January 2014

Categorical Bayesian Inference

David Orion Girardo
Worcester Polytechnic Institute

Abstract

The language and constructions of category theory have proven useful in unifying disparate fields of study and bridging formal gaps between approaches, so it is natural that a categorial eye should be turned to the theory of probability and its relation to formal logic. [1] Continuing from the foundational work of Lawvere and Giry [2] in developing a functorial theory of probability, Stuartz and Culbertson [3] detail the central importance of and connection between deterministic processes and stochastic processes. Fong expanded this theory to give a categorical account of Bayesian causality. Here we collect and summarize the rich body of research in categorical probability theory, and further develop mathematical machinery for applications in algorithmic Bayesian statistics. Particularly showing that stochastic networks of a certain type satisfy the structural properties of a framed 2-category.

Contents

1 Introduction

> "Much more interesting than the question of what is real is the question of what is causal, i.e. what leads to what."
>
> Daniel M. Ingram

Probability theory has broad applications in nearly every scientific and computational field of research and application. Originally formulated for the study of games of chance, as early as the sixteenth century, probability theory has been continuously developed by such classic mathematicians as Cardano, Fermat, Pascal, and Laplace, until the advent of so called "Modern probability theory", frequently attributed to Kolmogorov [4]. While the first study of probability theory emerged from the analysis combinatorial games of chance, probability theory has expanded to encompass such diverse examples of stochastic processes as coding theory, actuarial mathematics, biological and physical modeling, and decision theory.

One particular usage that has gained prominence in the last few decades with the increasing centrality of computers in modern society is the use of probability theory as a tool for rational automated decision making and knowledge discovery. The majority of information processed by the human mind deals with some degree of uncertainty, and while the theory of discrete logic has been developed extensively, a strong understanding of efficient and effective stochastic reasoning remains illusive. Such studies are deeply connected with the advancement of artificial intelligence, and the fields of formal logic and theoretical computer science.

The language and constructions of category theory have proven useful in unifying disparate fields of study and bridging formal gaps between approaches, so it is natural that a categorial eye should be turned to the theory of probability and its relation to formal logic. [1] Continuing from the foundational work of Lawvere and Giry [2] in developing a functorial theory of probability, Stuartz and Culbertson [3] detail the central importance of and connection between deterministic processes and stochastic processes. Fong expanded this theory to give a categorical account of Bayesian causality. Here we collect and summarize the rich body of research in categorical probability theory, and further develop mathematical machinery for applications in algorithmic Bayesian statistics. Particularly showing that stochastic networks of a certain type satisfy the structural properties of a framed 2-category.

2 A Brief Introduction to Bayesian Probability

> As soon as we recognize that
> *probabilities do not describe reality*
> *only our information about reality* –
> the gates are wide open to the optimal
> solution of problems of reasoning from
> that information.
>
> E. T. Jaynes

While probability theory has proved widely useful for the quantitative sciences in dealing with uncertainty, the epistemic interpretation of probabilities is not entirely obvious. There are many different interpretations and approaches to probability theory, the most prominent of which are the *Frequentist* and *Bayesian* perspectives. Despite their differences, nearly all approaches to probability theory are rooted in a fundamental theory, which we review here.

2.1 Random Variables and Stochastic Systems

Random Variables are the main object of study in probability theory. A random variable is a collection of outcomes with a probability (a real number taking values on the interval $[0, 1]$ associated with each outcome, and so that the sum of probabilities is always 1. We will give a rigorous treatment to random processes in section 4, so here it is best to proceed by example, as the nature of simple random variables is quite intuitive. A system that exhibit randomness and so whose components may be modeled by random variables is called a *stochastic system*.

Example 2.1. The simplest example of a random variable is the uniform distribution on two events, best embodied as a coin flip. Our collection of events are {Heads,Tails}, and if we are considering a fairly weighted coin, then our probabilities are {(Heads,0.5),(Tails,05)}. This simple example illustrates many of the key properties of random variables. Notice that the possible events {Heads,Tails} of the random variable are disjoint; a coin cannot land both heads and tails. It is clear from intuition that if we are considering the logical disjunction of disjoint events ("the probability that the coin lands heads *or* tails") , we should add the probabilities. Further, it is clear that the probability the disjunction of all events is 1 (The coin always lands either heads or tails). One may take this further to consider events with probability 1 as those which occur "almost definitely", and those with probability 0 occur "almost never".

Example 2.2. A slightly less trivial but still common example is the results of the roll of a common six-sided die. We may label the faces {1,2,3,4,5,6}, and if the die is fair then they have probability $\frac{1}{6}$ each. Whereas a coin has only two events, we may consider here more general groupings of events. It is clear that the probability of a roll of 1, 2, or 3, is the sum of those probabilities, but it should also be clear that the probabilities of (1 or 2) or (2 or 3) should not be the sum of their probabilities, or else we would double count 2. In the general case, so long as the sets of events are disjoint, the probability of either collection of events

occurring is the sum of that of each collection, but not so when the collections overlap. In this particular case, we can easily count the probabilities of individual events, but as we will see in the next example, probabilities are most naturally assigned to *collections of events* rather than individual events.

Not all random variables can be fully described by considering the probabilities of individual events. In systems with an incalculably vast event set it is unfeasible to consider individual events, so it becomes prudent to consider the limit in $\mathbb{R}$ of discrete probabilities. If the distribution follows some smoothness conditions, then the probability of any given event in the limit goes to 0, so we cannot compute the probability of a disjoint collection of events simply based on their individual probabilities as in the discrete case. Indeed the problem becomes much worse when not simply dealing with real approximations of discrete systems but physically continuous variables such as the complex amplitudes of quantum physics, which truly have no discrete analogue. In such cases it becomes necessary to define a *probability density function*, where infinitesimal probabilities are assigned to infinitesimal ranges of possible outcomes. The probability that our random variable takes a value in some interval is then computed by integrating the density function over that interval. Because we are integrating over an interval, this method is clearly easier when working with random variables containing a natural order. In practice, most uncountable random variables *do* have a natural order, but there are cases where they do not, and tools for computing the calculation (e.g. Integrating over possible histories with Feynman diagrams).

While determining a proper probability density function of a continuous system is in general more difficult than for a discrete system, maximum entropy methods have proven widely useful in practice. The intuition behind maximum entropy is fundamentally Bayesian, mandating that the probability density be the most general one that produces the observed results. Formally, if the probability of an event is $p(x)$, maximum entropy distributions are those which fit the observed data and maximize the entropy

$$\int_{-\infty}^{\infty} p(x) \log(p(x)) dx$$

which can be thought of as "unpredictability" in this instance. The simplest and most widely used maximum entropy distribution is the "Gaussian" or "Normal" distribution, which maximizes entropy given an observed mean and standard deviation. The normal distribution for mean μ and standard deviation σ is given by

$$\frac{e^{-1\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

Maximum entropy methods can also be useful for variables with merely countably infinite events. In statistical thermodynamics, Boltzmann distributions are commonly used because they maximize entropy subject to the physical constraints on the system.

Example 2.3. For example, the height of a human male in the united states as a continuous random variable. We know that the average (mean) male height is about 70 inches and that about 70% of males are between 67 and 73 inches. Suppose further that we may assume

human height to follow a roughly normal distribution. Then we have a probability density function given by the normal distribution with standard deviation about 3 inches.

$$D(x) = \frac{1}{3\sqrt{2\pi}} e^{-\frac{(x-70)^2}{18}}$$

This is a probability density function, so to calculate the actual probability that a male is between 70 and 71 inches we compute

$$\int_{70}^{71} D(x)dx \approx 13\%$$

But the probability that someone is exactly 70 inches is

$$\int_{70}^{70} D(x)dx = 0$$

This may not seem to connect with real experience ("of course there are people at exactly 70 inches!"), but consider that in performing height measurements, we are implicitly measuring an interval by rounding to the nearest inch, 10th of an inch, etc.

It is this drive for generality that will drive us to consider more abstract definitions, though most cases of interest can be extrapolated from the finite case via standard means.

Definition 2.1 (Conditional Probability)**.** When studying an unknown stochastic system, an attempt to characterize such a system is an attempt to give structure to the underlying probability distribution and infer relations between stochastic events in the system. As we recall from elementary probability theory, the probability of two unrelated stochastic events occurring simultaneously is simply the product of their probabilities $P(A \wedge B) = P(A)P(B)$. However, in more complex systems there may be relations between probabilities, a *flow of information* between them. We infer these relations when the joint probability $P(A \wedge B)$ is not simply the product of independent probabilities. In these cases, we may speak of the probability of an event given a context. The *Conditional Probability* of a random variable A *given* some context C, is written $P(A|C)$, and defined to be $P(A|C) = \frac{P(A \wedge C)}{P(C)}$. The context C is a collection of known information about the system, a collection of known outcomes of other random variables in the system. In the simple case of discrete outcomes, we write $P(A|B = b)$ for the probability distribution on A given that we know the outcome of variable B was b. In the general case, $P(A|B)$ represents the total distribution for all $A = a$ and $B = b$.

Example 2.4. Conditional probabilities are used frequently in interpreting medical data. One such example is the heated case of the early 1990s to determine if smoking really caused lung cancer. In this simplified example let us say that 20 From the definition of conditional probability, we readily obtain the celebrated inversion formula

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

2.2 Bayesianism

There are technical differences in the two approaches, but these differences arise from a difference in philosophy between Frequentist and Bayesian approaches in the interpreted meaning of "probability", and more practically, "conditional probability". Frequentists view probability as the fraction of experiments giving a particular outcome in the limit of infinite experiments, using the "law of large numbers". For example, when we flip a fair coin a large number of times, we expect the fraction of heads to approach one half. In this case, the probability is considered a property of the system to be measured. On the other hand, the Bayesian view takes probability as a measure of uncertainty introduced by incomplete information. We may consider again the flip of a fair coin. A coin flip is still subject to the laws of physics, and in principle the exact trajectory of the coin could be determined each time. Yet, if the trajectory is not measured, the coin is still subject to the laws of probability, reflecting an uncertainty about the result of a (predetermined) process. From a practical standpoint, Bayesian inference is born from Bayes' Rule:

$$P(a|b) = \frac{P(b|a) * P(a)}{P(b)} \tag{1}$$

Where $P(b|a)$ can be read as "the conditional probability of b given a". Bayes' Rule is derived from the identity $P(a|b) = \frac{P(a \wedge b)}{P(b)}$ defining the conditional probability in terms of the joint probability of $a \wedge b$. This gives a universal method for updating beliefs based on new information. These conditional probabilities are taken as the fundamental unit of Bayesian probability, allowing the construction of *probabilistic models*. A model takes some conditionals as given (ex. computed via experimentation) and infers all other probabilities using a chaining of Bayes' Rule. This method of imagining probabilities as states of knowledge elevates conditional probabilities to the "most natural" object of study, representing our expectations based on previous knowledge of the system. In this case, it makes sense to "update" our probabilities, without invoking philosophical issues of what it means for a probability to "change", we simply change our hypothesis based on new information.

2.3 Probabilistic Graphical Models

Definition 2.2 (Independence)**.** While determining causality is tricky, characterizing conditional probabilities is a powerful tool in understanding stochastic systems. In general it is highly valuable to classify variables in two ways, those that have some conditional relationship $P(A \wedge B) \neq P(A)P(B)$ we call *dependent*, whereas those whose joint probabilities are the product $P(A \wedge B) = P(A)P(B)$ we call *independent*. An equivalent characterization that makes the dependence relation more clear is using conditional probabilities to say that variables A and B are *independent* if $P(A|B) = P(A)$ (Or equivalently, $P(B|A) = P(B)$). We may denote that two variables are independent by writing $A \perp B$

Definition 2.3 (Conditional Independence)**.** Since conditional probabilities cannot directly characterize causality, what exactly do they suggest? Judea Pearl et al suggest that causal influence is the primitive unit of connectivity, but that we may only directly measure conditional probability as the transitive flow of causal information between variables. That is,

if a conditional probability does not represent a direct causal influence, then the conditional is mediated by a chain of direct causation of intermediate variables. To make this precise, we introduce the notion of *Conditional Independence*. We say that two variables A, B are conditionally independent given some context (i.e. a collection of random variables) C if the following equivalent conditions hold

1. $P(A|B,C) = P(A|C)$
2. $P(B|A,C) = P(B|C)$
3. $P(A \wedge B|C) = P(A|C)P(B|C)$

We denote that A, B are conditionally independent given a context C by writing $A \perp B|C$

Conditional independence says that causal influence between A, B is mediated by the context C. If there exists no mediating context C (that is, A, B are conditionally dependent given every context C), then we say there is *Potential Cause* between them. To determine the direction of causality (i.e. A causes B or B causes A), we must find an asymmetry in conditionality between some third variable X.

Definition 2.4 (Potential Cause)**.** A random variable A is said to be a *Potential Cause* of B if A, B are dependent in every context ($\forall C, A \not\perp B|C$) and there is a third variable X and a context C so that $A \perp X|C$ and $B \not\perp X|C$. Intuitively this says that there is some other variable X that varies independently of A, but influences B, while A also influences B; then there is no other way for A to influence B other than through direct causation.

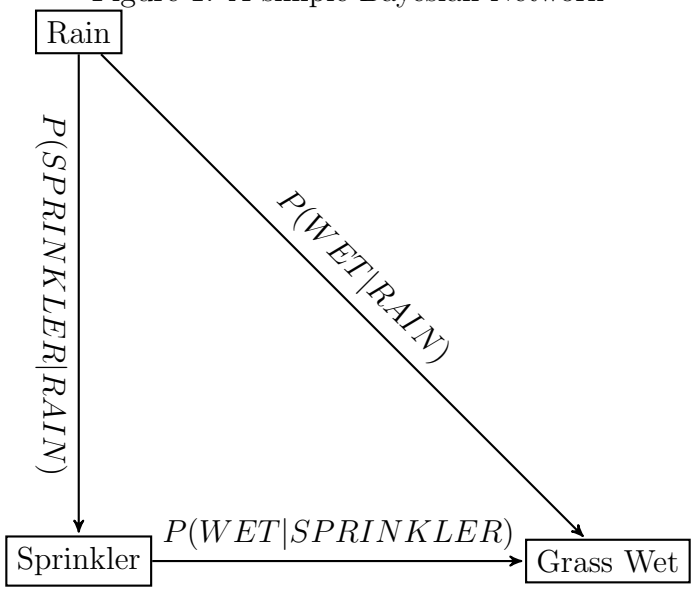
While this case seems strong, we can only say that A is a *potential cause* of B because we are performing the independence calculation with a chosen set of variables included in our model. In reality, there may be additional factors not yet considered. In practice it is impossible to design a model that includes all possible factors as variables for a sufficiently complex system, such as a biological organism, economic system, or many particle physical system. Despite this, we may in many cases determine *Direct Causal Influence* without explicitly considering all possible mediating factors.

Definition 2.5 (Genuine Cause)**.** A random variable A is said to be a *Genuine Cause* of a random variable B if A, B are the transitive closure of the following conditions: $\forall$ contexts $C, A \not\perp B|C$ and there exists a variable X and context C so that

1. X is a potential cause of A
2. $X \not\perp B|C$
3. $X \perp B|A, C$

These conditions can become difficult to keep track of as the number of variables and potential causal influences increases. *Bayesian Networks* were invented as a graphical language to keep track of conditional independence relations.

Figure 1: A simple Bayesian Network



Definition 2.6 (Bayesian Network)**.** A *Bayesian Network* is a directed acyclic graph (DAG) with nodes labeled by random variables and edges $a \to b$ denoting potential causal influence from a to b.

The graphical nature of Bayesian Networks makes them natural tools to facilitate reasoning about stochastic systems. As an example, consider a simple model for determining if it is raining as seen in figure 1

Rain, Sprinkler and Grass Wet are all binary variables, true or false. Intuitively, this model states that Rain causally affects the grass, as well as the sprinkler state (one wouldn't often leave the sprinkler on in the rain). Likewise, the sprinkler can also cause the grass to be wet.We augment this graph to a proper probabilistic model by including probability distributions obtained via experimentation. For example the following tables for $P(SPRINKLER|RAIN)$, $P(SPRINKLER)$, and $P(WET|SPRINKLER, RAIN)$ respectively:

	Sprinkler	
Rain	T	F
F	0.4	0.6
T	0.01	0.99

Rain	
T	F
0.2	0.8

		Grass wet	
Sprinkler	Rain	T	F
F	F	0.4	0.6
F	T	0.01	0.99
T	F	0.01	0.99
T	T	0.01	0.99

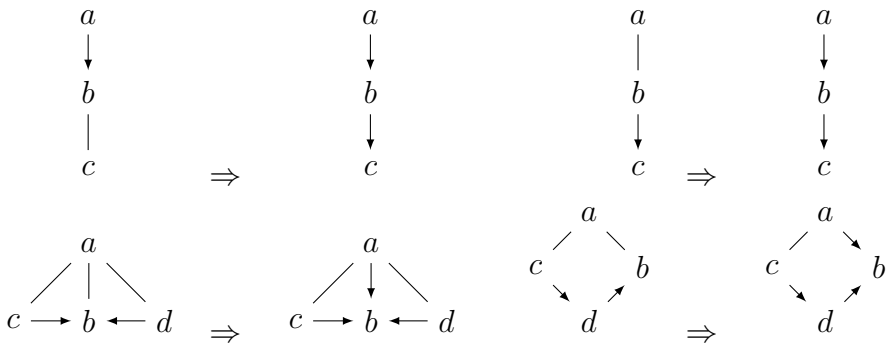
These probability tables may be regarded as simple matrices and multiplied for the purpose of composition. Bayesian networks may be augmented in a number of subtly different ways. We shall still call all such modifications *Bayesian Networks* where the augmentation is clear from context. One such augmentation we will consider here also considers potential ambiguity between cause and effect. In this form of network, an undirected edge $a - b$ represents a conditional dependence between a and b, whereas a directed edge $a \to b$ represents a potential causal influence form a to b.

Using this formalization, Verma and Pearl [5] developed an algorithm (Inductive Causation, or IC algorithm) for inferring potential causal influence from a joint probability distribution.

IC algorithm

Data: $\hat{P}$, a stable distribution on a set V of variables.
Result: A pattern $H(\hat{P})$ compatible with $\hat{P}$.
Start with an empty graph;
foreach *variable* a **do**
  add a vertex labeled a to the graph
end
foreach *pair of variables* a, b *in* V **do**
  search for a set S_{ab} such that $P(a|S_{ab}, b) = P(a|S_{ab})$ $(a \perp b|S_{ab})$;
  if *no such* S_{ab} *exists* **then**
    add an (undirected) edge $a - b$
  end
end
foreach *pair of nonadjacent variables* a, b *in* V **do**
  foreach *common neighbor* c*, of* a *and* b *in* V **do**
    if $c \in S_{ab}$ **then** continue
    else add directed arcs pointing at c i.e. $(a \to c \leftarrow b)$
  end
end
repeat
  foreach *undirected edge* $a - b$ **do**
    if *there is an arc* $c \to a$ **then**
      orient the edge as $a \to b$
    end
    else if *there are arcs* $a \to c$ *and* $c \to b$ **then**
      orient the edge as $a \to b$
    end
    else if *there are undirected edges* $a - c, a - b$*, and directed arcs* $c \to b, d \to b$ *with* c, d *nonadjacent* **then**
      orient the edge as $a \to b$
    end
    else if *there is an undirected edge* $a - c$ *and directed arcs* $c \to d$ *and* $d \to b$ *with* c, b *nonadjacent and* a, d *adjacent* **then**
      orient the edge as $a \to b$
    end
  end
until *there are no orientable edges*;

The conditions for orienting undirected edges in the IC algorithm ensures that any alternative orientation would yield a new v-structure[1] or a directed cycle. These conditions can be visually described by the (complete) subgraphs below.



The first two rules preserve v-structures and the last two prevent directed cycles. This algorithm provides (equivalence classes of) an underlying causal structure to a probability distribution.

This algorithm does not consider the distinction between potential cause and direct cause. Verma and Pearl also developed an augmented graphical language and algorithm for inferring information about direct causation between variables. A reasonable hypothesis of Verma and Pearl is that all conditional dependence relations are induced by direct causal influence of observed or unobserved variables. In this model, we have four types of edges

1. Marked directed arrows $a \xrightarrow{*} b$ representing genuine direct causation of b via a
2. directed edges $a \to b$ representing a as a potential cause of b. This represents either underlying causality $a \xrightarrow{*} b$ or some mediating unobserved variable $a \leftarrow c \to b$
3. undirected edges $a - b$ representing dependence between a and b in all contexts. Assuming that all relations are given by direct causality, this edge represents either direct causation $a \to b$, $b \to a$ or causation by a mediating unobserved variable $a \leftarrow c \to b$
4. bidirected edges $a \leftrightarrow b$ representing some hidden common cause $a \leftarrow c \to b$

Using this formalization, Verma and Pearl developed an extension of the IC algorithm to incorporate direct causal information.

[1]Chains of the form $a \to b \to c, a \to b \leftarrow c, a \leftarrow b \to c$ are distinct v-structures

IC* algorithm

Data: $\hat{P}$, a distribution on a set V of variables.
Result: A pattern $H(\hat{P})$ compatible with $\hat{P}$.
Start with an empty graph;
foreach *variable* a **do**
  add a vertex labeled a to the graph
end
foreach *pair of variables* a, b *in* V **do**
  search for a set S_{ab} such that $P(a|S_{ab}, b) = P(a|S_{ab})$ $(a \perp b|S_{ab})$;
  if *no such* S_{ab} *exists* **then**
    add an (undirected) edge $a - b$
  end
end
foreach *pair of nonadjacent variables* a, b *in* V **do**
  foreach *common neighbor* c*, of* a *and* b *in* V **do**
    if $c \in S_{ab}$ **then** continue
    else add directed arcs pointing at c i.e. $(a \rightarrow c \leftarrow b)$
  end
end
repeat
  foreach *nonadjacent pair* a, b **do**
    foreach *common neighbor* c *of* a, b *with* $(a \overset{*}{\rightarrow} c$ *or* $a \rightarrow c)$ *and* $b - c$ **do**
      orient the edge $b - c$ as $b \overset{*}{\rightarrow} c$
    end
  end
  foreach *undirected edge* $a - b$ **do**
    if *there is a marked directed path* $a \overset{*}{\rightarrow} x_1 \overset{*}{\rightarrow} x_2 ... \overset{*}{\rightarrow} b$ **then**
      orient the edge as $a \rightarrow b$
    end
  end
until *there are no orientable edges*;

The graphical language of Bayesian networks has proven quite useful for reasoning about complex stochastic systems, by allowing researchers to express and quickly infer dependencies between variables without calculating conditional independencies directly each time. Fully marked and directed Bayesian networks may also be used as a tool for stating hypotheses about causation, and suggest further experimentation where the current data suggests only undirected or bidirected edges. In addition to the obvious dependencies of direct causation, Bayesian networks also implicitly depict more complex *induced relations* between variables, where the relationships change depending on observed variables within the network. In general, determining all such induced relations is difficult, but many useful cases my be inferred by the rules of *d-separation*.

Definition 2.7 (d-separation)**.** Collections of variables A, B are considered *d-separated* by a an observed set X if for every (bidirected) path P between variables of A and variables of B, at least one of the following holds.

1. The path P contains a chain $a \to x \to b$ where $x \in X$
2. The path P contains a span $a \leftarrow x \to b$ where $x \in X$
3. The path P contains a collider $a \to y \leftarrow b$ where $x \notin X$

Under these conditions, one can infer a wide range of induced independence relations and suggest variables to observe in experiment to determine genuine cause or to manipulate to most efficiently induce a desired result from a stochastic system. The intuition behind the first condition is that a acts on b through some chain of events (for example, smoking causes tar accumulation in the lungs which cases lung cancer). If the one of the intermediate variables is observed and thus rendered independent of its parent (the probability of it being in the state it was, in fact, observed to be in becomes simply 1), then the chain is broken and b is no longer dependent on a. Following our example, this means that if we observe a patient to smoke heavily, but to happen to have a low level of tar in their lungs, then it does not matter that they have been smoking, only the level of tar is relevant once its value is known). Similarly for the second case, observation or manipulation of a mutual cause renders its children independent. For example, if a predisposition to both smoking and lung cancer were caused by a particular gene, then if one knew that they had the gene, it would not matter how much they smoked, their probability of lung cancer is already fixed by their genetics, though without already knowing their genetics, smoking would indicate a higher chance of the cancer gene.

The final condition is better interpreted as an *induced dependence*. Under normal circumstances, if there is a collider $a \to x \leftarrow b$, then $a \perp b$, we can infer no relation just because they have a common effect. However, if we know the result of their common cause, then the parents become causally related through their common effect, by *explaining away* the cause. For example, both smoking and radon exposure cause lung cancer, but it is clear that smoking does not increase exposure to radon, or radon increase ones desire to smoke, so normally they would be independent. However, if we know that a patient has lung cancer, then if we find that they are also a smoker, then it is much more likely that their lung cancer was caused by smoking, and so, because probabilities of disjoint events must sum to 1, it follows that the probability of radon exposure must decrease. Since the probabilities of all

possible events sum to 1, one can think of "probability mass" as a conserved resource, and explaining away a cause as "pulling probability mass" away from an alternative cause.

A critical feature of Bayesian networks as a graphical language for representing a probability distribution is the *Markov property*, which says that any two variables are conditionally independent given the context consisting of their immediate parents parents (including undirected and bidirected neighbors). Indeed, close observation of the IC and IC* algorithms shows (and has been proven by Verma and Pearl) that any model constructed from this algorithm must have the Markov property. The additional assumption implicit in Bayesian networks is that every conditional relationship emerges from some direct, asymmetric causation of (observed or hidden) variables. This additional assumption makes Bayesian networks a powerful tool for reasoning about temporal relationships and potential interventions. Among other things, the causal assumption requires that there be no directed cycles in the network, as mutually causation is logically unsound.

There exist other graphical models capable of modeling symmetric conditional relationships, and thus cyclic dependencies. One such example is the often used Markov Random Fields

Definition 2.8. A *Markov Random Field* is an undirected graph with vertices labeled by random variables and edges representing conditional dependencies between variables in all contexts, such that all variables obey the Markov property regarding their neighbors.

Markov Random Fields are useful for representing ensembles of mutually interacting variables. For example, the physical laws governing particle interaction are mainly symmetric, so the states of many interacting particles can be modeled by a Markov random field in the shape of a grid, where neighboring particles can interact (and thus obey a dependence relation). Such symmetric relationships of interacting particles are impossible to correctly model in Bayesian networks, because there is (as far as the current physical theories suggests) no mutual cause of the symmetric force of field interactions (ex. magnetism), the mutual interaction occurs simultaneously. On the other hand, Markov random fields cannot represent such complex causal relationships as induced dependencies. Markov random fields are also useful in thermodynamics because they represent exactly the stochastic systems that can be represented by a Gibbs (Boltzmann) distribution. This allows a Markov random field once specified to be generalized to an arbitrary number of interacting particles via transition to a continuous Gibbs distribution.

Both Bayesian networks and Markov random fields represent special cases of a more general principle, a *factorization* of a joint probability distribution. Recall that the joint probability of independent collections of variables is simply the product of their individual probabilities, but dependent variables may be related by more complex relationships. We may express these more complex relationships as functions on the joint variables. For example, if A, B are dependent, then we cannot simply write $P(A \wedge B) = P(A)P(B)$, we must say $P(A \wedge B) = f(A, B)$ for some function $f : A \times B \to \mathbb{R}$

Definition 2.9. A *Factorization* of a joint probability distribution $P(x_i...)$ is a decomposition into factor functions $f_j(\{x_i | i \in I_j\}$ taking values in the index set I_j, so that

$$P(x_i...) = \prod_j f_j(\{x_i | i \in I_j\})$$

For example, for a system with variables A, B, C and $A \perp C$, we have $P(A, B, C) = f_1(A, B)f_2(B, C)$ for some smaller joint functions f_1, f_2. As a causal model, this represents the graph

$$A \xrightarrow{f_1} B \xrightarrow{f_2} C$$

Such a factorization must be represented by a Markov random field because the factorization is given by independent pairwise joint distributions. If the graph above represented an augmented Bayesian network, then we could only say that a, b, c are dependent, and that the distribution cannot factor into an independent product. Thus, we could only say that it is represented by some joint function $f_1(a, b, c)$. This is an example of a case where Markov random fields may be more expressive than Bayesian networks in representing certain factorizations.

If instead of simply a conditional relationship, we have a causal relationship $a \to b$, then we are saying that a is causally independent of b, so for a Bayesian network

$$A \xrightarrow{f_1} B \xleftarrow{f_2} C$$

we can say that b is in fact a function of a and c, so that the probability distribution factors as $P(a)P(c)f_1(a, a)$, where $f_1(a, a)$ represents the value of b given its parents a, c. Such a causal factorization of a joint distribution is one that cannot be represented by Markov random fields, but can be captured by Bayesian networks.

The factorizing captured by Bayesian networks and Markov random fields can be merged into a generalization of both called *factor graphs*, representing arbitrary factorizations of probability distributions.

Definition 2.10 (Factor Graph)**.** In the classic formulation, a *Factor Graph* for a joint distribution $P(x_1, ..., x_n)$ is an undirected bipartite graph with one bipartite part of nodes uniquely labeled by the variables x_i and the other part uniquely labeled by functions $g_k : x_k \to \mathbb{R}$ where $\{x\}_k \subseteq \{x_i\}$ simply denotes the domain of g_k. As their name implies, factor graphs represent a factorization of the joint distribution with $P(x_1, ..., x_n) = \prod_{k=1}^{K} g_k(\{x\}_k)$. Factor graphs can be augmented into directed factor graphs to convey direct causal structure [6]

3 Overview of Category Theory

> By relieving the brain of all unnecessary work, a good notation sets it free to concentrate on more advanced problems, and in effect increases the mental power of the race.
>
> Alfred Whitehead

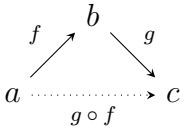
3.1 Besic Definitions

The functional quality, proximity to formal logic, and naturally graphical presentation of probability networks all point towards category theory as a potential tool for further analysis. Category theory is a natural setting for unifying diverse concepts in mathematics, and studying higher order relationships. In this manuscript we mainly follow Mac Lane's classic formalism [7] but refer to some more modern works where noted.

Definition 3.1 (Category)**.** A *Category* C is a collection (class) of *objects* $c \in C$, together with a collection of *morphisms* or *arrows* $f : X \to Y$ between objects of C. In addition, we have the following two constructions

1. (*Identity*) For each object $a \in C$, a unique identity morphism $\mathrm{id}_a = 1_a : a \to a$
2. (*Composition*) For each pair of morphisms $(f : a \to b, g : b \to c)$, a composite morphism $g \circ f : a \to c$

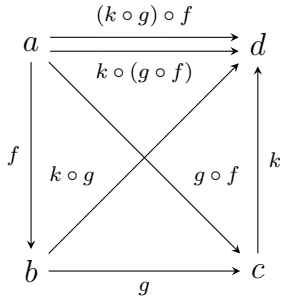
We may represent functional equations as a labeled directed graph called a *Diagram* such as



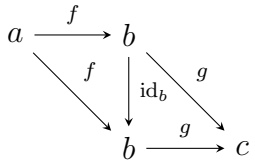
We say that a diagram *Commutes*, and so call it a *Commutative Diagram*, if we additionally hold that all paths in the diagram are equal as functions, and dotted arrows represent necessary morphisms given the others. For example, the above is a commutative diagram for the composition law.
The identity and composition constructions additionally satisfy two laws:

1. (*Associativity.*) For objects and and arrows like $a \xrightarrow{f} b \xrightarrow{g} c \xrightarrow{k} d$ it is always the case that $k \circ (g \circ f) = (k \circ g) \circ f$. This property can be represented by the commutative

diagram



2. (*Unit Law.*) For all arrows $f : a \to b$ and $g : b \to c$, $\mathrm{id}_b \circ f = f$ and $g \circ \mathrm{id}_b = g$. This can be represented by the commutative diagram



In many cases it is a useful intuition to consider morphisms as "generalized homomorphisms", though they they can represent much more.

Definition 3.2. Dealing with categories defined over proper classes can run into similar problems encountered in set theory, thus, we typically like to work with "smaller" categories. A category is *Small* if the class of objects is a set, and *Large* if it is a proper class. In many cases, we can get away with a weaker condition. A *Locally Small* category is one such that for any two objects $a, b \in C$, the set of morphisms $a \to b$ is a set.

Definition 3.3 (Functor)**.** A factor $F : C \to C'$ is a map on objects and morphisms. T maps each object c to an object Tc of C', and each morphism f to a morphism Tf of C' to an object of C' subject to the morphism constraints

$$T(1_c) = 1_{Tc}, \quad T(g \circ f) = Tg \circ Tf$$

Example 3.1. We may construct the category **Cat** of small categories with functors between them as morphisms in the category. While **Cat** is not small, it is locally small.

Definition 3.4 (Diagram)**.** We have informally discussed *commutative diagrams* already. Formally, a *Diagram in C of type J* is a (covariant) functor $D : J \to C$. We call J the index category, and sometimes refer to D as a *J-shaped diagram in C*. Intuitively, the objects and morphisms of J do not matter much, we only care about their relative interconnection as it maps into C. This definition gives strength to our informal definition of commutative diagrams. Commutative diagrams simply represent diagrams with the identity arrows undrawn. Note however that not all diagrams defined in this way commute; asserting that a diagram commutes is an additional property that we posit as mathematicians.

Definition 3.5 (Natural Transformation)**.** For any two categories C, B the *Functor Category*, written B^C has functors $C \to B$ as objects. The morphisms in B^C are called *Natural Transformations*. For two functors $S, T : C \to B$, a *Natural Transformation* $\tau : S \xrightarrow{\bullet} T$ maps each object $c \in C$ to an arrow $\tau_c : Sc \to Tc$ of B so that for each arrow $f : c \to c'$ in C, $\tau_{c'} \circ F(f) = G(f) \circ \tau_c$. Diagrammatically, the follow diagram commutes

$$\begin{array}{ccc} c & & Sc \xrightarrow{\tau_c} Tc \\ \downarrow f & & {\scriptstyle Sf}\downarrow \qquad \downarrow{\scriptstyle Tf} \\ c' & & Sc' \xrightarrow{\tau_{c'}} Tc' \end{array}$$

Intuitively, a natural transformation is a map between functors with the same domain and codomain, giving a way to transform the image of S into the image of T by pointwise "transport" of the objects. Considering S and T as C shaped diagrams in D, we are taking the diagram S and "pulling" it into the diagram T and creating a "3D" diagram.

3.2 Constructions

Example 3.2 (Arrow Category)**.** The idea of natural transformations being morphisms between functors can be generalized to the notion of an *Arrow Category* $C\!\uparrow$ for any category C. The arrow category has objects the morphisms of C, and morphisms $\alpha : (m : A \to B) \to (n : C \to D)$ are the pairs of functions (f, g) forming the commuting squares

$$\begin{array}{ccc} A & \xrightarrow{m} & B \\ {\scriptstyle f}\downarrow & \Downarrow \alpha & \downarrow{\scriptstyle g} \\ C & \xrightarrow[n]{} & D \end{array}$$

Equivalently, we may formally construct this category as the category C^I where I is the category with two objects $0, 1$ and a single non-identity morphism $0 \to 1$.The effect of such a functor $F \in C^I$ is to pick out an arrow $F(0) \to F(1)$ in C, and a natural transformation $F \to G$ is then a pair of arrows $F(0) \to G(0), F(1) \to G(1)$ so that the natural transformation laws hold.

Definition 3.6. A particular functor if interest is the functor $\mathbf{op} : C \to C^{\mathrm{op}}$ that makes each category to its *opposite category*, or *dual category*, the category with all arrows reversed. **op** is used so frequently that we call a functor $F : C^{\mathrm{op}} \to D$ *contravariant* to call attention to the fact that such a functor precomposes arrows in reverse, by analogy to tensor analysis. A "normal" functor is then called *covariant*. Since duality is an isomorphism, every construction on a category gives us an equivalent dual construction on its opposite category.

Definition 3.7 (Product Category)**.** For any two categories C, D we may construct their *product category* $C \times D$ with objects as pairs $\langle c, d\rangle \quad c \in C, d \in D$ and arrows $\langle f, g\rangle : \langle c, d\rangle \to \langle c', d'\rangle \quad f : c \to c', g : d \to d'$. For each product category $C \times D$ we have projection functors $\Pi_C : C \times D \to C$ and $\Pi_D : C \times D \to D$ which simply select the corresponding component in the objects or arrows.

Definition 3.8 (Multifunctor)**.** A *Multifunctor* is a functor from product categories, or may be thought of as functors in multiple arguments. If the functor is from a single product $F : B \times C \to D$ we call it a *Bifunctor*. Such functors are quite common, for example the Cartesian product bifunctor $\times : \mathbf{Set} \times \mathbf{Set} \to \mathbf{Set}$, or more generally the product functor : $\mathbf{Cat} \times \mathbf{Cat} \to \mathbf{Cat}$ mapping a pair of categories to its product category. We may have multifunctors from dual categories such as $F : B \times C^{\text{op}} \times D \to X$. In this case, we call F contravariant in C and covariant in B and D.

Definition 3.9 (Hom-functor)**.** One of the most common bifunctors is the *Hom-functor* for a category C, $\text{hom} : C^{\text{op}} \times C \to \mathbf{Set}$, mapping a pair of objects in C to the set of morphisms between them, and each pair of arrows $f : a \to a', g : b \to b'$ to the set function $\text{hom}(f, g) : \text{hom}(a', b) \to \text{hom}(a, b')$ by the action $g \circ h \circ f$ for $h \in \text{hom}(a', b)$. We often write the hom map as $C(-,-)$ for the category C. Note that hom is contravariant in its first argument because the application involves precomposition: it maps an arrow $f : a \to a'$ to a function in the opposite direction. This can be illustrated by fixing $g = \text{Id}$ and considering, for $f : a \to a', C(f, \text{Id}) : C(a', b) \to C(a, b)$. We say that a property holds *Locally* in a category C if it is true for all hom-sets $C(a, b) \quad \forall a, b \in C$

3.3 Enriched and Higher Catgories

Definition 3.10 (Enriched Category)**.** In many cases, the collection of arrows between objects in a category has more structure than a simple set, and in fact has the structure of some category $\mathcal{V}$. In this case we may replace the notion of the hom functor with a similar functor $C^{\text{op}} \times C \to V$. Here we say that C is *Enriched over* $\mathcal{V}$, or simply a $\mathcal{V}$*-Category*. Note that in order for such a replacement to make sense, the category $\mathcal{V}$ must have additional properties analogous to hom-set composition and the identity arrow, namely, it must be a Monoidal Category

Definition 3.11 (Strict Monoidal Category)**.** As the name suggests, a monoidal category is a category that is also a monoid. Formally, a *(Strict) Monoidal Category* is a category C together with

- A bifunctor $\otimes : C \times C \to C$ called the *tensor product*
- An identity object I

Satisfying

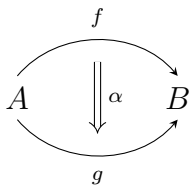
$$(A \otimes B) \otimes C = A \otimes (B \otimes C) \qquad \text{(Associativity)}$$

$$A \otimes I = I \otimes A = A \qquad \text{(Unit)}$$

Notice that the associativity and unit property of a monoidal category are exactly the properties required of arrows in a category. Indeed, this structure is sufficient to allow objects of a monoidal category $\mathcal{V}$ to act as "Arrow objects" for some category enriched in $\mathcal{V}$

Definition 3.12 (Strict 2-Category)**.** The notion of a category may be generalized to provide for morphisms between morphisms, by way of 2-categories. There are many equivalent and isomorphic definitions of 2-categories but often their construction Because we are dealing

only with strict 2-categories here (where associativity and identity laws are equalities, rather than only holding up to natural isomorphism), we may simply sat that a *(Strict) 2-Category* C is a category enriched over **Cat**. That is, the collection of morphisms from any object $a \in C$ to any other $b \in C$ in a 2-category are themselves objects in their own category $C(a, b)$. To avoid confusion in higher categories, we call the objects 0-cells, the morphisms 1-cells, and the morphisms between morphisms 2-cells. We can draw the 0-cell A, B, 1-cells $f, g : A \to B$, and 2-cell $\alpha : f \Rightarrow g$ as



The study of 2-categories and more general n-categories is an enormous field and represents a great expansion of category theory since the classic work of Maclane. Many concepts of classic category theory are elucidated or simplified extensively by a transition to a higher category perspective. We will need only very basic 2-categorical machinery here, but the interested reader is referred to [8], [9] and [10] for a more general definition, extensive study of 2-categories, and higher categories respectively.

Example 3.3. The prototypical example of a 2-category is **Cat** considered as a 2-category with 0-cells objects, 1-cells functors, and 2-cells natural transformations. This follows readily from enriching 1-**Cat** over the disjoint union of all functor categories $\mathcal{B}^{\mathcal{A}} \quad \forall \mathcal{A}, \mathcal{B} \in \mathbf{Cat}$

4 Categorical Probability Theory

> Category Theory is the subject where you can leave the definitions as exercises.
>
> John C. Baez

Here we develop two categories $\mathcal{M}$eas and **Stoch** to give a formal categorical basis for probability theory and then show that these two categories can be unified into a single construction, the framed 2-category $\mathcal{P}$rob. The categories $\mathcal{M}$eas and **Stoch** were originally characterized by Lawvere[2] and Giry [2], but here we follow the more modern interpretation of Culbertson & Sturtz [3] and Fong [11]. The basics of probability and measure theory discussed here may be found in any introductory text on the subject and the reader is referred to the texts of Ash [12] and Dudley [13] for a deeper coverage of these topics.

4.1 Measurable Sets and Functions

Definition 4.1 (Measurable Space)**.** A *σ-algebra* over a set X is a subset $\Sigma_X \subseteq 2^X$ of the powerset of X satisfying:

1. Nonempty: $\Sigma_X \neq \emptyset$
2. Complements: $A \in \Sigma_X \implies X - A \in \Sigma_X$
3. Unions: $A, B \in \Sigma_X \implies A \cup B \in \Sigma_X$

We call the elements $S \in \Sigma_X$ *Measurable Sets*, and the pair (X, Σ_X) a *Measurable Space*, and a function $f : (X, \Sigma_X) \to (Y, \Sigma_Y)$ between measurable spaces a *Measurable Function* if $f^{-1}(y \in \Sigma_Y) \in \Sigma_X$. For simplicity we refer to a measurable space simply as X if the context is clear, Σ_X is the σ-algebra on X, and σ_X is an element of Σ_X

Definition 4.2 (Meas)**.** Measurable spaces form a category $\mathcal{M}$eas, with objects as measurable spaces and morphisms as measurable functions. [3].

In addition, this category forms a monoidal category where the identity object is any singleton measurable space $\mathbf{1} \overset{\text{def}}{\equiv} (\{*\}, \{\emptyset, \{\}\})$ and the tensor product acts on objects $(A, \Sigma_A) \times (B, \Sigma_B) \mapsto (A \times B, \Sigma_A \times \Sigma_B)$ sending pairs of measurable spaces to the measurable spaces defined by their Cartesian product.

Note that the Cartesian product is not strictly associative, nor is the Cartesian product $A \otimes \mathbf{1}$ exactly equal to A, but merely isomorphic (by the map $(a, *) \leftrightarrow a$). Such a category whose coherence conditions only hold up to isomorphism are called *weak* monoidal categories. The more general definition is given in the appendix, but we will not concern ourselves with the technicalities here, as every weak monoidal category is equivalent to a strict one by Mac Lane's coherence theorem [7]. So we may proceed by simply remembering to use the appropriate coherence isomorphisms when necessary.

[2]Unpublished, Lawvere was Giry's advisor in the following work

Example 4.1 (Countably Generated σ-Algebra)**.** An important subclass of σ-algebras are those that are *Countably Generated*. That is, those that are constructed as the closure under complements and unions of a countable collection of "generating sets". Notably, this excludes the standard σ-algebra on $\mathbb{R}$ (the Lebesgue sets or the Borel sets).

Example 4.2 (Borel σ-algebra)**.** A particularly interesting class of measurable spaces are those generated by a topological space, called the *Borel σ-algebra*. Recall that a topological space is a set X and a collection τ of "open" subsets of X called a *topologoy* on X where τ satisfies the following properties

1. $\emptyset, X \in \tau_X$
2. τ is closed under arbitrary unions
3. τ is closed under finite unions

The borel closure $\mathfrak{B}(\tau_X)$ is given by completing τ_X with complements i.e. by including also the closed sets. Typically we just write τ for $\mathfrak{B}(\tau)$ or refer to the whole induced measurable space as simply X when the context is clear.

Example 4.3 (Discrete σ-algebra)**.** One example of a Borel algebra the we will use extensiviely is that of the *discrete σ-algebra*. The discrete algebra on a finite set X is simply the powerset of X i.e. the set of all subsets of $X, \{\sigma_X | \sigma_X \subseteq X\}$

Definition 4.3 (Measure)**.** A *Measure* on a σ-algebra Σ_X is a map $\mu : \Sigma_X \to \mathbb{R}$ satisfying:

1. Non-negative: $\forall A \in \Sigma_X : \mu(A) \geq 0$
2. Null empty set: $\mu(\emptyset) = 0$
3. σ-additivity: For $\sigma_X = \bigcup_i A_i$ for (countably many) mutually disjoint A_i then $\mu(\sigma_X) = \sum_i \mu_x(A_i)$

If in addition $\mu(X) = 1$ then we further call μ a *Probability Measure*

Definition 4.4 (Measure Space)**.** A *Measure Space* is a measurable space equiped with a measure. Formally, it is a triple (X, Σ_X, μ_X) with X a set, Σ_X a σ-algebra over X and μ_X a measure on X. If μ is a probability measure then we further call it a *Probability Space*. We will mainly concern ourselves with probability spaces here. When there is no ambiguity we refere to a measure space simply by its set (ex: X), and always refer to its measure by μ_X.

Measure spaces also form a category **Rand** with objects as *perfect*[3] measure spaces (X, Σ_X, μ_X) and morphisms $\phi : (X, \Sigma_X, \mu_X) \to (Y, \Sigma_Y, \mu_Y)$ measurable functions on the

[3]A measure space X is perfect if for any measurable function $f : X \to \mathbb{R}$, there exists a Borel set $E \subset f(X)$ such that $\mu_X(f^{-1}(E)) = \mu(X)$. This stipulation ensures that, among other things, we end up with "optimal" conditionals in the sense that we use as much information as possible to formulate conditional probabilities, and never "forget" information along the way. The formalization is tedious, and the results match standard intuition for classic Bayesian probability from section 2, so the interested reader is referred to [14] for a complete exposition. The intuition is that when performing Bayesian inference, we speak of "The" probability of some variable X, or "The" conditional probability between variables. This restricts our category to unique morphisms between any two objects representing the "optimal" probability given the data

underlying measurable space such that ϕ preserves the measure

$$\forall \sigma \in \Sigma_Y. \quad \phi_* \mu_X(\sigma) \stackrel{\text{def}}{\equiv} \mu_X(\phi^{-1}(\sigma)) = \mu_Y(\sigma) \tag{2}$$

We need only show that the composition of measure preserving maps is also measure preserving because associativity and the identity are inherited from **Set**. Let $X \stackrel{\phi}{\to} Y \stackrel{\psi}{\to} Z$ for X, Y, Z measure spaces. Then $\mu_X((\psi \circ \phi)^{-1}(\sigma_Z) = \mu_X(\phi^{-1}(\psi^{-1}(\sigma_Z))) = \mu_Y(\psi^{-1}(\sigma_Z)) = \mu_Z(\sigma_Z)$ by simply applying the measure preserving properties of ϕ and ψ in sequence.

Example 4.4 (Characteristic Function)**.** Define $\chi : X \times \Sigma_X \to \mathbb{R}$ be the characteristic function $\chi(x, \sigma_X) = \begin{cases} 1, & \text{if } x \in \sigma_X \\ 0, & \text{otherwise} \end{cases}$ It is easy to check that $\chi(x, \cdot)$ for a fixed x is a probability measure. Similarly, $\chi(\cdot, \sigma_X)$ for a fixed σ_X is a measurable function when considering $\mathbb{R}$ as a measure space equipped with its standard borel σ-algebra.For simplicity we write $\chi_x \stackrel{\text{def}}{\equiv} \chi(x, \cdot)$ and $\chi_{\sigma_X} \stackrel{\text{def}}{\equiv} \chi(\cdot, \sigma_X)$

Definition 4.5 (Atomic)**.** Given a measure space X, a set $\sigma \in \Sigma_X$ is an *atom* if

$$\forall (\epsilon \in \Sigma_X) \subset \sigma. \quad \mu_X(\sigma) > \mu(\epsilon) = 0$$

A measure space is said to be *atomic* if it has at least one atom.

Remark 4.1. Every countably generated σ-algebra is atomic, and the atoms are given by the intersection over generating sets containing a given point. For a family of generating sets $\mathcal{C}_X$, the atoms are

$$\text{Atoms} = \{A | x \in X, A = \bigcap_{C \in \mathcal{C}_X, x \in C} C, \mu_X(A) > 0\} \tag{3}$$

The proof follows readily from the well known theorem of Sierpiński [15] that if a measure space X is non-atomic then

$$\forall b \in \mathbb{R}, A \in \Sigma_X, \text{ s.t. } \mu_X(A) \geq b \geq 0. \quad \exists (B \in \Sigma_X) \subseteq A. \text{ s.t. } \mu(B) = b$$

Most importantly, the measure μ of a countably generated space is uniquely determined by its values on the atoms. To see this take the intersection of some set Further, the atoms of a countably generated σ-algebra themselves form a generating family for the algebra. It follows that the measure is uniquely determined by its values on atoms, since every set is formed through complement and countable union of atoms, and remembering σ-additivity

4.2 Categories of Stochastic Maps

Maps between measure spaces may alternatively be thought of as "measures parameterized by a set" as follows.

Definition 4.6 (Stochastic Map)**.** A *Stochastic Map* $k : (X, \Sigma_X) \nrightarrow (Y, \Sigma_Y)$ between measurable spaces is a function $k : X \times \Sigma_Y \to [0, 1]$ such that $k(x, \cdot) : \Sigma_Y \to [0, 1]$ is a probability measure on Y and $k(\cdot, \sigma) : X \to [0, 1]$ is measurable

The category **Stoch** has objects as measurable spaces and morphisms stochastic maps. Composition of Stochastic maps $f \circ g : (X, \Sigma_X) \xrightarrow{f} (Y, \Sigma_Y) \xrightarrow{g} (Z, \Sigma_Z)$ is via the integral:

$$f \circ g(x, \sigma_Z) \stackrel{\text{def}}{\equiv} \int_{y \in Y} f(y, \sigma_Z) g(x, dy) \tag{4}$$

Definition 4.7 (Product Measure)**. Stoch** is also a monoidal category with the same monoidal unit as $\mathcal{M}$eas ($I = \mathbf{1} \stackrel{\text{def}}{\equiv} (\{*\}, \{\emptyset, \{*\}\})$). The tensor product acts on objects in the same way as $\mathcal{M}$eas, that is, it sends pairs of measurable spaces to their product measurable space given by the Cartesian product of their σ-algebras. On morphisms, the tensor product sends pairs of stochastic maps $(M : A \nrightarrow B, N : C \nrightarrow D)$ to the product $M \times N : A \times C \nrightarrow B \times D$ defined by $M \times N((a, c), (\sigma_B, \sigma_D) = M(a, \sigma_B) * N(c, \sigma_D)$. Thus the tensor product represents the joint measure where the two objects are independent.

Example 4.5 (Countably Generated Measure Spaces)**.** If we restrict ourselves to countably generated measure spaces, then the integral takes a particularly simple form since we have a surjective mapping from the points $y \in Y$ to the atomic sets on Y. Since a measure is uniquely determined by its value on atoms, we may take our dy to range over the atoms A_y corresponding to the point y, so the integral reduces to

$$f \circ g(x, \sigma_Z) = \sum_{z \in \sigma_Z} \sum_{y \in Y} f(y, A_z) g(x, A_y) \tag{5}$$

Example 4.6. In the case of finite sets X with the discrete σ-algebra $\Sigma_X = 2^X$ (powerset), the atoms further reduce to be a singleton sets. Then the composition of stochastic maps simplifies to

$$f \circ g(x, \sigma_Z) \equiv \sum_{z \in \sigma_Z} \sum_{y \in Y} f(y, \{z\}) g(x, \{y\}) = [\sigma_Z]^T F G[\{x\}]$$

where F, G are matrices indexed by set elements (regarded to have some arbitrary fixed ordering) with $G_{y,x}, F_{z,y}$ the measures on $g(x, \{y\}), f(y, \{z\})$ and $[\sigma_Z]$ regarded as the length $|Z|$ characteristic row vector with elements $[\sigma_Z]_z = \chi(z, \sigma_Z)$

Remark 4.2. The integral in equation (4) is taken to be the appropriate Lebesgue integral on the space. For reasons to become clear shortly, we limit ourselves to countably generated measure spaces for the majority of this paper, where such integrals take this particularly simple form. All integration results hering are valid in the more general context so long as one checks well definedness and convergence conditions for the corresponding integrals.

Definition 4.8 (Deterministic Maps)**.** There is a special case of stochastic maps which may be called *deterministic*. These are the stochastic maps having values in only $\{0, 1\}$. Such maps are called deterministic because their input/output pair happen almost surely (1) or never (0).

The characteristic function is one example of a deterministic stochastic map, but we will now classify the more general cases

Definition 4.9 (Point Measure)**.** For a measurable function $f : (X, \Sigma_X) \to (Y, \Sigma_Y)$ define the *point measure*

$$\delta_f(x, \sigma_Y) \stackrel{\text{def}}{\equiv} (\chi_{\sigma_Y} \circ f)(x) \equiv \begin{cases} 1, & \text{if } x \in Y \\ 0, & \text{if } x \notin Y \end{cases}$$

Proposition 4.1. *For measurable functions $f, g : (X, \Sigma_X) \stackrel{f}{\to} (Y, \Sigma_Y) \stackrel{g}{\to} (Z, \Sigma_Z)$, the composition of their point measures is the point measure of their composition $\delta_g \circ \delta_f = \delta_{g \circ f}$.*

Proof: Expanding the definition of composition and point measure:

$$(\delta_g \circ \delta_f)(x, \sigma_Z) \equiv \int_{y \in Y} \chi(g(y), \sigma_Z) \chi(f(x), dy)$$

Since χ takes values $0, 1$, we may look at each component individually. $\chi(f(x), \cdot)$ is equivalent to the statement $f(x) \in dy$, but since dy is an infinitesimal, this is only true if $f(x) = y \in Y$. Similarly $\chi(g(y), \sigma_Z)$ but $y = f(x)$ so we have the integral over $g(f(x)) \in \sigma_Z$ $\square$

In fact, for countably generated measurable spaces, the point measures completely characterize deterministic maps. More precisely, [3] (Proposition 2.1) showed that

Proposition 4.2. *If (Y, Σ_Y) is a countably generated measurable space, then a stochastic map $X \stackrel{M}{\nrightarrow} Y$ is deterministic if and only if $M = \delta_f$ for some measurable function $f : X \to Y$.*

another important property, also from [3] is that

Proposition 4.3. *If $(X, \Sigma_X) \stackrel{M}{\leftrightarrows} (Y, \Sigma_y)$ is an isomorphism in* **Stoch** *then M is deterministic.*

Remark 4.3. Probability measures μ_X on a measurable space X may be viewed as morphisms $\hat{\mu} : \mathbf{1} \nrightarrow X$ in **Stoch**, where $\mathbf{1}$ is any (necessarily equivalent) single object measure space. To see this, recognized that such a morphism is a function $* \times \sigma_X \stackrel{\hat{\mu}_X}{\nrightarrow} [0, 1]$ where $*$ is the unique object of the measurable space $\mathbf{1}$. This extra $*$ provides no additional information to the function, so we can reduce it to $\sigma_X \stackrel{\hat{\mu}_X}{\nrightarrow} [0, 1]$ which is just a measure on X.

We can also convert measures on X into arrows *from* X to the object $\mathbf{2}$ with objects (WOLOG) $\{0, 1\}$. Evidently any $X \stackrel{M}{\nrightarrow} \mathbf{2}$ is a measurable function in the left argument by definition. But if we are dealing with countably generated spaces,then the measures are generated by their values on singletons. Since $M(x, \cdot)$ is a probability measure, we have that $M(x, 0) + M(x, 1) = 1$; that is, x is either true or not, $\text{prob}(x) = 1 - \text{prob}(x)$. These properties are not enough to ensure that M builds a probability measure over X. We must explicitly construct $\overline{\mu_X} : X \nrightarrow \mathbf{2}$ so that $\overline{\mu_X}(x, 1) = \mu(\{x\})$ and hence $\int_X \overline{\mu_X}(x, 1) = 1$. Then $\overline{\mu_X}$ can be interpreted as a probability measure.

So for a measure μ on X we have a correspondence between morphisms $\hat{\mu} : \mathbf{1} \nrightarrow X$ and $\overline{\mu} : X \nrightarrow \mathbf{2}$. This gives us a choice of two embeddings. Using $\overline{\mu}$, composition proceeds in the normal order as in equation (2). However, the construction of $\overline{\mu}$ is messy, and moreover, $\hat{\mu}$ has the nice property that *every* morphisms $\mathbf{1} \nrightarrow X$ is a probability measure on X. So we will proceed with $\hat{\mu}$

Note however that the order of composition for $\hat{\mu}$ is reversed from that of μ. More generally, it can be seen that any morphisms $x \times \sigma_Y \to [0,1]$ is actually *contravariant* in its right argument. This is an important property, and fits with the generalization of profunctors ($\phi : A \nrightarrow B \equiv A \times B^{\text{op}} \to \mathcal{V}$ where A, B are both enriched over $\mathcal{V}$) to *proarrows*, which we shall see in the next section.

Based on this interpretation of measures in **Stoch**, we see that the condition (2) is just the requirement that for measurable function $f : X \to Y$, we have

$$\delta_f \circ \hat{\mu}_X \equiv \int_X \chi(f(x), \sigma_Y)\hat{\mu}_X(*, dx) \equiv \int_{x \in f^{-1}(\sigma_Y)} \hat{\mu}_X(*, \{x\}) \equiv \hat{\mu}_X(*, f^{-1}(\sigma_Y)) = \hat{\mu}_Y(*, \sigma_Y)$$

Which is ensured by X, Y having perfect measures. The intuition here is that if we were performing Bayesian inference, we could not choose our P_X, P_Y, and $P_{Y|X}$ freely, at any single point in time we'd have $P_Y = P_{Y|X} * P_X$ or our probabilities would be inconsistent.

So we have that

Proposition 4.4. *There is an equivalence of categories between the subcategory of* **Stoch***, restricted to countably generated measurable spaces, with deterministic non-terminal morphisms, and the category* **Rand**.

From here on, we will restrict ourselves to the subcategories with countably generated measurable spaces, referred to as **CGStoch**, **CGMeas** and **CGRand** as the subcategories of **Stoch**,$\mathcal{M}$eas, and **Rand** respectively, following [11].

5 The Framed 2-Category of Probabilistic Mappings

> It is important to distinguish the difficulty of describing and learning a piece of notation from the difficulty of mastering its implications. [...] Indeed, the very suggestiveness of a notation may make it seem harder to learn because of the many properties it suggests for exploration.
>
> Kenneth E. Iverson

CGRand may be though of as "**CGMeas** lifted into **CGStoch** plus basic probabilities". Indeed the point measure forms an adjoint "free construction" functor lifting **CGMeas** into **CGStoch** as shown in [**?**]. The meaning of the "plus basic probabilities" will be made clear in this chapter. First, we will take time to expand on the relationship between **CGMeas** and **CGStoch**. We will see that the characteristic function forms a bridge between measurable functions and stochastic maps. Precisely, **CGMeas** may be merged with **CGStoch** in a particular way to form a category with two distinct types of morphisms between objects, called a *framed category*, which we discuss in the next section.

5.1 Framed Categories

Framed 2-Categories centralize the notion of a 2-category with two distinct types of interacting 1-cells.For a rich exposition of framed categories and related theory, the author recommends [16].

We start by defining a category with two types of 1-cells (a double category). Next we introduce framed categories properly by giving additional structure allowing one type of 1-cell to be "pulled back" or "base changed" along another. We show that the category of probabilistic maps is an example of a framed category, and develop useful constructions in this category through the machinery of framed categories.

Definition 5.1 (Double Category)**.** A **double category** $\mathbb{D}$ consists of a category $\mathbb{D}_0$ of objects with "vertical arrows" and a category $\mathbb{D}_1$ of "horizontal arrows" with natural transformations related by the following functors

$$\begin{aligned} U &: \mathbb{D}_0 \to \mathbb{D}_1 \text{ (Identity)} \\ L, R &: \mathbb{D}_0 \to \mathbb{D}_1 \text{ (Source,Target)} \\ \odot &: \mathbb{D}_1 \times_{\mathbb{D}_0} \mathbb{D}_1 \to \mathbb{D}_1 \text{ (Composition)} \end{aligned}$$

Where $\mathbb{D}_1 \times_{\mathbb{D}_0} \mathbb{D}_1$ is the pullback $\mathbb{D}_1 \xrightarrow{R} \mathbb{D}_0 \xleftarrow{L} \mathbb{D}_1$. The pairs of horizontal arrows with the source of the second matching the target of the first, and such that:

$$\begin{aligned} L(U_A) = R(U_A) &= A \\ L(M \odot N) &= L(M) \\ R(M \odot N) &= R(N) \end{aligned}$$

and equipped with natural isomorphisms (iso arrows in $\mathbb{D}_1$)

$$\begin{aligned} \mathfrak{a} &: (M \odot N) \odot P \xrightarrow{\cong} M \odot (N \odot P) \text{ (Associator)} \\ \mathfrak{l} &: U_A \odot M \xrightarrow{\cong} M \text{ (Left Unitor)} \\ \mathfrak{r} &: M \odot U_B \xrightarrow{\cong} M \text{ (Right Unitor)} \end{aligned}$$

such that L and R of the above natural isomorphisms are identities in D_0

Intuitively, we may think of a double category as a bicategory with two different types of 1-cells. We do this by considering the objects of $\mathbb{D}_0$ as 0-cells of $\mathbb{D}$, the arrows of $\mathbb{D}_0$ as "vertical 1-cells" of $\mathbb{D}$, and the objects of $\mathbb{D}_1$ as "horizontal 1-cells" of $\mathbb{D}$. As suggested by their names, we draw vertical 1-cells vertically and horizontal 1-cells horizontally with a slash to avoid confusion.

The arrows of $\mathbb{D}_1$ as the 2-cells of $\mathbb{D}$. 2-cells α may be visualized by the square formed by the image of L and R on α as in figure 2. Note that the horizontal composition operation $\odot$ is in the reverse order of typical arrow composition. We choose this notation following [16] because it simplifies many expositions, allowing $M \odot N$ to correspond to left-to-right juxtaposition of arrows M, N in a diagram. Another way to think about a double category

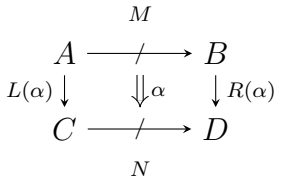


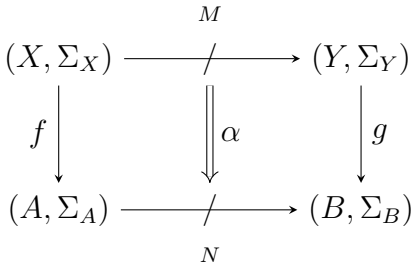
Figure 2: Relation between 2-cells and vertical/horizontal 1-cells in a double category

$\mathbb{D}$ is as a tuple consisting of a category $\mathbb{D}_0$ of objects and vertical 1-cells, (as above) and a bicategory $\mathcal{D}$ with the objects, horizontal 1-cells and 2-cells and a compatibility between the two types of 1-cells. This is often the most natural way to form a double category, whenever a set may be made into a (bi)category in two compatible ways.

Example 5.1. The motivating example of a double category is the category **Prof** with 0-cells (small) categories, vertical 1-cells functors, horizontal 1-cells profunctors, and 2-cells natural transformations between profunctors. Equivalently, **Prof** may be considered as the augmentation of the bicategory of profunctors with the usual **Cat** arrows.

Example 5.2. Perhaps of more interest to us is the double category $\mathbb{M}$**od** with 0-cells as rings, vertical 1-cells as ring homomorphisms, horizontal 1-cells $M : A \nrightarrow B$ as an (A, B)-bimodule, and 2-cells as $(L(\alpha), R(\alpha))$-bilinear maps (module homomorphisms). An (f, g)-bilinear map $M \to N$ is an abelian group homomorphism $\alpha : M \to N$ such that $\alpha(amb) = f(a)\alpha(m)g(b)$. Composition of modules is given by the tensor product of modules $M \odot N = M \otimes_B N$. It is then easy to check that the coherence conditions hold.

Definition 5.2 ($\mathcal{P}$rob)**.** Let us define the (conjectural) double category $\mathcal{P}$rob of probabilistic mappings.
The 0-cells of $\mathcal{P}$rob are measurable spaces (X, Σ_X).
The vertical 1-cells of $\mathcal{P}$rob are measurable functions $f : (X, \Sigma_X) \to (Y, \Sigma_Y)$.
The horizontal 1-cells of $\mathcal{P}$rob are stochastic maps $M : X \times \Sigma_Y \to [0, 1]$.
The 2-cells $\alpha : (M : X \nrightarrow Y) \Rightarrow (N : A \nrightarrow B)$ are triples (f, α, g) of measurable functions where $f : X \to A$, $g : Y \to B$ and group endomorphism $\alpha(M(\cdot, \cdot)) \mapsto N(f(\cdot), g(\cdot))$.
Notice that this definition of 2-cells is contravariant, in that an $\alpha : M \Rightarrow N$ actually lets us construct an (endomorphism of) M from an N by retraction along (f, g) as depicted below.



Remark 5.1. . More generally, if we don't require that our measures be probability measures, the 2-cells consist of triples (f, α, g) where α maps $M(x, \sigma_Y) \mapsto M_\alpha(f(x), g(\sigma_Y))$ and $\alpha(M)$ is a continuous group endomorphism of M on $\mathbb{R}^+$.The only such endomorphisms

on $\mathbb{R}^+$ are those of the form $f(x) = cx$ by the well known solution to the Cauchy equation $f(x+y) = f(x) + f(y)$ (equivalent to an endomorphism on $\mathbb{R}^+$. Since we require all probability measures to sum to 1, in that case the c is uniquely fixed as determined by the normalization requirements of α given (f, g) and the only allowable endomorphisms are inclusions and projections with normalization.

Theorem 5.1. $\mathcal{P}$rob *forms a double category with suitable structure maps* U, L, R

Proof: The 0-cells and vertical 1-cells of $\mathcal{P}$rob may be considered as the category **CGMeas** previously defined. Likewise, the horizontal 1-cells and 2-cells may be considered as the arrow category **CGStoch** $\uparrow$. Giving $\mathbb{D}_0 = \mathbf{CGMeas}, \mathbb{D}_1 = \mathbf{CGStoch} \uparrow$

We further equip $\mathcal{P}$rob with the necessary functors to define a double category. A functor U : **CGMeas** $\to$ **CGStoch** $\uparrow$ sending objects the the characteristic function $U((X, \Sigma_X))(x, \sigma_X) = \chi_{\sigma_X}(x)$ [4] and sending measurable functions $f : A \to B$ to the 2-cell ${}_fU_f : U(A) \to U(B)$ defined by $U(f)(x, \sigma_X) = (\chi_{f(\sigma_X)} \circ f)$. In general for a stochastic maps $M : A \to B$ and measurable functions $f : X \to A, g : Y \to B$, we define

$$ {}_fM_g(x, \sigma_Y) \stackrel{\text{def}}{\equiv} M(f(x), g(\sigma_Y)) \tag{6} $$

This notation is very convenient, and we will find later that it exhibits many important properties in this category.

Functors L, R : **CGStoch** $\uparrow \to$ **CGMeas** sending stochastic maps to their source and target respectively and 2-cells (f, α, g) to f, g respectively.

Composition of stochastic maps is given as above, for stochastic maps $M(x, \sigma_Y), N(y, \sigma_Z)$, the composition is given by

$$ (M \odot N)(x, \sigma_Z) \stackrel{\text{def}}{\equiv} \int_Y N(y, \sigma_Z) M(x, dy) \tag{7} $$

Note that in this case the structure maps $\mathfrak{a}, \mathfrak{l}, \mathfrak{r}$ are the identity 2-cell. So we call this a *strict* double category. Observe:

$$ (M \odot U_Y)(x, \sigma_Y) \equiv \int_Y \chi(y, \sigma_Y) M(x, dy) = \sum_{y \in Y} \begin{cases} M(x, dy), & \text{if } y \in \sigma_Y \\ 0, & \text{otherwise} \end{cases} $$

But this is just $M(x, \sigma_Y)$ as seen from example 4.1 that the value of a measure is constructed pointwise over the atoms around a point. Similarly for $(U_X \odot M)(x, \sigma_Y) \equiv \int_Y M(y, \sigma_Y) \chi(x, dy)$. Associativity is given by the associativity of integration.

Example 5.3. As a practical example of a diagram in the category $\mathcal{P}$rob, consider finite measure spaces $A = \{a, b, c\}$ and $X = \{x, y\}$ imbued with the discrete σ-algebra. and measures generated represented by the measures $\mu_A(a) = 0.1, \mu_A(b) = 0.2, \mu_A(c) = 0.7, \mu_X(x) = 0.25, \mu_X(y) = 0.75$, or alternatively as vectors

$$ \mu_A = \begin{bmatrix} 0.1 \\ 0.2 \\ 0.7 \end{bmatrix} \qquad \mu_X = \begin{bmatrix} 0.25 \\ 0.74 \end{bmatrix} $$

[4] Note that the matrix representation of $\chi_{\sigma_X}(x)$ for finite discrete Σ_X is exactly the identity matrix of dimension $|X|$

Then this subcategory of $\mathcal{P}$rob, we have objects $A, X, *, [0,1]$ the underlying measurable spaces, $*$ the terminal measurable space with a single object, and $[0,1]$ the measurable space with two objects. The unconditional probabilities of μ are represented by horizontal 1-cells $P_A : * \nrightarrow A$ and $P_B : * \nrightarrow B$. Equivalently, these are functions $* \times A \to [0,1] \cong A \to [0,1]$ and $* \times X \to [0,1] \cong X \to [0,1]$ where the latter two are probability measures (our original μ). Now, assume that we have additionally two more variables $B = \{1,2,3\}, Y = \{1', 2'\}$, and conditional probabilities given by

$$P(B|A) = \begin{bmatrix} a1 & b1 & c1 \\ a2 & b2 & c2 \\ a3 & b3 & c3 \end{bmatrix} \qquad P(Y|X) = \begin{bmatrix} x1' & y1' \\ x2' & y2' \end{bmatrix}$$

Then we have horizontal 1-cells $P_{B|A} : A \nrightarrow B$ and $P_{Y|X} : X \nrightarrow Y$ with $P_{B|A}(a, \sigma_B), P_{Y|X}(x, \sigma_Y)$ given by (sums over the elements of σ) the elements of their corresponding matrix. So for example, the joint probability $A \wedge B$ composition

$$P_A \odot P_{B|A} = \begin{bmatrix} a1 & b1 & c1 \\ a2 & b2 & c2 \\ a3 & b3 & c3 \end{bmatrix} \begin{bmatrix} 0.1 \\ 0.2 \\ 0.7 \end{bmatrix}$$

Suppose further that we have measurable functions

$$f : A \to X = \begin{cases} a \mapsto & x \\ b \mapsto & x \\ c \mapsto & y \end{cases} \qquad g : B \to Y = \begin{cases} 1 \mapsto & 1' \\ 2 \mapsto & 2' \\ 3 \mapsto & 1' \end{cases}$$

There is a convenient factorization ${}_fM_g = {}_f(U \odot M \odot U)_g = {}_fU \odot M \odot U_g$ of which we will get into more detail in the next section. But for now consider for our f, g, the matrix representations

$$[{}_fU] = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad [U_g] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Then we have a 2-cell $\alpha : P_{B|A} \Rightarrow P_{Y|X}$ with a new horizontal 1-cell, the normalization of the following matrix so that all columns sum to 1

$$[{}_f(P_{Y|X})_g] = \begin{bmatrix} x1' & x1' & y1' \\ x2' & x2' & y2' \\ x1' & x1' & y1' \end{bmatrix} = [U_g][P_{B|A}][{}_fU]$$

The homomorphism maps $P_{B|A}(a, \sigma_B) \mapsto \overline{P_{Y|X}(f(a), g(\sigma_B))}$ where $\overline{(-)}$ is the normalization $\overline{M(a, \sigma_B)} = \frac{M(a,\sigma_B)}{\int_B M(a,db)} = \frac{M(a,\sigma_B)}{M(a,B)}$

$\mathcal{P}$rob is essentially **CGStoch** lifted into a double category by including the arrows of **CGMeas**. By lifting **CGStoch**, we make clear the relation between measurable functions and measures, solidifying the fact that a measure space is defined *over* a collection of measurable spaces, in the same way that a vector space is defined over a field. Indeed, vector

spaces over fields (or more generally, bimodules over rings) form a natural double category as seen in example 5.2. But bimodules also exhibit a more interesting property that modules can be extended along ring homomorphisms. This extension or "base change" operation of horizontal 1-cells along vertical 1-cells can be generalized to a Framed Category, and we will see that $\mathcal{P}$rob also exhibits this property. There are many equivalent definitions of a framed 2-category. Along with the traditional definition (a double category where the L, R functors are fibrations), it has been shown equivalent to a proarrow equipment and a double category with a connection pair for every vertical 1-cell. The machinery of the classic definition is perhaps to heavyweight for our discussion here, so we give only the latter two definitions. For completeness we show that $\mathcal{P}$rob satisfies both of these conditions, though they each imply each other. Nevertheless, the fibrational definition of framed categories is highly elucidating, so the interested reader is referred to [16] for a deeper exploration. The first equivalent definition (See [16]. Thm 4.1) is given below.

Definition 5.3 (Framed 2-Category)**.** A double category $\mathbb{D}$ is a *Framed 2-Category* if additionally For every vertical 1-cell $f : A \to B$, we have the pair of horizontal 1-cells ${}_fB : A \nrightarrow B$ and $B_f : B \nrightarrow A$ with 2-cells

$$\begin{array}{ccc} & {}_fB & \\ A & \nrightarrow & B \\ {\scriptstyle f}\downarrow & \Downarrow \alpha & \| {\scriptstyle \mathrm{Id}} \\ B & \nrightarrow & B \\ & U_B & \end{array} \qquad \begin{array}{ccc} & B_f & \\ B & \nrightarrow & A \\ {\scriptstyle \mathrm{Id}}\| & \Downarrow \beta & \downarrow {\scriptstyle f} \\ B & \nrightarrow & B \\ & U_B & \end{array} \tag{8}$$

$$\begin{array}{ccc} & U_A & \\ A & \nrightarrow & A \\ {\scriptstyle f}\downarrow & \Downarrow \gamma & \| {\scriptstyle \mathrm{Id}} \\ B & \nrightarrow & A \\ & B_f & \end{array} \qquad \begin{array}{ccc} & U_A & \\ A & \nrightarrow & A \\ {\scriptstyle \mathrm{Id}}\| & \Downarrow \epsilon & \downarrow {\scriptstyle f} \\ A & \nrightarrow & B \\ & {}_fB & \end{array} \tag{9}$$

Such that for $\alpha, \beta, \gamma, \epsilon$ defined in equations (8) and (9), the following conditions hold.

$$\alpha\epsilon = U_f = \beta\gamma \tag{10}$$

$$\mathrm{Id}_{{}_fB} = (\epsilon \odot \alpha : U_A \odot {}_fB \to {}_fB \odot U_B) \tag{11}$$

$$\mathrm{Id}_{B_f} = (\beta \odot \gamma : B_f \odot U_A \to U_B \odot B_f) \tag{12}$$

5

[5]There is a misprint in [16] equation (4.4) where B_f is rendered as ${}_fB$. Our definition here is correct, as the types involved do not match up with ${}_fB$

Remark 5.2. Note the similarity between our B_f and the ${}_fM_g$ notation from equation (6). This similarity is intentional, as we will see them coincide shortly, but to avoid confusion, remember that if B is a 0-cell of a framed category then B_f is as defined here. If M is a horizontal 1-cell, then M_f is $M(\cdot, f(\cdot))$

5.1.1 Characterizing $\mathcal{P}$rob

Theorem 5.2 (Major Result)**.** *The double category* $\mathcal{P}$rob *is also a framed category with suitable* ${}_fB, B_f, \alpha, \beta, \gamma, \epsilon$

First introduce ${}_fB, B_f$ for our framed category

$$ {}_fB(x, \sigma_Y) = \chi(f(x), \sigma_Y) \qquad B_f(y, \sigma_X) = \chi(y, f(\sigma_X)) \tag{13}$$

And our $\alpha, \beta, \gamma, \epsilon$

$$\alpha : {}_fB(a, \sigma_B) \mapsto \chi(f(x), \sigma_B)$$
$$\beta : B_f(b, \sigma_A) \mapsto \chi(b, f(\sigma_A))$$
$$\gamma : \chi(a, \sigma_A) \mapsto B_f(f(a), \sigma_A)$$
$$\epsilon : \chi(a, \sigma_A) \mapsto {}_fB(a, f(\sigma_A))$$

For the α, β we have that the homomorphism component of α is just the identity, since already by definition ${}_fB(a, \sigma_B) = \chi(f(a), \sigma_B)$ and $B_f(b, \sigma_A) = \chi(b, f(\sigma_A))$.
For γ and ϵ, $B_f(f(a), \sigma_A)$ and ${}_fB(a, f(\sigma_A))$ both reduce to $\chi(f(a), f(\sigma_A)) = U(f) = {}_fU_f$ so the homomorphism components of γ and ϵ are just the restriction of $\chi = U_B$ to the image of f or could be taken as simply the identity.

To satisfy coherence law (10), see that

$$\alpha\epsilon(U_A)(a, \sigma_A) = \alpha(\chi)(f(a), f(\sigma_A)) = \chi(f(a), f(\sigma_A)) = U_f$$

$$U_f = \chi_(f(a), f(\sigma_A)) = \beta(\chi)(f(a), f(\sigma_A)) = \beta\gamma(U_A)(a, \sigma_A)$$

It is easy to see that coherence laws (11) and (12) hold because we have already seen in theorem 5.1 that $U_A \odot M = M \odot U_B = M$ for any $M : A \nrightarrow B$, so it must hold for specific instances. Particularly even if U has its domain or range restricted, if its restricted range is no larger than the domain of some M, then $U \odot M = M$, and similarly if its restricted domain is no smaller than the range of some N then $N \odot U = N$. Now observe

$$\epsilon(U_A) \odot \alpha({}_fB) = U(f) \odot {}_fB = {}_fB$$

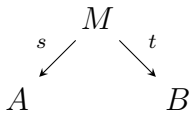
Because $U(f)$ is just U_B with its domain restricted to the image of f, and α is already the identity. A mirror argument may be used for coherence law (12). $\square$

Now let us take this framed category structure and lift it up into **CGRand**

5.2 The Framed 2-Category CGRand

There is a rich theory of profunctors and related constructs as detailed in [17]. Interesting point is that every profunctor can be reformulated as a particular span

Definition 5.4. A span for objects A, B is a diagram of the form

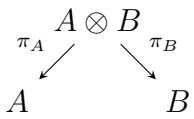


The names of the morphisms s, t have been chosen to suggest that M represents some "object of morphisms" and the morphisms s, t represent the source and target maps, but this is in no way a requirement of spans, which are much more general.

We often see that other "proarrows" have a similar formulation, such as bimodules [16]. Indeed, the spans in any category form "proarrow-like" morphisms, but it is not always meaningful to do so. Thus, it is interesting to see if there is any way we can reformulate our framed category $\mathcal{P}\mathrm{rob}$ as a category of spans. Here we do just that, generalizing and strengthening our model of Bayesian probability along the way.

One useful type of spans are *products* in a category.

Definition 5.5. A product $A \otimes B$ in a category is a span



such that every other span on A and B factor through π_A or π_B. $A \otimes B$ may be considered as pairs of elements of A, B if the category is suitably "setlike". The morphisms π_A, π_B are called the *projections* of the product. The factorization property says that the π are "universal" in the sense that they encode all the combined information about A and B

A particularly compelling view of this possibility is to recognize that in practice, we obtain marginal probabilities P_A and conditional probabilities $P_{B|A}$ from some measured joint distribution $P_{A \wedge B}$. The category **CGMeas** has a product object for every pair of objects, simply represented by the Cartesian product on the underlying set (This is exactly the monoidal tensor product structure on $\mathcal{M}\mathrm{eas}$). On the other hand, the category **CGStoch** does not have all products. Indeed, **CGStoch** has only *weak* products, given by the monoidal tensor product. These "product measures" fail to satisfy the uniqueness condition of a categorical product.

However, **CGStoch** still has product *measurable spaces* as objects, because we have an identity-on-objects functor $\delta : \mathbf{CGMeas} \to \mathbf{CGStoch}$, but these objects no longer have the universal properties of products in **CGStoch**. Notice that when we talk about a conditional probability $P_B|A$, we actually mean the probability of $A \wedge B$ given A, from which the probability of B can be readily recovered. This is apparent in the definition of conditional probability $P(B|A) \stackrel{\text{def}}{\equiv} \frac{P(A \wedge B)}{P(A)}$. This suggests that every morphism $A \nrightarrow B$ "factors through"

the joint object $A \wedge B$ in some sense. In fact, there are two types of morphisms at play here, the deterministic maps which do not factor (or factor trivially), and the non-deterministic maps defined through measures on the joint objects. We make this notion clear by elevating **CGRand** to a framed category.

Since probability measures may be regarded as morphisms from the terminal object in **CGStoch**, we can consider a reformulation of this category as the *under category* $\mathbf{1}/\,\mathbf{CGStoch}$ consisting of the morphisms out of the terminal object. We can formalize this as a double category as follows:

Definition 5.6. CGRand has

1. 0-cells as countably generated perfect probability measure spaces (X, μ_X) where X is a measurable space. The measure μ_X is regarded as a stochastic map $\mathbf{1} \nrightarrow X$ so that it can be composed with the 1-cells.

2. vertical 1-cells deterministic stochastic maps $\delta_f : A \nrightarrow B$ for a measurable function f so that $\mu_B = \delta_f \circ \mu_A$. Vertical 1-cells may be visualized as the commutative diagram

$$\begin{array}{ccc} A & \xrightarrow{\delta_f} & B \\ & {\scriptstyle \mu_A} \nwarrow \quad \nearrow {\scriptstyle \mu_B} & \\ & \mathbf{1} & \end{array}$$

3. horizontal 1-cells $M : A \rightsquigarrow B$ as spans of vertical 1-cells $(A, \mu_A) \overset{\mathcal{L}}{\nleftarrow} (F, \mu_F) \overset{\mathcal{R}}{\nrightarrow} (B, \mu_B)$. By the nature of vertical 1-cells, we have that $\mu_A = \mathcal{L} \circ \mu_F$ and $\mu_B = \mathcal{R} \circ \mu_F$. This property may be visualized by the commutative diagram

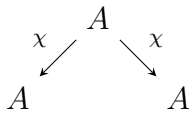
$$\begin{array}{ccccc} A & \xleftarrow{\mathcal{L}} & F & \xrightarrow{\mathcal{R}} & B \\ & {\scriptstyle \mu_A} \nwarrow & \uparrow {\scriptstyle \mu_F} & \nearrow {\scriptstyle \mu_B} & \\ & & \mathbf{1} & & \end{array}$$

4. 2-cells $\alpha : (F : A \rightsquigarrow B) \Rightarrow (G : C \rightsquigarrow D)$ as triples (l, α, r) of deterministic maps with $l : A \nrightarrow C, r : B \nrightarrow D, \alpha : F \nrightarrow G$ so that. In pictorial form, the following diagram commutes

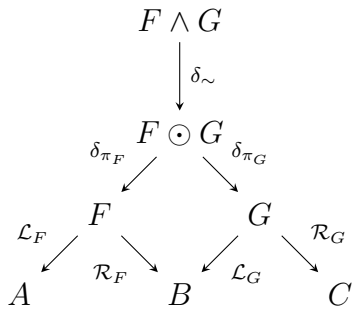
$$\begin{array}{ccccc} A & \xleftarrow{\mathcal{L}_F} & F & \xrightarrow{\mathcal{R}_F} & B \\ {\scriptstyle l}\downarrow & & \downarrow{\scriptstyle \alpha} & & \downarrow{\scriptstyle r} \\ C & \xleftarrow[\mathcal{L}_G]{} & G & \xrightarrow[\mathcal{R}_G]{} & D \end{array}$$

It is easy to check that **CGRand** with the following additional structure functors is a double category.

- The identity U as the span



- Source and target functors L, R mapping a span to its left and right targets respectively.
- Composition $F \odot G$ of spans $F : A \rightsquigarrow B$ and $G : B \rightsquigarrow C$ is given by the following pullback

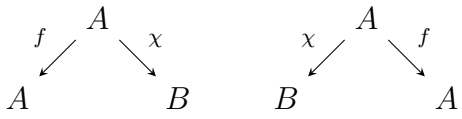


  Let $\mathcal{R}_F = \delta_f$ for some measurable f, and similarly $\mathcal{L}_G = {\delta_g}^6$, then the pullback is given by $(F \wedge_B G, \nu$ where $F \wedge_B G$ is the quotient subspace of the product measurable space $F \wedge G$ subject to the quotient $f(a) = g(b)$. The measure ν is simply the projection of the product measure, given by $\nu = \delta_\sim \circ \mu_{F \wedge G}$, where $\sim\colon F \wedge G \to F \odot G$ is the measurable function $(a, b) \mapsto (a, b)/f(a) = g(b)$ sending $F \wedge G$ to its quotient defined above.

Clearly the coherence laws for the L and R functors hold, as well as the left and right unitors. Associativity holds because we are computing strict pullbacks, so the commutivity requirement of the above diagram requires associative equality.

CGRand is also a framed category with:

- For each deterministic stochastic map $f : A \twoheadrightarrow B$, "base change" horizontal 1-cells $B_f : B \rightsquigarrow A$ and ${}_fB : A \rightsquigarrow B$ with B_f and ${}_fB$ as



  respectively

- The 2-cells $\alpha, \beta, \gamma, \epsilon$ are just the various liftings of f, $\alpha = (f, f, \mathrm{Id}), \beta = (\mathrm{Id}, f, f), \gamma = (f, \mathrm{Id}, \mathrm{Id}), \epsilon = (\mathrm{Id}, \mathrm{Id}, f)$ from which it is clear that the coherence laws hold strictly.

[6] We can do this because all deterministic stochastic maps on countably generated measurable spaces are induced by point measures of measurable functions

Remark 5.3. There are a particularly natural class of horizontal 1-cells $A \rightsquigarrow B$ of the form $(A, \mu_A) \overset{\delta_{\pi_A}}{\nleftarrow} (A \wedge B, \nu) \overset{\delta_{\pi_B}}{\nrightarrow} (B, \mu_B)$ where $A \wedge B$ is the categorical product of the measurable spaces A, B in the category $\mathcal{M}$eas. Such measure spaces are naturally imbued with deterministic maps ${}_{\pi_A}\chi : A \wedge B \nrightarrow A$ and ${}_{\pi_B}\chi : A \wedge B \nrightarrow B$ so one may take the 1-cell to be the span given below

$$\begin{array}{ccccc} & & A \wedge B & & \\ & {\scriptstyle {}_{\pi_A}\chi} \swarrow & & \searrow {\scriptstyle {}_{\pi_B}\chi} & \\ A & & & & B \end{array}$$

Such horizontal morphisms are entirely characterized by the measure ν. Note that every pair of 0-cells $(A, \mu_A), (B, \mu_B)$ have a canonical horizontal 1-cell given by $(A \wedge B, \mu_A * \mu_B)$ given by their product measures. In general, there are many probability measures on $A \wedge B$ that marginalize to μ_A and μ_B, so in practice we take these joint measures by experiment, and subsequently compute the marginals. We will call such spans *Primitive*.

Proposition 5.1. *Primitive 1-cells are universal in the sense that for every $N : C \rightsquigarrow D$, there is a primitive 1-cell $M : A \rightsquigarrow B$ and 2-cell $\alpha : M \Rightarrow N$.*

Proof: Since we have base-change morphisms, every $C \overset{\delta_f}{\nleftarrow} N \overset{\delta_g}{\nrightarrow} D$ factors into

$$\begin{array}{ccccc} C & & & & D \\ \downarrow{\scriptstyle \delta_f} & & & & \downarrow{\scriptstyle \delta_g} \\ N & \xleftarrow[\text{Id}]{} & N & \xrightarrow[\text{Id}]{} & N \end{array}$$

Since joint measurable spaces are true products in **CGMeas**, we can always construct the joint measure space $C \wedge N \wedge D$ so that it has Canonical projections

$$\begin{array}{ccccc} C & \xleftarrow{\delta_{\pi_C}} & C \wedge N \wedge D & \xrightarrow{\delta_{\pi_D}} & D \\ {\scriptstyle \delta_f}\downarrow & & \downarrow{\scriptstyle \delta_{\pi_N}} & & \downarrow{\scriptstyle \delta_g} \\ N & \xleftarrow[\text{Id}]{} & N & \xrightarrow[\text{Id}]{} & N \end{array}$$

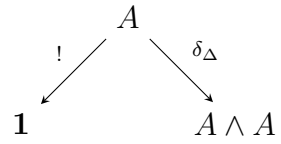
But $C \wedge N \wedge D$ is also a product of $(C \wedge D) \wedge N$, so we project onto $C \wedge D$ to get a primitive measure space $(C \wedge D, \pi_{C \wedge D}(\mu_{C \wedge N \wedge D}))$ as desired.

6 Conclusion and Further Work

> Whatever evidence an act might provide
> on what could have caused the act
> should never be used to help one decide
> on whether to choose that same act
>
> Judea Pearl

We have extended previous work on the relation between **Stoch** and $\mathcal{M}$eas, by showing that they also satisfy a framed category relation. This characterizes conditional probabilities as the "hom sets" of measurable functions, enriched over the category $[0, 1]$. We further this relation to reformulate conditional probabilities as restrictions on joint measures. In this paradigm, conditionals can be seen as "collections of deterministic maps", parameterized by the observed joint measure probabilities. Following the language of Fong [11], this category **CGRand** is the collection of **Stoch**-causal models in $\mathcal{M}$eas, mapped back into **Stoch** by the point measure functor δ.

There is also work to be done integrating this viewpoint with the theory of categorical matrix mechanics [18], detailed in the appendix. Particularly, as a corollary of proposition 5.1, **CGRand** is clearly a closed category, where each map $A \rightsquigarrow B$ is characterized by some joint measure space $(A \wedge B, \mu)$. It is a dagger category with the dagger of $(A \overset{\delta_f}{\leftarrow\!\!\!\prec} M \overset{\delta_g}{\nrightarrow} B)^\dagger$ as $(B \overset{\delta_g}{\leftarrow\!\!\!\prec} M \overset{\delta_f}{\nrightarrow} A)^\dagger$, and dagger compact with $A = A^*$ and $\eta = \epsilon^\dagger =$



Where ! is the unique terminal map, and Δ is the diagonal map $A \mapsto (A, A)$, sending A to the perfectly entangled pair of A.

The matrix mechanics interpretation gives a particularly compelling way to compute the marginals on a joint distribution. In this formalism, the marginals are the joint distribution composed with the frobenius counit over each marginalized variable. Viewed in this way, the belief propagation algorithm [19], or more generally the [20] on a tree shaped probabilistic graphical model is equivalent to a choice of ordering of the (commutative) counit compositions over a joint distribution. In the case of cyclic models, the marginals can be computed exactly using the same counit formulation, but "entangled" variables cannot be decomposed and computed in isolation, giving exponential asymptotic complexity in the cycle size. In this case, the approximate loopy belief propagation algorithm [21] corresponds to a truncation of the trace on this model. **CGRand** is a dagger compact category and so admits a unique canonical trace [22]. Further investigation determining if these constructions are coherent promise to be fruitful.

7 Appendices

7.1 Appendix A: Additional Category Theory

7.1.1 Proarrows

Perhaps a more natural way to think about a framed category is as a 2-category with a proarrow equipment. The two structures were developed separately for different purposes but shown to be equivalent [16]

Definition 7.1 (Proarrow Equipment)**.** A *Proarrow Equipment*, or *2-Category Equipped with Proarrows* is a pair of 2-categories K, M and a 2-functor $P : K$ so that

1. P is the identity on objects
2. For every 1-cell $f : A \to B$ in K, $P_f \overset{\text{def}}{\equiv} P(f)$ has a right adjoint ${}_fP$ meaning that there exits a 1-cell ${}_fP : P_f(B) \to P_f(B)$ in M and 2-cells
   - $\eta : 1_{P(B)} \Rightarrow {}_fPP_f$
   - $\epsilon : P_f{}_fP \Rightarrow 1_{P(A)}$
3. P is locally full and faithful. Meaning that for any two objects $x, y \in A$, the hom sets are isomorphic $A(x, y) \cong B(F(x), F(y))$

The intuition behind proarrow equipments is as a generalization of the framed category of profunctors. Indeed, when studying *internal category theory*, considering every object of a category itself as some category, it is seen that most horizontal 1-cells in a framed category are equivalent to enriched profunctors between some categories.

Proposition 7.1 (Equivalent Formulation)**.** *The categories* **CGMeas** *(considered as a 2-category with trivial 2-cells) and* **CGStoch** *form a proarrow equipment with* $P_f(x, \sigma_Y) = \chi(f(x), \sigma_Y)$ *and* ${}_fP(y, \sigma_X) = \int_{x \in f^{-1}(y)} \chi(x, \sigma_X)$*. This proarrow equipment is equivalent to* $\mathcal{P}$rob

Definition 7.2 (Weak Monoidal Category)**.** The similarity between the above two examples $\times$ and $\otimes$ begs a generalization. *Monoidal Category* is a category C together with

- A bifunctor $\otimes : C \times C \to C$ called the *tensor product*
- An identity object I
- 3 natural isomorphisms assuring that $\otimes$ satisfies the monoid laws (up to natural isomorphism)
  - **Identity**: Right and left identity maps $\rho_A : A \otimes I \cong A, \lambda_A : I \otimes A \to A$
  - **Associativity**: An *associator* isomorphism with components $\alpha_{A,B,C} : (A \otimes B) \otimes C \cong A \otimes (B \otimes C)$

- The construction must satisfy the coherence conditions (commutative diagrams)

$$\begin{array}{ccccc}
((A \otimes B) \otimes C) \otimes D & \xrightarrow{\alpha_{A,B,C} \otimes \mathrm{Id}_D} & (A \otimes (B \otimes C)) \otimes D & \xrightarrow{\alpha_{A,B\otimes C,D}} & A \otimes ((B \otimes C) \otimes D) \\
\downarrow{\scriptstyle \alpha_{A\otimes B,C,D}} & & & & \downarrow{\scriptstyle \mathrm{Id}_A \otimes \alpha_{B,C,D}} \\
(A \otimes B) \otimes (C \otimes D) & & \xrightarrow{\alpha_{A,B,C\otimes D}} & & A \otimes (B \otimes (C \otimes D))
\end{array}$$

$$\begin{array}{ccccc}
 & & A \otimes B & & \\
 & {\scriptstyle \rho_A \otimes \mathrm{Id}_B} \nearrow & & \nwarrow {\scriptstyle \mathrm{Id}_A \otimes \lambda_B} & \\
(A \otimes I) \otimes B) & & \xrightarrow{\alpha_{A,I,B}} & & A \otimes (I \otimes B))
\end{array}$$

Definition 7.3 (Adjunction)**.** Central to category theory is the notion of *Adjoint Functors*. Adjunctions can be thought of as the "best approximation" to a pair of inverse functors. Formally, an *adjunction* from C to D is a pair of functors $F : C \to D, G : D \to C$ and a family of natural isomorphisms ϕ so that for objects $c \in C$ and $d \in D$,

$$C(Fc, d) \overset{\phi_{c,d}}{\cong} D(c, Gd)$$

We denote this connection $F \dashv G$ and say that F is *left adjoint* to G, or that G is *right adjoint* to F. Equivalently, we have an adjunction $F \dashv G$ if we have natural transformations $\eta : \mathrm{Id} \Rightarrow FG$ and $\epsilon : GF \Rightarrow \mathrm{Id}$ called the *unit* and *counit* of the adjunction respectively.

Often adjunction pairs take the form of a forgetful functor and free construction. For example, the forgetful functor mapping a group to its underlying set is right adjoint to the functor mapping a set to the free group generated by the elements of the set. While this intuition is useful, it does not always hold in general. A more general tool may be to consider that the left adjoint *creates* structure while the right adjoint *forgets* structure, though this too may not help in all cases.

Definition 7.4 (Closed Monoidal Category)**.** A *Closed Monoidal Category* is a monoidal category such that for each $b \in C$, the functor $(- \otimes b) : C \to C$ mapping each object of C to its monoidal product with b, has a right adjoint $b \multimap - : C \to C$ so that

$$C(a \otimes b, c) \cong C(a, b \Rightarrow c)$$

. The object $b \multimap c$ is call the *internal hom* object, and "acts like" the set of morphisms from b to c. Equivalently, we may simply say that a monoidal category C is closed-monoidal if it can be enriched over itself, so that $\cdot \multimap \cdot : C^{\mathrm{op}} \times C \to C$ is exactly the enriched hom functor on C. Note that in general, there may be non monoidal closed categories with internal hom objects satisfying some additional properties. However, the definition of a closed monoidal category is more natural, with all relevant properties arising from the adjunction. Further, all relevant closed categories discussed here are also monoidal, so we omit the more general definition.

Definition 7.5 (Trace)**.** [23] A *Trace* in a (strict) symmetric monoidal category $\mathcal{C}$, is an operator

$$Tr^X_{A,B} : \mathcal{C}(A \otimes X, B \otimes X) \to \mathcal{C}(A, B) \tag{14}$$

For objects $A, B, X \in \mathcal{C}$, with the conditions

$$Tr^X_{A',B'}((k \otimes \mathrm{Id} \circ f \circ (h \otimes \mathrm{Id})) = k \circ Tr^X_{A,B}(f) \circ h \qquad \text{(Tightening)}$$

$$Tr^X_{C\otimes A, C\otimes B}(\mathrm{Id} \otimes f) = id \otimes Tr^X_{A,B}(f) \qquad \text{(Superposing)}$$

Superposing says that the trace doesn't interact with tensored identities. Tightening goes further to say that the trace passes through composition on tensored identities. The objects we take our trace over do matter and cannot be omitted in some cases. For example, the linear trace (below) would not make sense if the superposition constraint applied for all objects. Note that the typical Yanking and Exchange laws are trivial because we are dealing only with strict symmetric categories. For a more general definition, see [22]

The trace operation can be thought of as "tying the knot" on the extra X, feeding its output back to input. Indeed, in a computer science setting, traces correspond to fixed point operations. The typical trace of a linear transformation from linear algebra is a special case of this trace (Tr_L may be recovered by taking $Tr_L(f) = Tr(\mathrm{Id} \otimes f)$).

Remark 7.1. If a (strict) symmetric monoidal category is also compact closed, then it has the unique canonical trace

$$Tr^X_{A,B}(f) = (\mathrm{Id}_B \otimes \eta_X) \circ (f \otimes \mathrm{Id}_{X^*}) \circ (\mathrm{Id}_A \otimes \eta_X)$$

This formula has deep implications that we will only mention in passing. In computer science applications, the ϵ, η represent turning an output into an input, giving the most natural way to think about a trace as closing a recursive loop. In quantum physics applications, the ϵ, η represent dual particle annihilation and creation operators; in the context of a trace, they could be thought of as reversing the direction of time, turning a particle into its antiparticle. [1, 23]

Definition 7.6 ((Co)Monoid Object)**.** A (strict) monoid object in a monoidal category $\mathcal{C}$ (with monoidal unit I) is an object $A \in \mathcal{C}$ together with morphisms $\mathbf{m} : A \otimes A \to A$ and $\mathbf{u} : I \to A$ called the multiplication and unit respectively. so that the multiplication is associative ($\mathbf{u_A} \circ \mathbf{u_A} \otimes \mathrm{Id}_A = \mathbf{u_A} \circ \mathrm{Id}_A \otimes \mathbf{u_A}$) and the unit interacts well with the monoidal unit ($\mathbf{u_A} \circ \mathbf{m_A} \otimes \mathrm{Id}_A = \mathbf{u_A} \circ \mathrm{Id}_A \otimes \mathbf{m_A}$). A comonoid is the dual construction, with comultiplication $\mathbf{w} : A \to A \otimes A$ and counit $\mathbf{n} : A \to I$ satisfying the dual coherence conditions to the monoid.

Example 7.1. Let Δ be the diagonal measurable function $A \to A \otimes A$ sending $a \mapsto (a, a)$. Δ. In the category **CGMeas** every object is a monoid object, by taking $\mathbf{m}$ as the deterministic function χ_Δ so that $\mathbf{m}((a, a'), \sigma_A) = \begin{cases} 1, & \text{if } (a, a') \in \Delta(\sigma_A) \\ 0, & \text{otherwise} \end{cases}$ and $\mathbf{u}(*, \sigma_A) = \frac{|\sigma_A|}{|A|}$ representing the uniform distribution over A. Recalling that an "element" of an object A in $\mathcal{C}$ is simply

a morphism from the terminal object to A, then ${}_{\Delta}\chi$ can be interpreted as the measure that is 1 on objects $(a, a'), a = a'$ and 0 for other objects, as depicted below.

$$\mathbf{1} \xrightarrow{(a,a)} A \otimes A \underset{\pi_{A_2}}{\overset{\pi_{A_1}}{\rightrightarrows}} A$$

CGMeas is also a comonoid, with $\mathbf{w} = {}_{\Delta}\chi$, sending elements a to the "entangled state" where both copies of A always take the same value. Similarly, $\mathbf{n}$ is the stochastic map induced by the terminal (in **CGMeas**) map $a \mapsto *$. So $\mathbf{n}(a, \sigma_*) = \chi(*, \sigma_*)$

Definition 7.7 (Dagger Category)**.** A *Dagger Category* is a category $\mathcal{C}$ together with a functor $\dagger : \mathcal{C}^{\text{op}} \to \mathcal{C}$ that is the identity on objects. For a morphism $f : A \to B$ in $\mathcal{C}$, we have $f^\dagger \overset{\text{def}}{\equiv} \dagger(f) : B \to A$. In this case, the functor laws mean that we have

1. $\text{Id}^\dagger = \text{Id}$
2. $(g \circ f)^\dagger = f^\dagger \circ g^\dagger$
3. $f^{\dagger\dagger} = f$

$f^\dagger$ is typically called the *adjoint* morphism of f, by analogy the the motivating example of Hilbert spaces.

Remark 7.2. It is trivial to check that the requirement for dual pairs ${}_fB, B_f$ in a framed category are exactly the requirements for a dagger category. That is, for every framed category $\mathbb{D}$, the horizontal 1-cells induced by the vertical 1-cells have a dagger structure. Particularly, since **CGStoch** forms a framed category with vertical 1-cells measurable functions and horizontal 1-cells stochastic maps, we have that:

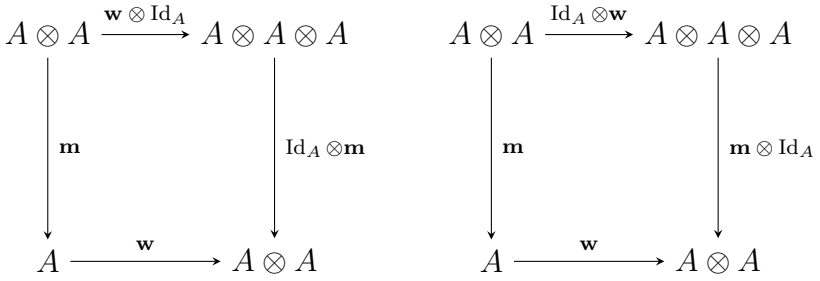
Proposition 7.2. *The subcategory of* **CGStoch** *restricted to deterministic stochastic maps has a dagger structure on all horizontal 1-cells.*

Proof: [3] show that every deterministic stochastic map with countably generated codomain is induced by the point measure ${}_f\chi$ for some measurable function f. Since ${}_f\chi : A \nrightarrow B$ has a dual pair $\chi_f : B \nrightarrow A$ as proved in theorem 5.2, we have an involutive mapping ${}_f\chi \leftrightarrow \chi_f$ and ${}_{\text{Id}}\chi = \chi_{\text{Id}} = \chi$, so requirements (1) and (3) for a dagger structure are obvious. The condition (2) is guaranteed since stochastic maps $A \nrightarrow B$ (which are really maps $A \times B^{\text{op}} \to \mathbb{R}$) are contravariant in their right argument.

Remark 7.3. Since a comonoid is the dual of a monoid (that is, a comonoid object in $\mathcal{C}$ is a monoid object in $\mathcal{C}^{\text{op}}$), we have that in a dagger category, monoid objects and comonoid objects are equivalent. Particularly, $w^\dagger = m$ and $n^\dagger = u$ (and vice versa).

Definition 7.8 (Dagger Frobenius Structure)**.** A *Dagger Frobenius Structure* in a dagger category $\mathcal{C}$ is a monoid/comonoid object that interacts nicely with the dagger structure.

Formally, in a dagger category $\mathcal{C}$, a dagger frobenius structure is a monoid object $(A, \mathbf{m}, \mathbf{u})$ (or equivalently a comonoid object $(A, \mathbf{w}, \mathbf{n})$ so that the following coherence diagrams commute.



Remark 7.4. A frobenius structure may equivalently be defined as a monoidal functor $\mathbf{1} \to \mathcal{C}$ where $\mathbf{1}$ is the category with one object and its identity morphism. In this light, the Frobenius algebras in **CGStoch** are a particular degenerate kind of stochastic causal models, using the terminology of [11]

References

[1] J. C. Baez and M. Stay. Physics, Topology, Logic and Computation: A Rosetta Stone. *ArXiv e-prints*, March 2009.

[2] Michéle Giry. A categorical approach to probability theory. In B. Banaschewski, editor, *Categorical Aspects of Topology and Analysis*, volume 915 of *Lecture Notes in Mathematics*, pages 68–85. Springer Berlin Heidelberg, 1982.

[3] J. Culbertson and K. Sturtz. A categorical foundation for Bayesian probability. *ArXiv e-prints*, May 2012.

[4] Andrey N. Kolmogorov. *Foundations of the Theory of Probability*. 1933.

[5] Judea Pearl. *Causality: Models, Reasoning, and Inference*. 2009.

[6] Brendan J Frey. Extending Factor Graphs so as to Unify Directed and Undirected Graphical Models. pages 257–264.

[7] Saunders Mac Lane. *Categories For the Working Mathematician*. Springer-Verlag, 1971.

[8] Ross Street. Encyclopaedia of mathematics, supplement ii. `http://maths.mq.edu.au/~street/Encyclopedia.pdf`, 2000.

[9] S. Lack. A 2-categories companion. *ArXiv Mathematics e-prints*, February 2007.

[10] T. Leinster. *Higher Operads, Higher Categories*. August 2004.

[11] B. Fong. Causal Theories: A Categorical Perspective on Bayesian Networks. *ArXiv e-prints*, January 2013.

[12] Robert B. Ash. *Real Analysis and Probability*. Academic Press, New York, 1972.

[13] R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2nd edition, August 2002.

[14] Arnold M. Faden. The existence of regular conditional probabilities: necessary and sufficient conditions. 13:288–298, 1985.

[15] Wacław Sierpiński. Sur les fonctions d'ensemble additives et continues. *Fundamenta Mathematicae*, 3(1):240–246, 1922.

[16] M. A. Shulman. Framed bicategories and monoidal fibrations. *ArXiv e-prints*, June 2007.

[17] B Jean. Distributors at Work. (June):1–28, 2000.

[18] B. Coecke and R. W. Spekkens. Picturing classical and quantum Bayesian inference. *ArXiv e-prints*, February 2011.

[19] Jonathan S Yedidia, William T Freeman, and Yair Weiss. Understanding Belief Propagation and its Generalizations. 2001.

[20] Frank R Kschischang, Brendan J Frey, and Hans-andrea Loeliger. Factor Graphs and the Sum-Product Algorithm. 47(2):498–519, 2001.

[21] Jonathan S Yedidia, William T Freeman, and Yair Weiss. Constructing Free Energy Approximations and Generalized Belief Propagation Algorithms. 2004.

[22] B Y Andre Joyal, Quebec Hc, and Ross Street. Traced monoidal categories. pages 447–468, 1996.

[23] S. Abramsky, R. Blute, and P. Panangaden. Nuclear and Trace Ideals in Tensored *-Categories. *ArXiv Mathematics e-prints*, May 1998.