Discussion of [**todo**: *need name for model*] Semantics and Consistency

Oliver Richardson oli@cs.cornell.edu

July 18, 2019

Recall that we have a model of nodes, chained together with Markov kernels / conditional probability distributions / stochastic matrices. Of course, there is already an enormous literature about diagrams which look somewhat like this, made up of nodes and conditional probability tables: Baysian Networks (BNs) and their many variants. The diagrams that we employ look very similar, but are intended for a different purpose, and hence have a different semantics. Whereas BN's are a factorization of a particular joint probability distribution and consistent by design, these models are merely constraints on the distribution, which might be so strong so as not to admit any joint distribution.

We will go into the difference more carefully in the next section, but first here are some reasons to prefer this as a way of modeling agents,

1. This representation more naturally matches what humans are aware of, encoding small locally consistent models rather than one giant probability distribution (see sections 1.3,1.4)
2. It is cleaner from a mathematical standpoint: utilities and probabilities are related, we get a notion of belief composition, and we can make use of both category and information theory (sections 1.7, 1.5, and 1.1).
3. This allows composition of arrows to be defined, and gives meanings to paths (section 1.5).
4. When bits of these diagrams are added and removed, the meaning and form of the rest of the diagram remains the same (section 1.6).
5. We can now represent inconsistency, which we will use to drive preference change. While we agree with the classical picture in that inconsistency is bad, now we can talk about it
6. Due to the compositionality, it is possible to add typing rules to dynamically change beliefs, knowledge representations, and so forth (section 1.1)
7. It is a strictly more general representation— we can easily convert BNs to these diagrams (section 1.2)

Recap: ⟨*model name*⟩ Definitions and Semantics

Definition 0.1. A ⟨*model name*⟩ model $(\mathcal{A}, \mathcal{L})$ is a collection of variables $\mathcal{A}$, attached to each of which is some preference information: (this could be an order, utility, pairwise utility matrix, a supremum function), plus a collection of probabilistic links $\mathcal{L}$ between some (but not necessarily all) pairs of them, where $L : A \to B \in \mathcal{L}$ is a (sub) Markov kernel $A \to \Delta(B \cup \{\bullet\})$[1] representing beliefs about how a setting of A to $a \neq \bullet$ will impact the value of B.

I have no idea with a "(sub) Markov kernel" is. Do we lose something if we just say "mapping"

In the finite case but we have $\mathbb{R}$-valued utilities

and I want to use the standard technical terminology so people working on related things like probabalistic programming can find it

Variables can be thought of equivalently as:

- random variables X which can take on values $\{x_i\}$ (or possibly none, which we denote $\bullet$)
- sets X with elements $\{x_i\}$
- partitions of the universe of outcomes into disjoint (but possibly not exhaustive) events $\{x_i\}$

You gain one audience, and lose 10 others. I think that's a bad tradeoff! I don't see why it's not a mapping even in the infinite case. If you really want to connect with the probabilistic programming folks, add a footnote.

[1]The additional "phantom element" $\bullet$ absorbs probability density that we don't want to equivocate on, allowing our model to capture partial families of conditional probabilities, by extension things like implication, and giving agents more tools to avoid inconsistency. This has the effect of making our links substochatic matrices/kernels rather than stochastic ones. However, if we restrict to beliefs which assign zero probability to $\bullet$, everything in the model works as before. See section 1.8 for details

In this document, we will focus on models where the preference information is encoded as a special utility domain U, with links from other variables. While there may be some propositions tying this case to utilities or preference orders, we will avoid talking explicitly about the more general setting of preference matrices and choice functions out for now.

1 Defense of ⟨*model name*⟩ as a Static Representation

1.1 The Possibility of Type Constructors

The classical picture also features a fixed set of variables. In addition to allowing new concepts to form for exogenous reasons, we would like to have inductive ways of introducing new ones logically, as combinations of the existing variables. Interpreting variables as types, whose possible assignments are terms, the syntax of which variables we can construct and what beliefs we have about the resulting picture is a type theory. There are some obvious ways one might want to combine existing domains; the one we are concerned with here is products.

oops... ?? events are to occur

In terms of beliefs, you might already have beliefs about how likely two A and B, then suddenly wonder about how they interact. For instance, you may already have beliefs about how likely you are to leave your keys in the ignition, and also how often your car is dead, and then wonder if there's a connection. In terms of preferences, you might think $a \succcurlyeq a'$ and $b \succcurlyeq b'$, but now wonder what you would do if you had to choose between ab' and $a'b$. In both cases, you are slightly enlarging your picture to consider the relevant features that a classical agent would already have.

For this reason, we introduce the ability to create a domain $A \times B$, if A and B are nodes in the picture, whose elements are the cartesian product of A and B. This is represented graphically as:

As I said, if you have an arrow B-> A, then you also have a natural arrow B->AxB

If you have both a setting of A and B (i.e. a selection from A×B), it determines in a natural way the values of A and B. The reverse is not true.

$$A \xleftarrow[\pi_A]{} A \times B \xrightarrow[\pi_B]{} B$$

Why do the arrows go in this direction? The other direction seems more natural to me (and is equivalent to B -> A

1.1.1 Products vs Unions

I don't really understand this section. To me, products and unions are doing very different things. They're both useful; it's not a question of either/or.

We could have achieved a similar thing by considering $\{A\}$ and $\{B\}$ as singleton sets of variables, and then adding $\{A, B\}$ to the picture. Here we would interpret as a set of random variable taking values in the range of the product of its elements possible values. The two accounts differ when there is an overlap between the sets. Should we be allowed to represent $A \times A$?

Taking unions does protect us from doing certain bad things: it would be a mistake to assign a distribution $\Pr(a, a') > 0$ for $a \neq a'$, for instance — but we're already in the business of allowing this kind of inconsistency — in the picture below all the marginals are fine as far as the syntax is concerned, but there is no way to assign a probability distribution on all of the nodes.

I wasn't talking about unions of domains, but set of variables!

$$\begin{tikzcd} & \boxed{1} \arrow[d, "p(A \times A)"] & \\ & \boxed{A \times A} \arrow[dl] \arrow[dr] & \\ \boxed{A} \arrow[rr, equal] & & \boxed{A} \end{tikzcd}$$

Maybe, but we were talking in meetings about how they might be equivalent, and how you thought re-framing in terms of sets of variables would be better. I'm arguing that products make more sense in the way that I'm using them.

It can also be very useful to keep extra copies of A around even if (part of) one already exists buried in another variable, because we can delay integration using this trick, as in example 1.7 for instance. The most compelling reason for me is that one might not know that you're looking at two copies of A; maybe they've been framed differently — but still you should be able to take a product and get two things, rather than a union, which immediately leaks the truth of their equivalence to an agent.

1.1.2 Additional Types

[**todo**: *Gestures towards embedding logics; equivalence between logic and type theory; in particular: conditionals, coproducts, negation, higher order nodes (beliefs about prefs, prefs about beliefs)*]

1.2 Converting BNs

The semantics of a Baysian network ensure that there is no inconsistency: the arrows into a node taken together collectively determine a single well-defined probability distribution. Formally they consist of a set of nodes $\mathcal{N}$, and for each $N \in \mathcal{N}$, a set of parents $\mathrm{Par}(N)$, and a conditional probability distribution $\Pr(N \mid \mathrm{Par}(N))$, which is a distribution over the values of N for each setting of every variable in $\mathrm{Par}(N)$. While each of our arrows can be interpreted by itself as a marginal, a collection of arrows into a single node must be taken together to have any meaning in a BN.

Sorry, bad shorthand.

The procedure for converting to a BN is simple: we simply take every node's incoming arrows, and insert the product of its parents as a node before it. With this procedure, if a node N has just one parent P, we replace $P \to N$ with $P \to N \;=\; N$, which is redundant so we don't draw this. If a node had zero parents (i.e., the BN just gives it a probability distribution not dependent on anything), we insert the product of zero things, i.e., the singleton node $1 = \{*\}$, a variable which only takes one value, and set $\Pr(N \mid *) = \Pr(N)$.

What is P-> N = N??

It's a path

$P \to N \to N$ where the $N \to N$ link is the identity on N.

I'm still confused …

This sound much more complicated than it is. Consider the example below, where the left is a BN, and the right is the corresponding ⟨*model name*⟩ .

I definitely want to talk about inconsistent pictures (which are not all the same) so "the set of distributions consistent w/ picture" is not enough information.

I defined "the center" to capture the generic case, fully explain connection to BNs, and get uniqueness for products, which gives us extra diagrammatic reasoning tools. For example, it lets us infer the dotted arrow here.

$$\begin{tikzcd} & A \arrow[dl] \arrow[dr] & \\ B \arrow[dr] & & C \arrow[dl] \\ & D & \end{tikzcd}$$

$$\begin{tikzcd} & 1 \arrow[d] & \\ & A \arrow[dl] \arrow[dr] \arrow[dd, dashed] & \\ B & & C \\ & B \times C \arrow[ul] \arrow[ur] \arrow[d] & \\ & D & \end{tikzcd}$$

We have effectively changed two things: first, visually encoded the probability distribution of A as the arrow $1 \to A$ (which we are now allowed to omit; sometimes you don't want priors on things, such as your own actions). Second, we have combined the two arrows $B \to D$ and $C \to D$ into a single one, $B \times C \to D$. Though certainly more verbose, this is arguably visually clearer if want to follow arrows: you cannot compute D from B; you need both B and C.

In order to fully get the joint representation given by the BN we would also need to make the final assumption that $B \perp\!\!\!\perp C \mid A$. This is possible to do with an extra arrow, but this solution doe not scale well and clutters the diagram. Instead, we will leave the picture as it is, and tackle the independence in a weaker way.

Definition. The *center* of a ⟨*model name*⟩ $(\mathcal{A}, \mathcal{L})$ is the set of joint distributions on $\mathcal{A}$ that come closest to satisfying $\mathcal{L}$, which are of maximum entropy.

This is an alternate way of capturing the conditional independence of variables that are not required by the

Oliver, unless I'm missing something, there's a gap here/issue to clarify: what is the semantics of one of your diagrams. I thought it was the set of probability measures consistent with the diagram. If that's what you intend,you need to say so. But given the definition above, I'm guessing that you now want to take the semantics of a graph to be the distribution that maximizes entropy among all those consistent with the graph. If that's the case, you have to say so explicitly. In particular, you need a careful discussion of the semantics of a graph (something that I asked you to do before!) Note that in general, given a set of distributions there isn't a unique distribution that maximizes entropy. There is if the set of distributions is convex. I suspect that the set of distributions consistent with a graph is convex. In any case, this needs to be proved. (I think that this is what you'll need to prove Conjecture 1.1.)

(no need in general, just for conjecture 1.1.)

I don't know what more you're looking for: it's a collection of marginals representing a belief / pref state.

model to be related, without ever conditioning on ancestors, which include products[2]. In some sense this is the worst case outcome for an agent intending to narrow down possibilities: a distribution in the center requires the maximum amount of information to determine the state of the world, but at the same time cannot leverage this assumption to simplify things. The principle of maximum entropy is well established[3] and there are many other, more complete arguments for it **[todo:** *choose references***]**. We also get:

Conjecture 1.1. *The center of the* ⟨model name⟩ *obtained by transforming a Bayesian Network as described above (i.e., by inserting an extra node* $X := \prod_{P \in \mathrm{pa(A)}} P$ *before every node, with the appropriate projections), is a singleton set consisting of exactly the probability distribution that the BN represents.*

OK, we should take it out for the economists, but I want to keep it around somewhere in some form b/c it's both interesting & useful.

Conjecture 1.2. *Products (defined in section 1.1) in the center of a model* $(\mathcal{A}, \mathcal{L})$ *have the usual uniqueness to make them categorical limits.*

Trust me, we really don't want this in the paper.

We will clarify this definition and explain the connection to thermodynamics more carefully (section 3.3) once we have a numerical definition of consistency. In the mean time, we continue to argue for this collection of marginals as a way of representing beliefs.

I agree that people have only partial information. But marginals represent just one form of partial information. You may want to capture, for example, "all I know is that the probability is between 1/3 and 2/3".

1.3 Human Beliefs as Marginals

This can be handled by •, the 'something else' bucket — though I'm sure there are other forms which cannot be.

From a modeling perspective, the biggest reason for keeping track of marginals rather than joint probability distributions, is that it can represent lots of information tracked by people, that BNs regard as incomplete. As seen earlier, the major feature we admit that is prohibited in a Baysian Network is a merging of two paths, where neither is a projection.

In our case, we also have the possibility of branching and merging. In our picture, a diagram $A \to C \leftarrow B$ represents two probability distributions $\Pr(C \mid A)$ and $\Pr(C \mid B)$. Having this kind of information is both common and not representable as a (single) probability distribution. Scientific studies never control for everything that is relevant, so you're left with marginals which may not tell the whole story.

I don't see how the "something else" bucket gives you intervals.

Example 1.1. After reading a number of empirical studies, you come to believe that smokers have a 70% chance of developing cancer, compared to 20% for non-smokers. At the same time, you believe that those who use tanning beds have a 80% chance of developing cancer, compared to 18 % for those who do not use them. You have no information about how the two interact. △

Example 1.2. You are on a game show, and offered a choice between several levers (A); your choice will determine how much money you receive. You are uncertain what each lever does, but you do have a vague intuition about the mechanism, giving you a distribution over amounts for each lever ($\Pr(C \mid A)$). You also had enough time to read statistics about how well people have done in the past ($\Pr(C)$). You do not have any information about what levers they've chosen though, nor do you have a complete joint probability $\Pr(A, C)$. In fact, having an accurate probability on A alone would seem to undermine your agency. △

So information of this form may not be entirely complete, be contradictory, or make it seem as though the choice is an illusion. This "outside view" is also important for constructing Newcomb's problem:

Example 1.3 (Newcomb)**.** There are two boxes. Box 1 is clear and visibly contains \$1k; box 2 is opaque, and will contain \$1M if a predictor (which you know has been very accurate in the past) predicts you will

[2] to see why this is problematic, consider that the product of all variables can always be formed, and is the ancestor of all variables. If we condition on it, then everything is always conditionally independent, as we're looking at a degenerate distribution consisting of a single outcome.

[3] However, preferences are still thought of as things which need to be totally nailed down rather than maximum entropy; this project can be thought of as pushing in this direction

leave box 1, taking just this box, and will contain nothing if the predictor predicts you will take the first box. You have to choose whether or not to take the visible box (there's no reason not to take box 2).

In the diagram below, you and your past choices determine the prediction P, as given by the blue line, but you do not have access to this information. They also determine your action A in some sense, which is the right dashed arrow, which you only know for certain after you make your decision. Because the two processes cannot exchange information (i.e., the predictor cannot decide after your action has been made, and you don't get to know the prediction), these two processes determine the process Identity $\to A \times P$.

$$\begin{tikzcd}
& \text{Identity} \arrow[dl, dashed] \arrow[d, dotted] \arrow[dr, dashed] & \\
\cancel{\text{Action}} \; A \arrow[ur, dashed, bend left] \arrow[rr, bend left] & A \times P \arrow[l] \arrow[r] \arrow[dl] \arrow[d] & \cancel{\text{Prediction}} \; P \\
\$\$ & A \overset{?}{=} P \; = \mathbb{1}_{A \overset{?}{=} P} & 1 \arrow[l]
\end{tikzcd}$$

(1) Yes
(2) Yes
(3) Yes
(4) Here
(5)

I don't understand this graph. Some points of confusion: (1) Are A and Action the same? (2) What about P and prediction? (3) What does A=P mean? Is it a random variable that is 1 iff the prediction is correct? (4) Above you said that "you and your past choices* determine P. Where are your past choices in this picture? (5) What if you believe that your action determines the prediction. How is that captured in the graph?

Then you get the pink arrow, which to me only makes sense by inverting the 'Identity → A' arrow, dashed above.

Now, on the one hand, you have a logical picture of how an action and prediction together will impact whether the predictor is right (the arrow $A \times P \to A \overset{?}{=} P$). On the other hand, you also have an outside view of how likely the predictor is to be right (the arrow $1 \to A \overset{?}{=} P$). We have not resolved the paradox; we have merely represented the beliefs in the setup, which make very little sense as a BN. △

I think this need *much* more discussion

At least classically, beliefs are only half of the picture. Keeping around marginals is also more psychologically plausible for preferences.

1.4 Human Preferences as Marginals

Is this issue critical to anything else you're doing? If not, cut it. Less is more. Why bother bringing up a controversial psychological issue. It will be a distraction.

In this section we will argue that the distinction between 'pure' utilities and 'expected' ones does not make much psychological sense. Clearly from a mathematical perspective the two are different: the pure utilities can be used to make sure the whole picture is consistent in the classical setting, and are unchanged by beliefs. In our view these are both bad things for modeling people, and drawing this distinction also causes the math to be less unified in subtler ways. In any case

I'm arguing that we don't actually need the standard utility function, but can get away with marginals. In addition to the math, it feels more plausible.

Example 1.4. Suppose you really like ice cream, and assign high utility to it. But now your family and government collaborate to make sure that whenever you eat ice cream you get an unpleasant electric shock, and feel awful for a day. It's bad enough that it's definitely not worth eating ice cream. But in this new life of yours, do you like it? The classical answer is yes. You will always like ice cream, it's just that you dislike shock. But we know that experiences that co-occur blend into one another a lot; a bad meal or bad date can ruin a restaurant, and that's much more ephemeral than a shock collar that is now part of your life. It's even harder to argue this if you consider the collar being put on before you've ever had ice cream. Now ice cream tastes like pain, and you would avoid it either way.

What does "pain" taste like? How do you know? How would you test this? You're getting into murky psychological/philosophical waters.

It gets worse still: what if instead of an external device, the procedure was a hypnosis that made you feel this way? It seems hard to argue that the new version of you still likes ice cream. △

This isn't clear to me. In fact, I find this whole discussion far too fuzzy. Moreover, I htink that making it precise would take us too far afield.

ok that was just prose, the point is that we know that co-occurring events bleed into each other (and have mechanisms explaining why) – classical conditioning, Hebbian dynamics, ...

When you can't separate two effects, there's no reason to talk about them, and no way to differentiate between them; such a representation should fade into the background. There's also no reason to restrict ourselves to trivial things such as tastes. The more important things, too (and perhaps especially, since they're further away from qualia) should be treated in context.

Let's not get into qualia …

Fine, rewording: further away from sensation, incorporating more reasoning / beliefs?

Maybe but there's still valuable intuition here: what co-occurs with what constrains preferences, and this illustrates that you can't solve the problem by drawing a line between external & internal.

The point is that the preferences an agent forms are constrained by experiences/reasons, and therefore dependent on the state of the world. Moreover, this still (and perhaps especially) happens when we're talking about important things like freedom — not just ice cream.

Example 1.5. You think freedom is objectively valuable. Governments and organizations that restrict freedom give worlds negative utility in your view, and tools that give people more freedom are of positive utility. But why do you like freedom? Presumably it's in part based on experience, and reading, and thinking about the structure of how societies are set up. It's not just a preference which comes from nowhere. If you lived in an alternate reality where people were much more malevolent and freedom was always associated with murder— where free societies collapsed immediately and tools that empowered people were invariably used for evil— would you still think freedom is good? △

Maybe not, but it's not clear to me what point you're trying to make here. I would cut this example too.

What you really mean when you say, colloquially, that something has high utility, or that you like one thing more than another, is that you like it *in context*. It's not that you would universally like it to be true over the alternative, as CP nets assume, or that you have some fixed platonic additive component of a utility function inside of you that gives you some number of points for freedom, as the classical view of economics would suggest. We're really talking about expected utility — or more generally, the marginal distribution of "good" given different choices, given your current beliefs. Even this insistence on collapsing the distribution to its mean is problematic, and makes problems like risk aversion trickier to talk about. The risk associated with not feeling good even though the state of the world is exactly what you thought it would be is almost never accounted for, and yet seems like a big deal, especially for people who don't have clear purpose in life.

I agree that context matters. I see no reason that this should mean that we're "really" talking about expected utility or marginals. At this point, I think you need a clear definition of context and what it means in your graphical setting.

We can now get back to more technical benefits of looking at the world as a collection of conditional distributions.

Ah I see why this looks like a leap. Let me try again: a marginal $A \to U$ encodes context about all variables $B \neq A$, implicitly. Expected utility is a special case. I'm still arguing that these make more sense than a complete utility function, b/c human prefs are in context.

1.5 Composition Of Arrows

One feature we really would like to have is the ability to chain these conditional distributions together. Among other things, a well-behaved notion of composition will:

- give meanings to the visual paths
- allow us to represent integrating out variables graphically
- unlock parallels to other structures through category theory
- think of expected utility, belief propagation, and inductive inference as simple juxtapositions of arrows

Since arrows are Markov kernels / matrices, we already have a natural way of doing this, by integration / summation over the intermediate variable — if $f : A \times B \to \mathbb{R}$ and $g : B \times C \to \mathbb{R}$ (this is the finite case) we can obtain a stochastic matrix

Slow down and give intuition. Don't introduce new notions (Markov kernel/matrix) out of the blue! (I would prefer that you didn't bring them into this part of the paper at all.)

Right. I didn't intend for this to be a stand-alone document. This is in the previous one.

Oops, assumed this was standard review

$$g \circ f : A \times C \to \mathbb{R} = \sum_{b \in B} f(b \mid a) g(c \mid b)$$

It depends on your goal, which you haven't enunciated. Even if you're not "done" (research is never done), it might be a good time to assess the consequences of the current formalism.

So are we done? We now have a notion of composition that lines up with matrix multiplication and gives us a marginal of the appropriate kind. Unfortunately, there are a few good reasons people are deterred from this formulation.

To the extent that you're right that people represent things using marginals, then the fact that there are possibly multiple ways of calculating things should have testable consequences (and is a source of inconsistency, which according to you is what drives (preference) change

Not necessarily, as per example 1.9.

Probably. Maybe worth doing. But we don't even need to go that far — examples make this clear (e.g. 1.10)

1.5.1 Should paths be equal?

This design decision also has the effect of proving multiple ways to calculate things, which leads to the somewhat counter-intuitive fact that that not all diagrams commute, even ignoring preferences entirely. We will begin by explaining why this might not be what one would expect.

Behave which way? This is far from clear (and may require an unreasonable state space).

It does.

1. Markov processes behave this way, and most things can be modeled as Markov processes. Unfortunately, to do this, we require control over the state representation — and many of our variables might have ~~way less information capacity~~ than would be required to make this work.

   compositionally

   This is the premise behind Hamiltonian dynamics / symplectic geometry, which has been very successful in modeling most everything in physics.

   not big enough. State space too small.

   What is the "information capacity" of a variable? Please don't use undefined notions.

2. If your distribution can be described as a Bayesian network without any merging, then all diagrams commute (this is simply a result of the associativity of composition).

3. If everything were deterministic, all diagrams would commute

Equal in what sense?

??? the two paths must be equal. Sorry for unneeded category theory jargon.

4. Even in our setting, many of them have to commute, and in fact several axioms of probability can be expressed as requirements that large classes of them must.

Example 1.6 (Marginalization)**.** Recall how a probability can be obtained by marginalization:

$$p(b) = \sum_{a \in A} p(a \wedge b) = \sum_{a \in A} p(a) \frac{p(a \wedge b)}{p(a)} = \sum_{a \in A} p(a) p(b \mid a)$$

below is an illustration of this fact:

$$\begin{tikzcd} & 1 \arrow[dl, "p(A)"'] \arrow[dr, "p(B)"] & \\ A \arrow[rr, "p(B|A)"'] & & B \end{tikzcd}$$

The left part of the diagram represents the right side of the equation and vice versa. △

This can be used inductively to show that every pair of paths from the singleton object 1 is equivalent, but before that we will deal with another important case:

Example 1.7 (Bayes Rule)**.** We can also represent Bayes' rule, $p(a \mid b)p(b) = p(b \mid a)p(a)$ as an assertion that two paths from:

$$\begin{tikzcd} & 1 \arrow[dl, "p(A)"'] \arrow[dr, "p(B)"] & \\ A \arrow[dr, "p(B\times A|A)"'] & & B \arrow[dl, "p(B\times A|B)"] \\ & B \times A & \end{tikzcd}$$

This can be seen as two applications of marginalization, one for each half. On the left, we have

$$p(b, a) = \sum_{a' \in A} p(a') p(a, b \mid a') = \sum_{a' \in A} p(a') p(b \mid a') \delta_{a,a'} = p(a) p(b \mid a)$$

and similarly, the right gives $p(b, a) = p(b)p(a \mid b)$. One thing to take away is that one can avoid the integration over a variable by simply considering the conditional distribution to a different variable: namely, one which is the product of the input and output. △

However, not all paths generated by composition of probabilities are strictly equal!

Example 1.8. In an extreme case, we can forget all of our information with the Markov assumption by going through a singleton object:

Why not write "1"? this is the standard notation for an identity morphism. There may be some confusion: P(A|A) is a matrix (finite case)

$$\begin{tikzcd} A \arrow[rr, "p(A|A)=\mathrm{id}_A"] \arrow[dr] & & A \\ & 1 \arrow[ur, "p(A)"'] & \end{tikzcd}$$

$$P(A|A) = \begin{array}{c|cccc} & a_0 & a_1 & a_2 & \cdots \\ \hline a_0 & 1 & & & \\ a_1 & & 1 & 0 & \\ \vdots & 0 & & 1 & \end{array}$$

OK; I understand. The (horrible!) notation p(AIA) confused me.

Was trying not to say "all diagrams commute" $\triangle$

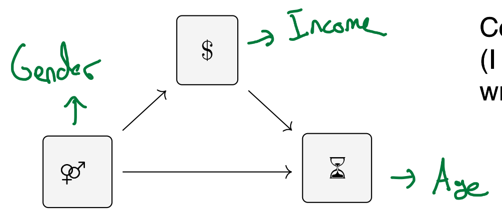
This is a strange way of putting it

For this to happen, the only thing we need is to allow composition and provide a probability on A; there's nothing inconsistent about this picture. Therefore, the measure of consistency is weaker than "all paths are equal". Still, there are some blatantly inconsistent pictures one could draw — anything that violates Bayes' rule or marginalization, for example (see section 2 for more).

I don't understand what it means for two paths to be equal!

Our singleton example is a little bit annoying, but at least it's the best prediction that could be made after forgetting all of the information. It is natural to ask: is ignoring everything the *worst* we can do? Is every bit of signal helpful? The answer is no.

Example 1.9. Men earn more than women, and people who earn more are generally older, but women live longer than men, so the top composition in the picture below



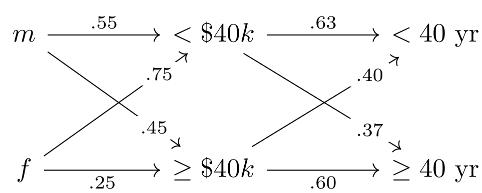
Could you explain the labels? (I would also prefer letters, since writing Pr($) seems very strange to me.)

Sure, thought it might be more compact / require less interpretation.

is worse than ignoring all information and just predicting age. Here it is with numbers. Suppose the truth is a conditional probability distribution Pr($, ⌛ |⚥) given by

	♂		♀	
	<40 yr	≥40 yr	<40 yr	≥40 yr
< $40k	.225	.05	.185	.19
≥ $40k	.075	.15	.065	.06

We can now construct our chain, ⚥ → $ → ⌛



Now, consider the following three arrows ⚥ → ⌛, as estimates of Pr(⌛ |⚥):

(1) the truth, Pr(⌛ |⚥):

	<40 yr	≥40 yr
♂	.6	.4
♀	.5	.5

(2) ⚥ → 1 → ⌛:

	<40 yr	≥40 yr
♂	.55	.45
♀	.55	.45

(3) ⚥ → $ → ⌛:

	<40 yr	≥40 yr
♂	.53	.47
♀	.57	.43

We can see that (2), which kills all signal, is closer to the truth than (3) in every way. Still, the picture is entirely consistent. Moreover, all of the important details of the joint distribution are saved in the three arrows, and subjectively, I used the arrows to construct the joint distribution I wanted, rather than the other way around.

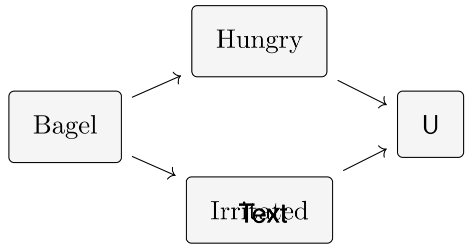
What does "equivalent" mean?

Even though this triangle does not commute, still every pair of paths from 1 to another node are equivalent; for instance, marginalizing out gender gives the same distribution on age in all three of the cases above. $\triangle$

For every variable A, any way of composing paths $1 \to \dots \to A$ gives the same marginal

There are more persuasive examples using preferences. Multiple paths from a domain to U can be thought of as pieces of a pros/cons list; in this case, people are intimately familiar with the fact that paths are not always equal.

Example 1.10. In expectation, if I eat this bagel, I'm less likely to be hungry, and when I'm not hungry, I'm likely to be happier. On the other hand, suppose I'm allergic to gluten, and if I eat the bagel, I'm also likely to be irritated and uncomfortable, and when irritated and uncomfortable, I'm likely to be less happy.



The two paths are opposite polarity and hence the two paths cannot be equal. △

The very fact that we write down both pros and cons implies paths aren't in general equal. So. Why even bother with composition then, if it doesn't give you the truth? People still write down arguments in support of and refuting positions, and often this is helpful. In example 1.10 the intuition is still that somehow by weighting the reasons appropriately and finding the centroid we're likely to reach a good decision.

Also, as mentioned in the beginning of the section, we can do some cool things with composition. The beliefs we model clearly should not necessarily be closed under composition, but being able to form them is still useful:

In what sense? → The maximum entropy sense. I don't know if I want to get into the details here.

- In the absence of additional information, the composition of two links is in some sense the best available estimate, and is compatible with any distribution in the center of the ⟨*model name*⟩ .
- If multiple, disjoint paths, agree this is good evidence that the final estimate is good — like obtaining the same value from two different fermi estimates.
- In some cases, the composition is guaranteed to equal or approximate the true probability.

Did you mean to add "agree"?

oops

I have no idea what a "fermi estimate" is.

In what sense? Why is this true?

→ (1) if processes don't exchange info,
(2) if we make indep. assumptions.
(3) limit as variable incorporates more relevant info
(4) limit as we weight more jos (related to ex. 1.10, where we add all relevant paths)
I definitely didn't have time to expand this, good point.

1.5.2 Some Results

Props 1.3, 1.4

Still, path equality is often expected; we would like to characterize when and why. Below are some necessary conditions for consistency, although more exploration is required.

Proposition 1.3. *If π is a path of conditional probabilities $1 = A_0 \to A_1 \to \cdots \to A_N = X$, then the composition π° of links in π is equal to* $\Pr(X)$.

Proof. This can be done by induction on the result from example 1.6, which is also the base case. The inductive step is as follows: if there is some joint probability p, and $\pi^{\circ k} := 1 \to \cdots \to A_k = p(A_k)$, then since we've assumed $p(A_{k+1} \mid A_k) = \pi_k$, we also have

$$\pi^{\circ k+1} = \sum_{a \in A_k} p(A_k)(a) p(A_{k+1} \mid A_k(a)) = p(A_{k+1})$$

again by the example. □

Similarly, we have the dual result for deterministic functions:

Proposition 1.4. *If a variable Q is completely determined by both A and B, i.e., $g : A \to Q$ and $h : B \to Q$ are deterministic, and $f : A \to \Delta B$ is* $\Pr(B \mid A)$, *then $h \circ f = g$.*

Perhaps write \Pr_f for the probability measure corresponding to f.

Maybe? But I do mean to require that f is actually the marginal probability of B given A.

Again, Pr(BIA) is horrible notation. It looks a single conditional probability, but that's not what you intend.

Proof. If there is a non-zero probability that $A = a$ while $B = b$, then it must be the case that $g(a) = h(b)$, since both a and b determine Q. So

$$h \circ f(a, q) = \sum_{b \in B} f(a, b)\delta_{h(b),q} = \sum_{b \in B} f(a, b)\delta_{g(a),q} = \delta_{g(a),q} \sum_{b \in B} f(a, b) = \delta_{g(a),q} = g(a, q)$$

$\square$

,

We can also use information theory to obtain some bounds. Degenerate, deterministic marginals (zero entropy) must commute, and because the paths are guaranteed to be centered (see how even in example 1.9 the means of the three compositions are the same) the possible deviation between paths is bounded by the entropy of the components. As a result, we will be able to show something in the spirit of the following (though it probably needs some revision):

Conjecture 1.5. *A composition of arrows can only be as far off as its entropy permits: if we have marginals* $A \xrightarrow{L} B \xrightarrow{P} C$, *then*

$$\forall a \in A. \left| P \circ L(a) - \Pr(C \mid a) \right| \leq H(L) + H(P)$$

where $\Pr$ *is any distribution consistent with* L *and* P.

To summarize: we have a picture containing conditional distributions, particularly the ones which are most useful and actionable. Each conditional is a constraint on the world. We have a natural way of composing distributions, but sometimes the composition of two distributions will not be consistent with the rest of your beliefs.

1.6 Altering Diagrams

1.7 Relations between Utilities and Probabilities

[**todo**: *This is in the other documents. I can lift it if this becomes something more stable and we want it here*]

1.8 Sub-stochastic Transitions and Conditionals

Often, an otherwise very useful variable might not apply. The variable describing whether or not your answer is correct doesn't make sense if you weren't solving problems; the amount of money in your wallet doesn't make sense if you don't own one, and so forth. So now, when you're trying to predict the probability of certain amounts of money in your wallet, some of the probability mass needs to go into the "actually this doesn't make sense" bucket.

While I have no problem with having a "something else" bucket (as opposed to "this doesn't make sense") the fact that some useful variable may not apply seems completely orthogonal

it's still good motivation but yes.

Sure.

Usually, this is not really a problem, because we can always just add that bucket as a real option that a variable can take: a variable which might not make sense can always take a `null` value, and so now the set of possibilities is once again exhaustive. For us, this resolution poses a problem: our marginals now require us to estimate the distribution of many things given a null value— suppose you are trying to represent the belief that you're happier when you get the right answer as a marginal link $L[\text{RightAns} \to ☺]$. We now need a distribution on happiness when you get the right answer, when you get the wrong answer, and also for when the question doesn't apply. But why doesn't it apply? Are you not solving problems because you're skiing? Because you've been injured? Maybe you are solving problems but there are multiple right answers? You can't just answer with a prior over happiness if you want to have consistent beliefs, because solving problems and happiness might be correlated. To be fair, this is something you certainly *could* have, but it's annoying that we cannot provide a belief about "does the right answer make you happy?" without also answering the much more difficult question, "are you happier when 'the right answer' is an applicable concept to your life?".

I don't think there should be a "the question doesn't apply" category.

I'm on board with calling it a "something else" category.

This seems to be one of many problems with taking a completely Bayesian position. If you could, e.g., use intervals (like [0,1]), you may have fewer problems

In addition to being psychologically implausible, it dramatically reduces the number of things we can represent; take implication, for instance. If A, B are binary variables (taking values $a, \bar{a}$ and $b, \bar{b}$ respectively), we can easily represent $A = B$, $A = \neg B$ as stochastic matrices

What I do is similar, and might be equivalent. I'll have to look into it. Is there a standard way of composing them?

I suspect so.

$$
p(B \mid A) = \begin{array}{cc} & \begin{array}{cc} b & \bar{b} \end{array} \\ & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{array}{c} a \\ \bar{a} \end{array} \end{array} \qquad \text{and} \qquad p(B \mid A) = \begin{array}{cc} & \begin{array}{cc} b & \bar{b} \end{array} \\ & \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{array}{c} a \\ \bar{a} \end{array} \end{array}
$$

but you cannot (via stochastic matrices) believe that $A \Rightarrow B$ without also believing a prior over B given $\bar{a}$. Maybe the best strategy is a uniform prior (principle of maximum entropy, used in theorem proving in [**logicalinduction**]), but this makes your beliefs inconsistent if it happens that B is always false for other, unrelated reasons.

For this reason, we drop the requirement that our null element, $\bullet$, indexes a distribution in marginals. Below is an example of transition matrix $A \to B$ including the extra element. As mentioned, the last row is not something we are keeping track of.

$$
\begin{array}{cc}
\begin{array}{ccc} b_0 & b_1 & \bullet \end{array} & \\
\left[\begin{array}{ccc} .2 & .1 & 0.7 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ \hline .2 & .6 & 0 \end{array}\right] & \begin{array}{c} a_0 \\ a_1 \\ a_2 \\ \bullet \end{array}
\end{array}
$$

Furthermore, because the final column is just whatever is necessary to make the rows sum to 1, we don't need to keep that either; as a result, it is sufficient to keep a smaller matrix without any $\bullet$-indices; the only price that we pay is that this matrix is *sub*-stochastic rather than stochastic: its row entries sum to at most 1, rather than exactly 1. Composition works just as before; the product of sub-stochastic matrices is sub-stochastic.

Oops maybe too strong. But I still don't see how.

I'm not sure that I agree

There is no way to get a Bayesian Network to do this, because we require the look-up tables to exactly match all possible values, so we get a real distribution at the end, rather than a weaker ~~sub-distribution.~~ constraint on distributions.

There are several other technical results that we will eventually need to make sure that no tricks are being pulled when we later use standard formulae for probability distributions, but I'm going to skip fully typesetting and verifying them for now.

1.8.1 Substochastic Sanity Results

[**todo**: *show we can still use Bayes rule, marginalization etc., under certain circumstances. Most should follow just by adding $\bullet$ as an absorbing state and interpreting in a larger space*]

2 Inconsistency

We will start with a simple predicate definition of ⟨*model name*⟩ consistency, of preference consistency, and give some examples before stating a less-brittle continuous generalization that will allow us to deal with inconsistent models.

Definition 2.1 (consistency)**.** A ⟨*model name*⟩ $(\mathcal{A}, \mathcal{L})$ is *consistent* if there exists some joint probability $\Pr(\mathcal{A})$ on all of the variables, that is consistent with every link marginal $L \in \mathcal{L}$.

Definition 2.2. A ⟨*model name*⟩ $(\mathcal{A}, \mathcal{L})$, in which preferences are represented by a single utility node $U \in \mathcal{A}$, is *pref-inconsistent* if it is inconsistent, but only because of the utility node — i.e., the model obtained by deleting U and all links to it, $\big(\mathcal{A} \setminus U, \;\; \mathcal{L} \setminus \{L : \square \to U\}\big)$ is consistent.

We will now show how they correspond to other intuitions of inconsistent preferences through some examples and specific cases.

Only in special cases:

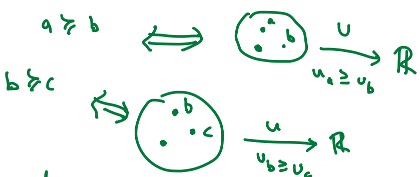
- The framing problem example in § 2.2
- Rigidly enforcing Max Entropy
  - Violations of props 1.5, 1.4
    - multiple paths outside one another's entropy cone

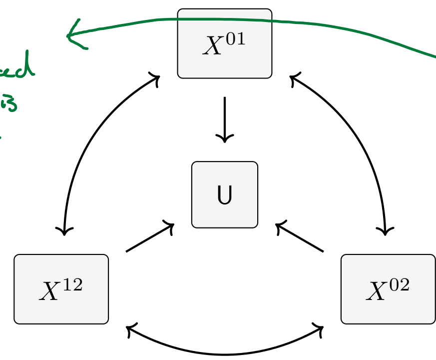
2.1 Intransitivity of Preferences as Inconsistency

Couldn't you get intransitivity by focusing on different paths?

While restricting ourselves to use preferences represented by a utility domain, we cannot represent intransitive preferences directly in the usual way. However, we can use the • and sub-stochasticity to break one domain into many smaller ones, each one representing a binary forced choice. For instance, suppose we have one variable X, which can take the three values $\{x_0, x_1, x_2\}$. Then we can form domains X^{01}, X^{12}, and X^{02}, where X^{01} contains elements corresponding to x_0 and x_1, and so on. We also assume it has preferences on each subdomain represented by utilities. This gives our agent the following picture:

Argument against ad-hoc: each of these domains is a forced choice between 2 elements; this is just the way to talk about binary relations with minimal additional assumptions.





This seems to me rather forced and ad hoc. I would prefer a solution that focused on different paths: when you're looking at one path, you get A > B > C; when you're looking at a different path, you get C > A. If this works, it also provides an "explanation" of the intransitivity.

This works if the paths are of low entropy and not close, but ex 1.9 shows that's not how it works in general.

The only thing that might be unnatural is a shared U domain.

The complete domain X, which we won't include in the agent's picture, could of course replace the entire clique as a simpler representation. The constraints between the domains are just the logical ones: $L[A \to B]_{a,b} = \delta_{a,b}$. We can expand further to give a more concrete sense of the shape of what's going on here— each matrix between the subdomains is of the form

$$L[X^{ab} \to X^{cd}] = \begin{array}{cc} & \begin{array}{cc} c & d \end{array} \\ \begin{bmatrix} \delta_{a,c} & \delta_{a,d} \\ \delta_{b,c} & \delta_{b,d} \end{bmatrix} & \begin{array}{c} a \\ b \end{array} \end{array}$$

Let's discuss this. I don't see what low entropy has to do with it.

And more concretely still:

$$L[X^{01} \to X^{12}] = \begin{array}{cc} & \begin{array}{cc} x_1 & x_2 \end{array} \\ \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} & \begin{array}{c} x_0 \\ x_1 \end{array} \end{array} \qquad L[X^{12} \to X^{02}] = \begin{array}{cc} & \begin{array}{cc} x_0 & x_2 \end{array} \\ \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} & \begin{array}{c} x_1 \\ x_2 \end{array} \end{array} \qquad L[X^{02} \to X^{01}] = \begin{array}{cc} & \begin{array}{cc} x_0 & x_1 \end{array} \\ \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} & \begin{array}{c} x_0 \\ x_2 \end{array} \end{array}$$

The other three are the transposes of these. Note that without the utility node, the ⟨*model name*⟩ is perfectly consistent: any distribution which assigns the same weight to identified values — that is to say, any distribution on $X^{01} \times X^{12} \times X^{02}$ which sense as a distribution on X — will work.

Now we can use this structure to encode any complete binary relation. Suppose that the agent's preferences on X are intransitive. Without loss of generality, suppose this occurs as $x_0 \succcurlyeq x_1 \succcurlyeq x_2 \succ x_0$ (that is, $x_0 \succcurlyeq x_1$ and $x_1 \succcurlyeq x_2$ but not $x_0 \not\succcurlyeq x_2$). This scenario is represented with utility marginals such that, this gives us

$$\mathbb{E}U_{X^{01}}(x_0) \geq \mathbb{E}U_{X^{01}}(x_1) \qquad \mathbb{E}U_{X^{12}}(x_1) \geq \mathbb{E}U_{X^{12}}(x_2) \qquad \text{and} \qquad \mathbb{E}U_{X^{02}}(x_2) > \mathbb{E}U_{X^{12}}(x_0) \tag{1}$$

where $\mathbb{E}U_Y(y)$ is the mean of the distribution $U_Y(y)$ over $\mathsf{U} \cong \mathbb{R}$. Now, in search of a contradiction, suppose that there is a joint probability distribution p over $X^{01} \times X^{12} \times X^{02} \times \mathsf{U}$ which marginalizes out to each of the links above, and whose marginals on U satisfy (1). Due to the logical links, we know that the same probability must be assigned to the same event in different domains; for example,

$$p(X^{01} = x_0) \;=\; \frac{p(X^{01} = x_0 \mid X^{02} = x_0)}{p(X^{02} = x_0 \mid X^{01} = x_0)} p(X^{02} = x_0) \;=\; p(X^{02} = x_0) =: p(x_0)$$

and also, since whenever $X^{01} = x_0$, $X^{02} = x_0$ with probability 1 and vice versa, conditioning on the two events is equivalent. Therefore,

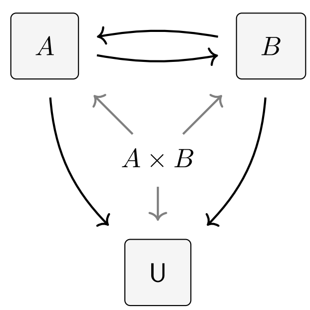
$$\begin{aligned} u_0 := \int_{u:\mathbb{R}} u \cdot p(u \mid x_0)\ \mathrm{d}\mu = \mathbb{E}U_{X^{01}}(x_0) \ &\geq\ \mathbb{E}U_{X^{01}}(x_1) = \int_{u:\mathbb{R}} u \cdot p(u \mid x_1)\ \mathrm{d}\mu \\ = \mathbb{E}U_{X^{12}}(x_1) \ &\geq\ \mathbb{E}U_{X^{12}}(x_2) = \int_{u:\mathbb{R}} u \cdot p(u \mid x_2)\ \mathrm{d}\mu \\ = \mathbb{E}U_{X^{02}}(x_2) \ &>\ \mathbb{E}U_{X^{02}}(x_0) = \int_{u:\mathbb{R}} u \cdot p(u \mid x_0)\ \mathrm{d}\mu = u_0 \end{aligned}$$

which is a contradiction. Therefore, this intransitive set of preferences is pref-inconsistent.

Without much effort this can easily be extended to show that in this encoding, an arbitrary sets and intransitive binary relations on it, results in a pref-inconsistent model.

2.2 Framing Problems as Inconsistency

Consider once again a framing problem: there are two variables A and B which you have preferences over; maybe you think $a \succcurlyeq \bar{a}$ and $\bar{b} \succcurlyeq b$. Unfortunately, you later discover that they're the same variable. Below is a diagram of this:



It is easy to see that the logical correspondence between A and B alone admits plenty of joint distributions; any distribution on A will extend to B and vice versa. However, adding the utility makes it impossible to satisfy. This too should be clear: if a corresponds exactly to b, then for any probability measure p on $A \times B \times \mathsf{U}$,

$$\mathbb{E}_p(U \mid a) > \mathbb{E}_p(U \mid \bar{a}) = \mathbb{E}_p(U \mid \bar{b}) > \mathbb{E}_p(U \mid b) = \mathbb{E}_p(U \mid a)$$

which is a contradiction.

One might wonder if this is only true in the degenerate case— but it is easy to be inconsistent even if the marginals have full support. Fix some small $\epsilon > 0$, and consider the case that is ϵ way from the one we just described. Any joint distribution on all three variables must factor into $p = \lambda a.\lambda b.p(a, b)p(u|a, b)$. The distribution $p(A, B)$ must therefore look something like

$$\begin{array}{cc} \begin{array}{cc} a & \bar{a} \end{array} & \\ \begin{bmatrix} ? & \leq \epsilon \\ \leq \epsilon & ? \end{bmatrix} & \begin{array}{c} b \\ \bar{b} \end{array} \end{array}$$

Once again, obviously consistent. So now, if U_B is the marginal of U on B and U_A is the marginal on A, and the utility classically is a function $u : A \times B \to U$, we can write:

$$\mathbb{E}U_B(a') = \int_B p(a, b, u(a', b))u(a', b)\mathrm{d}b = \sum_b \overbrace{p\Big(u(a, b) \mid a, b\Big)}^{1} p(a', b)u(a', b) = p(a', b)u(a', b) + p(a', \bar{b})u(a', \bar{b})$$

And similarly,

$$\mathbb{E}U_A(b') = \int_A p(a, b', u(a, b'))u(a, b')\mathrm{d}b = p(a, b')u(a, b') + p(\bar{a}, b')u(\bar{a}, b')$$

Defining a' to be the variable which corresponds to b' in the case of U_B, and b' to be the one that corresponds to a' in the case of U_A, i.e.,

$$a' := \begin{cases} a & \text{when } b' = b \\ \bar{a} & \text{when } b' = \bar{b} \end{cases} \qquad\qquad b' := \begin{cases} b & \text{when } a' = a \\ \bar{b} & \text{when } a' = \bar{a} \end{cases}$$

we then have $\mathbb{E}U_B(b') \approx u(a', b') \approx \mathbb{E}U_A(a')$, which gives us a chain of inequalities leading to a contradiction:

$$u(a, b) \approx \mathbb{E}U_A(a) \gg U_A\bar{a} \approx u(\bar{a}, \bar{b}) \approx U_B\bar{b} \gg U_B b = \mathbb{E}U_B(b) \approx u(a, b)$$

Note that this works even if utilities are classical functions rather than distributions! By replacing the platonic ideal of utility with expected utility, which interacts with the real world, as we've argued for (see section 1.4), we can quickly get preferences that conflict purely due to beliefs.

2.3 Learning Problems

[**todo**: *adapt from previous document. Most everything there is correct, but in context it doesn't flow. I don't think I got feedback on that section yet*]

3 Dynamics

3.1 A Better Definition

What does it mean for a marginal to be consistent?

The definition of consistency provides no insight for how to reduce it, and leaves us requiring all of the marginals to be consistent in order to do anything, just as in the classical case.

i.e. admitting at least one joint distribution!

min score; sum across marginals; weight over possible inputs; relative entropy between link + what p marginalizes to.

$$\zeta(\mathcal{A}, \mathcal{L}) := \inf_{p:\mathrm{Prob}(\mathcal{A})} \sum_{L[A\to B]\in\mathcal{L}} \mathop{\mathbb{E}}_{a\sim p(A)} \left[D_{\mathrm{KL}}\Big(L(a) \;\Big\|\; p(B \mid a) \Big) \right] \tag{2}$$

I have no idea what's going on here. I don't know what the symbols mean or the words in the next sentence. Slow down, and give *lots* more intuition.

OK will do next version.

Although this looks like a complicated formula full of arbitrary symbols, it is in fact just the minimum total information needed to encode a distribution given your current beliefs. Note that if there is a distribution p that marginalizes out to match each link L, then the KL divergence will be zero for each link and input, and therefore the inconsistency will be zero; thus this is a special case of our previous definition.

Although it is not obviously differentiable, this function at least admits sub-gradients, which is a place to look in order to implement gradient descent. One can always also estimate a gradient by sampling, and ζ is at least continuous. Not sure I understand.

Useful because it unifies belief + preference revision, validates our notion of inconsistency, and ties to existing literature so now maybe we can use examples / intuitions from belief revision?

3.2 Belief Revision

later in section ← ?? How?? Why is this useful??

Belief revision, both through Bayes' and Jeffrey's rules, can be thought of as the addition of a new marginal to a distribution, and then a resolution of inconsistency. In Dietrich, List, Bradly [**dietrich2016belief**], a belief revision is an update $p \mapsto p_I$ of a belief state p (in their formulation, these are still distributions) to a new one consistent with the input I.

I don't understand. Why is this the right/ best way to look at it.

For us, belief revision is simply adding marginals to the picture, and then resolving inconsistencies. In fact, our representation makes this rather pretty: the observation of a Jeffrey input is simply a factorization of existing links through a new finite random variable; observing a Bayesian input is the the particular case

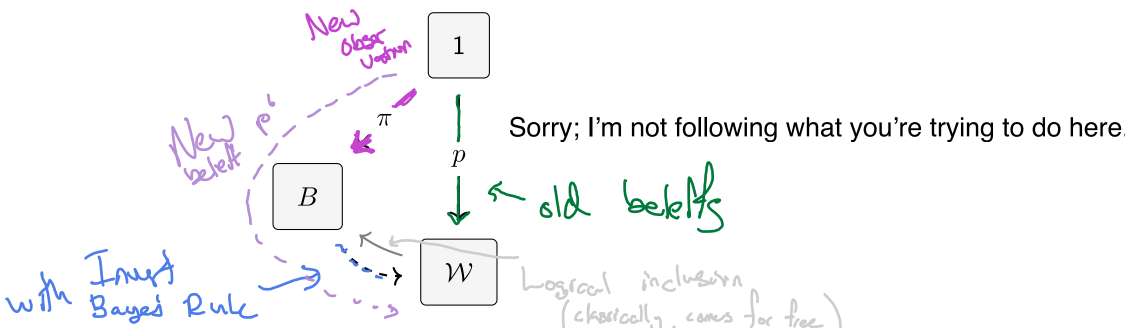
I'm just claiming its a way to look at it.

I think it's more general and the graphical models are nice but it might not be the best framing.

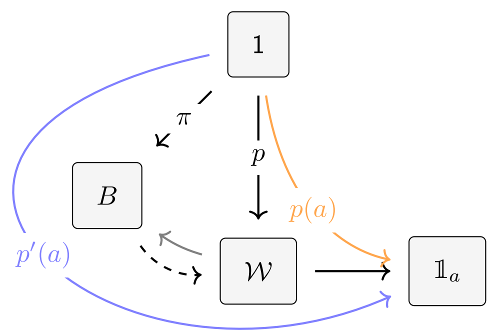
where the variable is binary and the observation is certain. Jeffrey's rule prescribes a posterior probability p' by:

$$p'(a) = \sum_{b \in B} p(a \mid b)\pi_b \qquad \text{for all outcomes } a \subseteq \mathcal{W}$$

Since variables can be thought of as partitions of outcomes, and at this point we're looking at the classical picture, where p is a distribution on all variables, we can draw a much cleaner picture, where the integration is implicit:



Now, $p' := p(\mathcal{W} \mid B)$ is just the left-most path. The gray arrow on the bottom left is just a projection / computation from the state of the world, and the dashed one is its inversion given by Bayes' rule, which is why the conditioning works out as in the formula. If we really want to match the formulation exactly we can put a into the picture— but rather than a subset of the outcomes, a is now a value that some variable can take. We can even create a special indicator variable for it, $\mathbb{1}_a$.



Visually it's much clearer what's going on: we've replaced the probability distribution $1 \to \mathcal{W}$ with the one that factors through B via the new observation π. In terms of evaluation, this means the orange path to $\mathbb{1}_a$ has been replaced by the blue one. This presentation also suggests a natural way we can generalize this to our setting, where we don't necessarily have full distributions but only a collection of marginals: we simply try to factor every distribution $1 \to *$ through B via π, as done with $\mathcal{W}$ above. The other marginals can stay the same, and the difference propagates through via composition.

In our case, there's also a simpler thing we could do, that's even more psychologically plausible: just add the new marginal π to our collection. Sure, it's probably inconsistent, but we can let the our inconsistency reduction take care of that. One might worry that it is likely we will now violate the responsiveness axiom [**dietrich2016belief**], as we could reject π — but I argue that this is not a concern. So long as an agent keeps observing or remembering the observation, we are effectively continuously reapplying consistency reduction while anchoring the new observation, until the responsiveness axiom is satisfied. This actually makes more sense than the standard belief revision picture: if a person doesn't spend long enough looking at it or thinking about it, they may forget or partially reject implications of this new view.

[**todo**: *Spend time converting the conservativeness axiom to this framework*]

One more benefit: belief revision no longer needs to happen immediately; we can add the marginal to our picture and deal with it later. This makes for an account which is much better suited to cognitively bounded agents, who might have more pressing matters than sorting through beliefs, and who might do them out of order.

$\sum$ bond energy, chemical energy, not taking entropy into account.

3.3 Preference Thermodynamics

??

We now have a quantity that behaves like enthalpy (inconsistency, ζ), as well as information theoretic entropy. We now ask the question: is it reasonable to use these and a temperature parameter to define an effective free energy for belief revision?

free energy

temp

$$G := \zeta - TH$$

inconsistency. entropy

At low temperatures, $G = \zeta$, so gradient descent is just minimization of inconsistency. At high temperatures, using gradient descent on G is mostly just entropy maximization: this corresponds to a softening of beliefs and preferences.

Of course, temperature could also be localized more to some links over others: you could be more willing or primed to reconsider beliefs about things you've just learned, for instance. Temperature can therefore function as the entrenchment parameter we've been looking for, and actually fits well with the physical analogy. When things are hot, they're maleable and are easily molded by the environment; as they cool off, they keep their history and identity. This seems like a fitting way of viewing agents as well.