal{P}^+I, A\subseteq B\right), \tag{4.6}$$

and we also define

$$Y_{\supsetneq A} = \left(Y_B : B\in\mathcal{P}^+I, A\subsetneq B\right). \tag{4.7}$$

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

$$Y_{\ni i} = Y_{\cap\{i\}} = Y_{\supseteq\{i\}}. \tag{4.8}$$

4.1. **Morphisms.** We will also define morphisms of random variable models. The point of this is only to allow the use of satisfying language; in future definitions and theorems, things that feel like maps will be called "morphism", and things that feel like equivalences will be called "equivalence".

Informally, the idea of a morphism will be to introduce new distinctions. This will be the case reading "from right to left". That is, in a morphism $\left(\Omega,(X_i)_{i\in I}\right)\to\left(\Lambda,(Y_j)_{j\in J}\right)$, the left space Ω may make more distinctions—it may have "more"

measurable sets—than the space Λ. Equivalently, reading from left to right, we could say that we forget measurable sets. This is similar to the situation in probability theory; if $\pi\colon \Omega \to \Lambda$ is a probability-preserving map of probability spaces, then $\pi^{-1}E$ is measurable for every measurable set E in Λ, but not every measurable set of Ω need be of this form. We can think of Ω as an extension of Λ. A morphism of random variable models will thus be similar to a probability-preserving map, but it will 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 4.5. A *morphism* of random variable models has the form

$$\left(\pi, \iota, (f_j)_{j\in J}\right) : \left(\Omega, (X_i)_{i\in I}\right) \to \left(\Lambda, (Y_j)_{j\in J}\right), \tag{4.9}$$

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 that for all $j \in J$, we have $Y_j = f_j\left(X_{\iota(j)}\right)$ 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 there exists a set E of probability one in Ω such that, for all $j \in J$, we have $Y_j = f_j\left(X_{\iota(j)}\right)$ everywhere on E.

(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 4.6. Given two morphisms

$$\left(\pi, \iota, (f_j)_{j\in J}\right) : \left(\Omega, (X_i)_{i\in I}\right) \to \left(\Lambda, (Y_j)_{j\in J}\right) \tag{4.10}$$

$$\left(\rho, \nu, (g_k)_{k\in K}\right) : \left(\Lambda, (Y_j)_{j\in J}\right) \to \left(\Pi, (Z_k)_{k\in K}\right), \tag{4.11}$$

their *composite* is the morphism

$$\left(\rho\circ\pi, \iota\circ\nu, \left(g_k \circ f_{\nu(k)}\right)_{k\in K}\right) : \left(\Omega, (X_i)_{i\in I}\right) \to \left(\Pi, (Z_k)_{k\in K}\right). \tag{4.12}$$

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 $k \in K$,

$$\rho^* Z_k = g_k\left(Y_{\nu(k)}\right) \tag{4.13}$$

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

$$\begin{aligned}(\rho\circ\pi)^* Z_k = \pi^*\rho^* Z_k = \pi^* g_k\left(Y_{\nu(k)}\right) = g_k\left(\pi^* Y_{\nu(k)}\right)\\ = g_k \circ f_{\nu(k)}\left(X_{\iota\circ\nu(k)}\right).\end{aligned} \tag{4.14}$$

as desired.

Proposition 4.7. *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

$$\left(\mathrm{id}_\Omega, \mathrm{id}_I, (\mathrm{id}_{\mathcal{R}\, X_i})_{i\in I}\right), \tag{4.15}$$

which we can see serves as an identity morphism. Further, we can see that composition is associative by checking associativity for any three morphisms
(4.16)

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

The composite of the first two morphisms is

$$\left(\rho\circ\pi, \iota\circ\nu, \left(g_k\circ f_{\nu(k)}\right)_{k\in K}\right) \tag{4.17}$$

and that of the last two is

$$\left(\sigma\circ\rho, \nu\circ o, \left(h_\ell\circ g_{o(\ell)}\right)_{\ell\in L}\right), \tag{4.18}$$

so the triple composite, interpreted in either order, is

$$\left(\sigma\circ\rho\circ\pi, \iota\circ\nu\circ o, \left(h_\ell\circ g_{o(\ell)}\circ f_{\nu(o(\ell))}\right)_{\ell\in L}\right). \tag{4.19}$$

□

Because of the phenomenon of equality almost everywhere, it will be convenient to have available a notion of *equivalence* of random variable models that is more general than isomorphism.

Proposition 4.8. *A morphism of random variable models*

$$\left(\pi, \iota, (f_j)_{j\in J}\right) : \left(\Omega, (X_i)_{i\in I}\right) \to \left(\Lambda, (Y_j)_{j\in J}\right) \tag{4.20}$$

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

$$\pi^* Y_{\iota^{-1}(i)} = f_{\iota^{-1}(i)}\left(X_i\right) \tag{4.21}$$

almost everywhere, so

$$f^{-1}_{\iota^{-1}(i)}\left(Y_{\iota^{-1}(i)}\right) = \left(\pi^{-1}\right)^* X_i, \tag{4.22}$$

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 4.9. 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

$$f_j\left(X_{\iota(j)}\right) = Y_j = g_j\left(X_{\nu(j)}\right) = g_j\left(X_{\iota(j)}\right) \tag{4.23}$$

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

$$\boldsymbol{\pi} = \left(\pi, \iota, (f_j)_{j\in J}\right) \colon \mathcal{M} \to \mathcal{N} \tag{4.24}$$

$$\boldsymbol{\rho} = \left(\rho, \nu, (g_i)_{i\in I}\right) \colon \mathcal{N} \to \mathcal{M} \tag{4.25}$$

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) $f_i\left(X_i\right) = \pi^* Y_i$ and $g_i\left(Y_i\right) = \rho^* X_i$ 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 4.11. *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*

$$\boldsymbol{\pi}, \boldsymbol{\rho}\colon \mathcal{L} \to \mathcal{M}, \qquad \boldsymbol{\sigma}, \boldsymbol{\tau}\colon \mathcal{M} \to \mathcal{N} \tag{4.26}$$

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 almost everywhere. 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 probability, and σ and τ similarly agree on such a set $F \subseteq \Lambda$. The set $E \cap \pi^{-1}\left(F\right)$ has probability one, and for all ω in this set

$$\sigma\circ\pi\left(\omega\right) = \tau\circ\pi\left(\omega\right) = \tau\circ\rho\left(\omega\right), \tag{4.27}$$

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 4.12. *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

$$(\boldsymbol{\rho} \circ \boldsymbol{\tau}) \circ (\boldsymbol{\sigma} \circ \boldsymbol{\pi}) = \boldsymbol{\rho} \circ \boldsymbol{\pi} = \mathrm{id}_{\mathcal{L}} \tag{4.28}$$

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

5. 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? In this section, we will introduce the notion of *perfect condensation*, characterizing when the conditioned scores of a latent variable model are as low as reasonably possible. Under the hypothesis of perfect condensation, we will prove that different latent variable models must correspond in an appropriate sense in Theorem 5.15.

First, we note that we do indeed have uninteresting latent variable models with bad scores.

Example 5.1. 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,

$$\mathcal{L}_1 = \left(\left(\Omega, (Y_A)_{A \in \mathcal{P}^+ I}\right), \mathrm{id}_\Omega\right) \qquad \mathcal{L}_2 = \left(\left(\Omega, (Z_A)_{A \in \mathcal{P}^+ I}\right), \mathrm{id}_\Omega\right) \tag{5.1}$$

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 be the product variable $Z_I = X_I$, and let Z_A be constant for $A \subsetneq I$. We have

$$\sigma_{\mathcal{L}_1}(A) = \sum_{B \cap A \neq \emptyset} H(Y_B) = \sum_{i \in A} H(X_i) \tag{5.2}$$

$$\chi_{\mathcal{L}_1}(A) = \sum_{B \cap A \neq \emptyset} H(Y_B \mid Y_{\supsetneq B}) = \sum_{i \in A} H(X_i) \tag{5.3}$$

$$\sigma_{\mathcal{L}_2}(A) = \sum_{B \cap A \neq \emptyset} H(Z_B) = H(Z_I) = H(X_I) \tag{5.4}$$

$$\chi_{\mathcal{L}_2}(A) = \sum_{B \cap A \neq \emptyset} H(Z_B \mid Z_{\supsetneq B}) = H(X_I). \tag{5.5}$$

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. Since lower scores are "better", this gives a bound on how "good" scores can be.

Proposition 5.2. *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*

$$\sigma_{\mathcal{L}}(A) \geq \chi_{\mathcal{L}}(A) \geq H(Y_{\cap A}) \geq H(X_A). \tag{5.6}$$

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

$$\left\{B \mid B \in \mathcal{P}^+I, B \cap A \neq \emptyset\right\} \tag{5.7}$$

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

$$H\left(Y_B \mid (Y_C)_{C\supsetneq B}\right) \geq H\left(Y_B \mid (Y_C)_{C\prec B}\right). \tag{5.8}$$

Now we can establish the next inequality by a calculation:

$$\begin{aligned}\chi_{\mathcal{L}}(A) &= \sum_{B\cap A\neq\emptyset} H\left(Y_B \mid (Y_C)_{C\supsetneq B}\right) \\ &\geq \sum_{B\cap A\neq\emptyset} H\left(Y_B \mid (Y_C)_{C\prec B}\right) \\ &= H(Y_{\cap A}).\end{aligned} \tag{5.9}$$

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

This motivates a definition of perfect condensation. In addition, we will define simple-perfect condensation, using the simple score in place of the conditioned score. Perfect condensation will be the main object of study in this section, with simple-perfect condensation serving as a simpler model—the things that we have to say about it will all be analogous to things that we say about perfect condensation. It follows from the previous proposition that simple-perfect condensation is a more restrictive property than perfect condensation.

Definition 5.3. A latent variable model $\mathcal{L}$ *perfectly condenses* a random variable model $\mathcal{M} = \left(\Omega, (X_i)_{i\in I}\right)$ if $\chi_{\mathcal{L}}(A) = H(X_A)$ for all $A \subseteq I$. Further, $\mathcal{L}$ *simply-perfectly condenses* $\mathcal{M}$ if $\sigma_{\mathcal{L}}(A) = H(X_A)$ for all $A \subseteq I$.

Example 5.4. 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 $\mathcal{P}^+I$, 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 associated with $\mathcal{M}$ via the identity map id_Ω. For each $i \in I$, we define X_i to be the product random variable

$$X_i = Y_{\ni i} = (Y_A : i \in A \subseteq I). \tag{5.10}$$

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

$$H(X_A) = H(X_i : i \in A) = H(Y_B : B \cap A \neq \emptyset) = \sum_{B \cap A \neq \emptyset} H(Y_B), \tag{5.11}$$

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

In order to acquaint ourselves with perfect and simple-perfect condensation, and to lay the ground for the theorem on the correspondence of perfect condensations, we will prove a variety of statements about these concepts.

Lemma 5.5. *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, for any $A \in \mathcal{P}^+ I$, the following are equivalent.*

1. *$H(Y_{\cap A}) = H(X_A)$*
2. *For all $i \in A$, 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.*

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

$$H(X_i) = H(Y_{\ni i}), \tag{5.12}$$

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

$$H(Y_{\ni i} \mid X_i) = H(Y_{\ni i}) - H(X_i) = 0. \tag{5.13}$$

Hence, using Proposition 2.4, $Y_{\ni i}$ is almost everywhere a function of X_i. Since $i \in A$, it follows that Y_A is almost everywhere a function of X_i as well.

($\Longleftarrow$) We know that $H(Y_{\cap A}) \geq H(X_A)$ by Proposition 5.2. Further, fixing any $i \in A$, the random variable $Y_{\cap A}$ is almost everywhere a function of X_i, and is therefore almost everywhere a function of X_A, so

$$H(Y_{\cap A} \mid X_A) = 0 \tag{5.14}$$

$$H(Y_{\cap A}) \leq H(Y_{\cap A}, X_A) = H(X_A). \tag{5.15}$$

□

Corollary 5.6. *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 5.2, this follows immediately. □

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

Proposition 5.7. *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 $\left(\Lambda, (Y_{\ni i})_{i\in I}\right)$ 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

$$g_i(x) = \left(f_A^i(x) : A \subseteq I, i \in A\right); \tag{5.16}$$

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

$$X_i = h_i(Y_{\ni i}) \tag{5.17}$$

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

$$Y_{\ni i} = g_i(X_i) \tag{5.18}$$

almost everywhere. Let p_A^i be the coordinate projection

$$\mathcal{R}Y_{\ni i} = \prod_{B\ni i} \mathcal{R}Y_B \to \mathcal{R}Y_A. \tag{5.19}$$

Then,

$$Y_A = p_A^i(Y_{\ni i}) = p_A^i(g_i(X_i)), \tag{5.20}$$

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

Corollary 5.6 tells us something about perfect, and hence simple-perfect, condensations. By imposing further conditions, we can define stronger properties, which will give us equivalent characterizations of perfect and simple-perfect condensation. First, we give a definition about probabilistic independence, which can be seen as a form of the Markov condition from the theory of Bayesian networks [SGS01].

Definition 5.8. 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 $Y_{\supsetneq A}$.

We can also state this in a more global equivalent way.

Proposition 5.9. *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) The family Y satisfies the ordered Markov condition.
(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

$$H\left(Y_{\mathcal{G}} \mid Y_{\mathcal{F}}\right) = H\left(Y_{\mathcal{G}} \mid Y_{\mathcal{F}\cap\mathcal{G}}, Y_{\mathcal{F}}\right) = H\left(Y_{\mathcal{G}} \mid Y_{\mathcal{F}\cap\mathcal{G}}\right). \tag{5.21}$$

As in the proof of Proposition 5.2, we will choose a linear order $\preceq$ on $\mathcal{P}^+ I$ such that whenever $A \supseteq B$, we have $A \preceq B$. We can expand

$$\begin{aligned} H\left(Y_{\mathcal{G}} \mid Y_{\mathcal{F}}\right) &= \sum_{A\in\mathcal{G}-\mathcal{F}} H\left(Y_A \mid Y_{\mathcal{F}}, \left(Y_B : B \in \mathcal{G}-\mathcal{F}, B \prec A\right)\right) \\ H\left(Y_{\mathcal{G}} \mid Y_{\mathcal{F}\cap\mathcal{G}}\right) &= \sum_{A\in\mathcal{G}-\mathcal{F}} H\left(Y_A \mid Y_{\mathcal{F}\cap\mathcal{G}}, \left(Y_B : B \in \mathcal{G}-\mathcal{F}, B \prec A\right)\right), \end{aligned} \tag{5.22}$$

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 Y_A is conditionally independent of $Y_{\mathcal{I}_A}$ given $Y_{\mathcal{S}_A} = Y_{\supsetneq A}$. From what we know about $\mathcal{F}$, $\mathcal{G}$, A, and $\preceq$, we have both

$$\mathcal{S}_A \subseteq \mathcal{F} \cup \{B \in \mathcal{G} - \mathcal{F} \mid B \prec A\} \subseteq \mathcal{S}_A \cup \mathcal{I}_A \tag{5.23}$$

and

$$\mathcal{S}_A \subseteq (\mathcal{F} \cap \mathcal{G}) \cup \{B \in \mathcal{G} - \mathcal{F} \mid B \prec A\} \subseteq \mathcal{S}_A \cup \mathcal{I}_A, \tag{5.24}$$

so

$$\begin{aligned} H\left(Y_A \mid Y_{\mathcal{F}}, \left(Y_B : B \prec A\right)\right) &= H\left(Y_A \mid Y_{\supsetneq A}\right) \\ &= H\left(Y_A \mid Y_{\mathcal{F}\cap\mathcal{G}}, \left(Y_B : B \prec A\right)\right) \end{aligned} \tag{5.25}$$

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. □

Now, we can use the ordered Markov condition to characterize perfect condensation.

Theorem 5.10. *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 almost everywhere. 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 5.2 that for each $A \subseteq I$,

$$\sum_{B \cap A \neq \emptyset} H\left(Y_B\right) = \sigma_{\mathcal{L}}\left(A\right) \geq \chi_{\mathcal{L}}\left(A\right) \geq H\left(Y_{\cap A}\right) \geq H\left(X_A\right) = \sigma_{\mathcal{L}}\left(A\right) \tag{5.26}$$

so all these quantities are equal. In particular,

$$\chi_{\mathcal{L}}\left(A\right) = H\left(X_A\right), \tag{5.27}$$

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

$$\sum_{B \in \mathcal{P}^+ I} H\left(Y_B\right) = \sigma_{\mathcal{L}}\left(I\right) = H\left(Y_{\cap I}\right) \tag{5.28}$$

that the latent variables $\left(Y_A\right)_{A \in \mathcal{P}^+ I}$ 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

$$\mathcal{H} = \mathcal{P}^+ I - \mathcal{F} - \mathcal{G}. \tag{5.29}$$

Then, we can calculate

$$\begin{aligned} H\left(Y_{\cap I}\right) &\leq H\left(Y_{\mathcal{F} \cup \mathcal{G}}\right) + H\left(Y_{\mathcal{H}}\right) \\ &\leq H\left(Y_{\mathcal{F}}\right) + H\left(Y_{\mathcal{G}}\right) + H\left(Y_{\mathcal{H}}\right) \\ &\leq \sum_{B \in \mathcal{P}^+ I} H\left(Y_B\right) = H\left(Y_{\cap I}\right), \end{aligned} \tag{5.30}$$

so all these are equal, and so

$$I\left(Y_{\mathcal{F}}; Y_{\mathcal{G}}\right) = H\left(Y_{\mathcal{F}}\right) + H\left(Y_{\mathcal{G}}\right) - H\left(Y_{\mathcal{F} \cup \mathcal{G}}\right) = 0. \tag{5.31}$$

(A3 $\implies$ A2) This is immediate from Corollary 5.6.

(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

$$H\left(Y_{\cap B}\right) \leq H\left(X_B\right). \tag{5.32}$$

Now, using also the independence hypothesis,

$$\sigma_{\mathcal{L}}\left(B\right) \geq H\left(X_B\right) \geq H\left(Y_{\cap B}\right) = \sigma_{\mathcal{L}}\left(B\right), \tag{5.33}$$

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

(B1 $\implies$ B2) The first part of this is simply the statement of Corollary 5.6. Next, choose a linear order $\preceq$ on $\mathcal{P}^+ I$ such that whenever $B \supseteq C$, we have $B \preceq C$. Consider any $A \in \mathcal{P}^+ I$ and let $\mathcal{F}$ be the collection of sets incomparable in the inclusion order to A. In this case, in order to demonstrate the ordered Markov condition, we want to choose $\preceq$ so that every set in $\mathcal{F}$ 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

$$\begin{aligned} H\left(Y_{\mathcal{P}^+I}\right) &= \sum_{B\in\mathcal{P}^+I} H\left(Y_B \mid Y_C : C \prec B\right) \\ &\leq \sum_{B\in\mathcal{P}^+I} H\left(Y_B \mid Y_{\supsetneq B}\right) \\ &= \chi_{\mathcal{L}}\left(I\right) = H\left(Y_{\mathcal{P}^+I}\right), \end{aligned} \tag{5.34}$$

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

$$H\left(Y_A \mid Y_C : C \prec A\right) = H\left(Y_A \mid Y_{\supsetneq A}\right), \tag{5.35}$$

and since

$$\left\{C \mid C \prec A\right\} \supseteq \mathcal{F} \cup \left\{D \mid D \supsetneq A\right\}, \tag{5.36}$$

it follows that

$$H\left(Y_A \mid Y_{\mathcal{F}}, Y_{\supsetneq A}\right) = H\left(Y_A \mid Y_{\supsetneq A}\right). \tag{5.37}$$

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

$$H\left(Y_{\cap A}\right) = H\left(X_A\right). \tag{5.38}$$

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

$$H\left(X_A\right) = H\left(Y_{\cap A}\right) = \sum_{B\cap A\neq\emptyset} H\left(Y_B \mid Y_C : C \prec B\right) \tag{5.39}$$

For each B in this sum, the set $\left\{C \mid C \prec B\right\}$ 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 the ordered Markov condition,

$$H\left(Y_B \mid Y_C : C \prec B\right) = H\left(Y_B \mid Y_D : D \supsetneq B\right) \tag{5.40}$$

for each such B. Hence,

$$H\left(X_A\right) = \sum_{B\cap A\neq\emptyset} H\left(Y_B \mid Y_D : D \supsetneq B\right) = \chi_{\mathcal{L}}\left(A\right) \tag{5.41}$$

as desired. □

Theorem 5.10 gives a strong characterization of perfect condensation, suggesting that it is a rare and rigid property. We expect that many random variable models have no perfect condensation, and that those that do have few, in some appropriate sense. In this way perfect condensation, the condition that $\chi\left(A\right) = H\left(X_A\right)$ for all A, contrasts with the weaker condition that $H\left(Y_{\cap A}\right) = H\left(X_A\right)$, which we characterized in Lemma 5.5. We can always construct latent variable models satisfying $H\left(Y_{\cap A}\right) = H\left(X_A\right)$ for all A, as we did in with the latent variable model $\mathcal{L}_1$ from Example 5.1, in which $Y_{\{i\}} = X_i$ and Y_A is constant for all other A.

Further, we can see that there are meaningfully different latent variable models satisfying the condition $H\left(Y_{\cap A}\right) = H\left(X_A\right)$. 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 5.10. Then, $\mathcal{M}$ would admit a very different random variable model as in Example 5.1 with latents Z, and both these latent variable models would satisfy the same condition:

$$H\left(Y_{\cap A}\right) = H\left(Z_{\cap A}\right) = H\left(X_A\right) \tag{5.42}$$

for all subsets A of the index set.

Theorem 5.15 on the correspondence of perfect condensations will express this rigidity of perfect condensations. Given a random variable model $\mathcal{M}$ and associated latent variable models $\mathcal{L}_1 = \left(\Lambda_1, (Y_A)_{A\in\mathcal{P}^+I}\right)$ and $\mathcal{L}_2 = \left(\Lambda_2, (Z_A)_{A\in\mathcal{P}^+I}\right)$, both of which perfectly condense $\mathcal{M}$, 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, for multiple reasons. As discussed in section 3.3, we will instead say that Y_A is a function of $Z_{\supseteq A}$, and vice versa, but this will only make sense if we can interpret Y and Z as defined on the same probability space.

It may be that the underlying measure spaces of our two latent variable models differ in a way that does not interact with the random variables of interest, but that already provides an obstacle to the existence of an equivalence. Maybe different points of Λ_1 can always be distinguished by some latent variable, but Λ_2 is the product of Λ_1 by a space with a Bernoulli distribution (i.e. a coin flip), 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 comparing $\widetilde{\mathcal{L}_1}$ and $\widetilde{\mathcal{L}_2}$ rather than $\mathcal{L}_1$ and $\mathcal{L}_2$. If these new latent variable models share an underlying measurable space, then we can pull back both the Y and Z variables to that space, and thus make sense of the statement that Y_A is a function of $Z_{\supseteq A}$. The definition of amalgamation will do all this.

Definition 5.11. 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

$$\begin{array}{ccc} & & \Lambda_1 \\ & & \downarrow{\scriptstyle \pi_1} \\ \Lambda_2 & \xrightarrow{\pi_2} & \Omega \end{array} \tag{5.43}$$

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

$$\begin{array}{ccc} \Lambda_0 & \xrightarrow{\rho_1} & \Lambda_1 \\ \downarrow{\scriptstyle \rho_2} & & \downarrow{\scriptstyle \pi_1} \\ \Lambda_2 & \xrightarrow{\pi_2} & \Omega \end{array} \tag{5.44}$$

of probability-preserving maps commutes.

Definition 5.12. 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 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

$$\boldsymbol{\rho}_1 = \left(\rho_1, \mathrm{id}_{\mathcal{P}^+ I}, \left(\mathrm{id}_{\mathcal{R} Y_A}\right)_{A \in \mathcal{P}^+ I}\right) \colon \widetilde{\mathcal{L}_1} \to \mathcal{L}_1 \tag{5.45}$$

$$\boldsymbol{\rho}_2 = \left(\rho_2, \mathrm{id}_{\mathcal{P}^+ I}, \left(\mathrm{id}_{\mathcal{R} Z_A}\right)_{A \in \mathcal{P}^+ I}\right) \colon \widetilde{\mathcal{L}_2} \to \mathcal{L}_2. \tag{5.46}$$

Lemma 5.13. *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*

$$\widetilde{\mathcal{L}_1} = \left(\left(\Lambda_0, (\rho_1^* Y_A)_{A \in \mathcal{P}^+ I}\right), \pi_1 \circ \rho_1\right) \qquad \widetilde{\mathcal{L}_2} = \left(\left(\Lambda_0, (\rho_2^* Z_A)_{A \in \mathcal{P}^+ I}\right), \pi_2 \circ \rho_2\right) \tag{5.47}$$

and

$$\boldsymbol{\rho}_1 = \left(\rho_1, \mathrm{id}_{\mathcal{P}^+ I}, \left(\mathrm{id}_{\mathcal{R} Y_A}\right)_{A \in \mathcal{P}^+ I}\right) \tag{5.48}$$

$$\boldsymbol{\rho}_2 = \left(\rho_2, \mathrm{id}_{\mathcal{P}^+ I}, \left(\mathrm{id}_{\mathcal{R} Z_A}\right)_{A \in \mathcal{P}^+ I}\right). \tag{5.49}$$

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

$$S = \left\{(\lambda_1, \lambda_2) \mid \lambda_1 \in \Lambda_1, \lambda_2 \in \Lambda_2, \pi_1(\lambda_1) = \pi_2(\lambda_2)\right\}. \tag{5.50}$$

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

$$\rho_k(\lambda_1, \lambda_2) = \lambda_k, \tag{5.51}$$

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

$$\pi_0 = \pi_1 \circ \rho_1 = \pi_2 \circ \rho_2. \tag{5.52}$$

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

$$\mathbf{P}_\Omega(\{\omega\}) = \mathbf{P}_1(\pi_1 = \omega) = \mathbf{P}_2(\pi_2 = \omega), \tag{5.53}$$

we can define the conditional probabilities $\omega \mapsto \mathbf{P}_1(\cdot \mid \pi_1 = \omega)$ and $\omega \mapsto \mathbf{P}_2(\cdot \mid \pi_1 = \omega)$. Since Ω is discrete, we can thus define the integral

$$\widetilde{\mathbf{P}_0}(E) = \int_\Omega \left[\mathbf{P}_1(\cdot \mid \pi_1 = \omega)\right] \times \left[\mathbf{P}_2(\cdot \mid \pi_2 = \omega)\right](E) \, \mathrm{d}\mathbf{P}_\Omega \tag{5.54}$$

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

$$\left[\mathbf{P}_1(\cdot \mid \pi_1 = \omega)\right] \times \left[\mathbf{P}_2(\cdot \mid \pi_2 = \omega)\right] \tag{5.55}$$

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.

$$\begin{aligned}(5.56)\qquad \widetilde{\mathbf{P}_0}\left(\bigcup_{k=1}^{\infty} E_k\right) &= \int_\Omega \sum_{k=1}^{\infty} \left[\mathbf{P}_1\left(\cdot \mid \pi_1 = \omega\right)\right] \times \left[\mathbf{P}_2\left(\cdot \mid \pi_2 = \omega\right)\right](E_k)\ \mathrm{d}\mathbf{P}_\Omega \\ &= \sum_{k=1}^{\infty} \int_\Omega \left[\mathbf{P}_1\left(\cdot \mid \pi_1 = \omega\right)\right] \times \left[\mathbf{P}_2\left(\cdot \mid \pi_2 = \omega\right)\right](E_k)\ \mathrm{d}\mathbf{P}_\Omega \\ &= \widetilde{\mathbf{P}_0}\left(\bigcup_{k=1}^{\infty} E_k\right).\end{aligned}$$

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

$$\begin{aligned}(5.57)\qquad \widetilde{\mathbf{P}_0}(S) &= \int_\Omega \left[\mathbf{P}_1\left(\cdot \mid \pi_1 = \omega\right) \times \mathbf{P}_2\left(\cdot \mid \pi_2 = \omega\right)\right]\left(S \cap \pi_0^{-1}(\omega)\right)\ \mathrm{d}\mathbf{P}_\Omega \\ &= \int_\Omega \mathbf{P}_1\left(\pi_1^{-1}(\omega) \mid \pi_1 = \omega\right) \mathbf{P}_2\left(\pi_2^{-1}(\omega) \mid \pi_2 = \omega\right)\ \mathrm{d}\mathbf{P}_\Omega \\ &= \int_\Omega 1\, \mathrm{d}\mathbf{P}_\Omega = 1,\end{aligned}$$

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

$$\begin{aligned}(5.58)\qquad \mathbf{P}_0\left(\rho_1^{-1}(E)\right) &= \int_\Omega \left[\mathbf{P}_1\left(\cdot \mid \pi_1 = \omega\right) \times \mathbf{P}_2\left(\cdot \mid \pi_2 = \omega\right)\right]\left(\rho_1^{-1}(E) \cap \pi_0^{-1}(\omega)\right)\ \mathrm{d}\left(\pi_1\right)_* \mathbf{P}_1 \\ &= \int_\Omega \mathbf{P}_1\left(E \cap \pi_1^{-1}(\omega) \mid \pi_1 = \omega\right) \cdot 1\ \mathrm{d}\left(\pi_1\right)_* \mathbf{P}_1 \\ &= \mathbf{P}_1(E),\end{aligned}$$

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 (5.44) 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. □

We need one more lemma in order to prove the correspondence theorem.

Lemma 5.14. *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

$$\Omega_c = \{\omega \in \Omega \mid C = c\} \tag{5.59}$$

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

$$X = f_1(Y_1) = f_2(Y_2) \tag{5.60}$$

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

$$Y_1(\upsilon) = Y_1(\omega_0) \qquad Y_2(\upsilon) = Y_2(\omega), \tag{5.61}$$

then

$$\begin{aligned} X(\omega) &= f_2(Y_2(\omega)) = f_2(Y_2(\upsilon)) \\ &= f_1(Y_1(\upsilon)) = f_1(Y_1(\omega_0)) \\ &= X(\omega_0). \end{aligned} \tag{5.62}$$

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 υ. 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 5.15 (Correspondence 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. There is an amalgamation of $\mathcal{L}_1$ and $\mathcal{L}_2$, and for any such amalgamation, letting $\widetilde{\mathcal{L}_1} = \left(\left(\Lambda_0, (Y_A)_{A \in \mathcal{P}^+ I}\right), \widetilde{\pi}_1\right)$ 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, 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 5.13. 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 5.10, 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 5.14 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 5.9. Since

$$\bigcap_{i \in A} \mathcal{F}_i = \{B : B \supseteq A\}, \tag{5.63}$$

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

This theorem gives a precise meaning to the idea that perfect condensations are rigid. We have concluded that perfect condensation is such a strong condition that any two perfect condensations have to resemble each other.

However, a disadvantage of rigidity, revealed in Theorem 5.10, is that perfect condensation is rare. This will be addressed in §6, where Theorem 6.8 will consider approximate correspondences, which will giving us meaningful control over latent variable models even when the associated random variable models do not admit any perfect condensations.

We can also reformulate Theorem 5.15 as an equivalence.

Corollary 5.16. *Let $\mathcal{M}$ be a random variable model with random variables $(X_i)_{i \in I}$, suppose that $\mathcal{L}_1$ and $\mathcal{L}_2$ perfectly condense $\mathcal{M}$, and 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 of an amalgamation of $\mathcal{L}_1$ and $\mathcal{L}_2$. Then, there is an equivalence between the latent variable models $\widehat{\mathcal{L}_1} = \left(\left(\Lambda_0, (Y_{\supseteq A})_{A \in \mathcal{P}^+ I}\right), \widetilde{\pi}_1\right)$ and $\widehat{\mathcal{L}_2} = \left(\left(\Lambda_0, (Z_{\supseteq A})_{A \in \mathcal{P}^+ I}\right), \widetilde{\pi}_2\right)$ consisting of morphisms of the form $\left(\mathrm{id}_{\Lambda_0}, \mathrm{id}_{\mathcal{P}^+ I}, (f_A)_{A \in \mathcal{P}^+ I}\right)$ and $\left(\mathrm{id}_{\Lambda_0}, \mathrm{id}_{\mathcal{P}^+ I}, (g_A)_{A \in \mathcal{P}^+ I}\right)$.*

6. Correspondence of latent variable models

6.1. **Suggestive examples.** We can generalize the ideas of the correspondence theorem, Theorem 5.15, beyond the hypothesis of perfect condensation. As we have seen in results like Theorem 5.10, perfect condensation is a significant constraint on the structure of a latent variable model. However, we will see in Theorem 6.8 that an analogue of Theorem 5.15 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 6.1. 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

$$\begin{aligned} \widetilde{Y}_{\{1,\ldots,n\}} &= L \\ \widetilde{Y}_i &= X_i \qquad (i = 1 \text{ to } n), \end{aligned} \tag{6.1}$$

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

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

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 idea that such models are approximately unique will be realized in Theorem 6.8. We will also mention some more diverse examples before continuing.

Example 6.2. 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

$$Y_A = (b(L_j) : j \in S), \tag{6.2}$$

for some bucketing function b, and let

$$Y_{\{i\}} = X_i. \tag{6.3}$$

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.

We can also revisit Example 3.5, using the real definition of latent variable models now.

Example 6.3. Given a Bayesian network [SGS01; PJS17] in which each variable has finite entropy, 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

$$Y_A = (X_j : \exists i \in A.\, j \blacktriangleleft i). \tag{6.4}$$

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

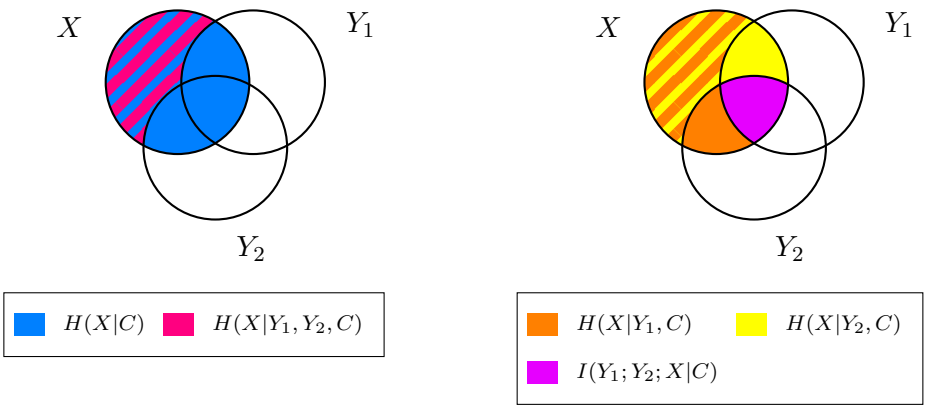

FIGURE 6.1. Information diagrams for Lemma 6.4

recovered from $\mathcal{M}_I$. Thus, we cannot expect an analogue of Theorem 5.15 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 Bayesian networks.

6.2. **Comparison of latent variable models.** We now proceed toward Theorem 6.8, starting with a quantitative variant of Lemma 5.14.

Lemma 6.4. *Let X, Y_1, Y_2, and C be random variables on some probability space, each of which has finite entropy. Then,*

$$H(X \mid C) \leq H(X \mid Y_1, C) + H(X \mid Y_2, C) + I(Y_1; Y_2 \mid C), \tag{6.5}$$

and further, we can make the exact statement

$$\begin{aligned} H(X \mid C) = {} & H(X \mid Y_1, C) + H(X \mid Y_2, C) - H(X \mid Y_1, Y_2, C) \\ & + I(Y_1; Y_2; X \mid C). \end{aligned} \tag{6.6}$$

Proof. We can verify (6.6) with a straightforward if unenlightening calculation. This can be clarified to a certain extent pictorially, as in Figure 6.1.

$$\begin{aligned} & H(X \mid Y_1, C) + H(X \mid Y_2, C) - H(X \mid Y_1, Y_2, C) + I(Y_1; Y_2; X \mid C) \\ & \quad = H(X \mid Y_1, C) + I(X; Y_1 \mid Y_2, C) + I(X; Y_1; Y_2 \mid C) \\ & \quad = H(X \mid Y_1, C) + I(X; Y_1 \mid C) \\ & \quad = H(X \mid C). \end{aligned} \tag{6.7}$$

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

$$\begin{aligned} H(X \mid C) &= H(X \mid Y_1, C) + H(X \mid Y_2, C) - H(X \mid Y_1, Y_2, C) \\ &\qquad + I(Y_1; Y_2; X \mid C) \\ &\leq H(X \mid Y_1, C) + H(X \mid Y_2, C) \\ &\qquad + I(Y_1; Y_2 \mid C) - I(Y_1; Y_2 \mid X, C) \\ &\leq H(X \mid Y_1, C) + H(X \mid Y_2, C) + I(Y_1; Y_2 \mid C). \end{aligned} \tag{6.8}$$

□

Definition 6.5. The *polar* of a subset $\mathcal{F}$ of $\mathcal{P}^+ I$ is the collection

$$\mathcal{F}^\circ = \left\{B \in \mathcal{P}^+ I : \forall A \in \mathcal{F}.\, A \cap B \neq \emptyset\right\}. \tag{6.9}$$

In order to state the approximate correspondence theorem, Theorem 6.8, 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 6.6. 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

$$\ell(u) = \ell(v) \cap \ell(w). \tag{6.10}$$

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

$$v \mapsto (\{a_v, b_v\}, \ell(v)), \tag{6.11}$$

indexed by 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 6.7. *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 Theorem 6.8. This will involve introducing various sets of indices, and then using them to state an inequality, (6.13), as we anticipated in §3. We can compare this inequality to Theorem 5.15 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 5.9. 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 5.15. 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 $\{C : C \supseteq A\}$. 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 $\{C : C \supseteq A\}$ 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 (6.14). This is in some sense stronger, but it has a less clear interpretation, without the analogy to Theorem 5.15. The merit of (6.13) 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, (6.14) 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 6.8 (Approximate correspondence 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 variables 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*

$$\mathcal{G} = \left\{C \in \mathcal{P}^+ I : \forall B \in \mathcal{F}.\, B \cap C \neq \emptyset\right\}; \tag{6.12}$$

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 $\{C \in \mathcal{P}^+ I \mid B \cap C \neq \emptyset\}$ 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

$$H\left(Y_{\supseteq A} \mid Z_{\mathcal{G}}\right) \leq \left[\sum_{B \in \mathcal{F}} H\left(Y_{\supseteq A} \mid X_B\right)\right] + \left[\sum_{v \in N} I\left(Z_{\mathcal{L}(v)}; Z_{\mathcal{R}(v)} \mid Z_{\mathcal{I}(v)}\right)\right]. \tag{6.13}$$

Further, we can make the exact statement

$$\begin{aligned} H\left(Y_{\supseteq A} \mid Z_{\mathcal{G}}\right) = &\left[\sum_{v \in L} H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(v)}\right)\right] - \left[\sum_{v \in N} H\left(Y_{\supseteq A} \mid Z_{\mathcal{L}(v) \cup \mathcal{R}(v)}\right)\right] \\ &+ \left[\sum_{v \in N} I\left(Z_{\mathcal{L}(v)}; Z_{\mathcal{R}(v)}; Y_{\supseteq A} \mid Z_{\mathcal{I}(v)}\right)\right]. \end{aligned} \tag{6.14}$$

Proof. We will first prove (6.14), which we can do following the inductive idea of Theorem 5.15, but now repeatedly applying Lemma 6.4. 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 (6.14) for T_v, which will be

(6.15)
$$H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(v)}\right) = \left[\sum_{w \in L(v)} H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(w)}\right)\right] - \left[\sum_{w \in N(v)} H\left(Y_{\supseteq A} \mid Z_{\mathcal{L}(w) \cup \mathcal{R}(w)}\right)\right] + \left[\sum_{w \in N(v)} I\left(Z_{\mathcal{L}(w)}; Z_{\mathcal{R}(w)}; Y_{\supseteq A} \mid Z_{\mathcal{I}(w)}\right)\right].$$

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 6.4—let s and t be the parents of v.

(6.16)
$$\begin{aligned} H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(v)}\right) &= H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(s)}\right) + H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(t)}\right) - H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(s) \cup \mathcal{I}(t)}\right) \\ &\quad + I\left(Z_{\mathcal{I}(s)}; Z_{\mathcal{I}(t)}; Y_{\supseteq A} \mid Z_{\mathcal{I}(v)}\right) \\ &= \left[\sum_{w \in L(v)} H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(w)}\right)\right] - \left[\sum_{w \in N(s) \cup N(t)} H\left(Y_{\supseteq A} \mid Z_{\mathcal{L}(w) \cup \mathcal{R}(w)}\right)\right] \\ &\quad + \left[\sum_{w \in N(s) \cup N(t)} I\left(Z_{\mathcal{L}(w)}; Z_{\mathcal{R}(w)}; Y_{\supseteq A} \mid Z_{\mathcal{I}(w)}\right)\right] \\ &\quad - H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(s) \cup \mathcal{I}(t)}\right) + I\left(Z_{\mathcal{I}(s)}; Z_{\mathcal{I}(t)}; Y_{\supseteq A} \mid Z_{\mathcal{I}(v)}\right) \\ &= \left[\sum_{w \in L(v)} H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(w)}\right)\right] - \left[\sum_{w \in N(v)} H\left(Y_{\supseteq A} \mid Z_{\mathcal{L}(w) \cup \mathcal{R}(w)}\right)\right] \\ &\quad + \left[\sum_{w \in N(v)} I\left(Z_{\mathcal{L}(w)}; Z_{\mathcal{R}(w)}; Y_{\supseteq A} \mid Z_{\mathcal{I}(w)}\right)\right]. \end{aligned}$$

Specializing this equation to the root, we establish (6.14).

Equation (6.14) follows by a term-by-term comparison. We have

$$\sum_{v \in L} H\left(Y_{\supseteq A} \mid Z_{\mathcal{I}(v)}\right) = \sum_{B \in \mathcal{F}} H\left(Y_{\supseteq A} \mid Z_{\cap B}\right), \tag{6.17}$$

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

$$H\left(Y_{\supseteq A} \mid Z_{\cap B}\right) \leq H\left(Y_{\supseteq A} \mid X_B\right) \tag{6.18}$$

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

$$-H\left(Y_{\supseteq A} \mid Z_{\mathcal{L}_v \cup \mathcal{R}_v}\right) \leq 0 \tag{6.19}$$

and

$$\begin{aligned} &I\left(Z_{\mathcal{L}(v)}; Z_{\mathcal{R}(v)}; Y_{\supseteq A} \mid Z_{\mathcal{I}(v)}\right) \\ &\quad = I\left(Z_{\mathcal{L}(v)}; Z_{\mathcal{R}(v)} \mid Z_{\mathcal{I}(v)}\right) - I\left(Z_{\mathcal{L}(v)}; Z_{\mathcal{R}(v)} \mid Y_{\supseteq A}, Z_{\mathcal{I}(v)}\right) \\ &\quad \leq I\left(Z_{\mathcal{L}(v)}; Z_{\mathcal{R}(v)} \mid Z_{\mathcal{I}(v)}\right). \end{aligned} \tag{6.20}$$

□

6.3. **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 6.8.

Corollary 6.9. *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 $(\{\mathcal{L}(v), \mathcal{R}(v)\}, \mathcal{I}(v))_{v\in N}$, we have

$$H\left(Y_{\supseteq A} \mid Z_{\mathcal{G}}\right) \leq \left[\sum_{B\in\mathcal{F}} H\left(Y_{\supseteq B} \mid X_B\right)\right] + \left[\sum_{v\in N} I\left(Z_{\mathcal{L}(v)}; Z_{\mathcal{R}(v)} \mid Z_{\mathcal{I}(v)}\right)\right]. \tag{6.21}$$

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

$$H\left(Y_{\supseteq A} \mid Z_{\mathcal{G}}\right) \leq \sum_{B\in\mathcal{F}} H\left(Y_{\supseteq B} \mid X_B\right). \tag{6.22}$$

Proof. The first form follows from 6.8 since, whenever $B \subseteq A$,

$$H\left(Y_{\supseteq A} \mid X_B\right) \leq H\left(Y_{\supseteq B} \mid X_B\right). \tag{6.23}$$

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 6.10. *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 $H\left(Y_{\supseteq C} \mid X_C\right) \leq \alpha$ 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 $(\{\mathcal{L}(v), \mathcal{R}(v)\}, \mathcal{I}(v))_{v\in N}$, we have*

$$\mathcal{G} = \left\{C \subseteq I \mid C \text{ contains at least all but } n-1 \text{ elements of } A\right\}, \tag{6.24}$$

and

$$H\left(Y_{\supseteq A} \mid Z_{\mathcal{G}}\right) \leq \binom{|A|}{k} \cdot \alpha + \left[\sum_{v\in N} I\left(Z_{\mathcal{L}(v)}; Z_{\mathcal{R}(v)} \mid Z_{\mathcal{I}(v)}\right)\right]. \tag{6.25}$$

In general, the structure of a latent variable model may lead us to apply Theorem 6.8 to whichever pairs $(A, \mathcal{F})$ we find appropriate, but the form chosen in Corollary 6.10 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.

[Fin37] Bruno de Finetti. "La prévision: ses lois logiques, ses sources subjectives". In: *Annales de l'institut Henri Poincaré* 7.1 (1937), pp. 1–68.

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

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

[Kal05] Olav Kallenberg. *Probabilistic Symmetries and Invariance Principles*. Probability and Its Applications. New York: Springer-Verlag, 2005.

[KS06] G. Jay. Kerns and Gábor J. Székely. "Definetti's Theorem for Abstract Finite Exchangeable Sequences". In: *Journal of Theoretical Probability* 19.3 (Dec. 1, 2006), pp. 589–608.

[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*. 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.

[WL23] John Wentworth and David Lorell. "Natural Latents: The Math". In: (Dec. 27, 2023). URL: https://www.alignmentforum.org/posts/dWQWzGCSFj6GTZHz7/natural-latents-the-math.

[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.