Merged: Dynamic Beliefs and Preferences

Oliver Richardson oli@cs.cornell.edu

October 8, 2019

A Quick Overview

In some sense, the goal of this project is to provide a compelling picture of preference revision — but in doing so, we will paint a very different, generalized picture of decision theory.

The core representation we suggest, the marginal constraint graph, is more flexible than the standard ones, as agents need not always have a perfectly consistent picture of reality, intermediate representations have meanings, and we can capture phenomena such as learning preferences from data and revising them in various places. In special cases, we recover standard results, including expected utility calculations, belief updating by conditioning.

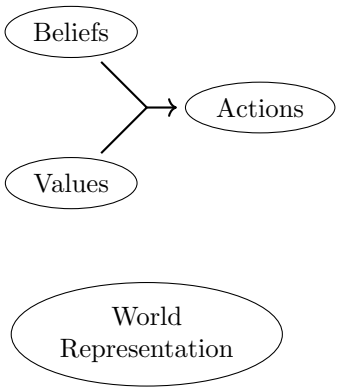
We can roughly separate the work into two pieces:

1. Representation: the sufficiency of our new representation as a unification of existing results. In particular, we think of an agent's values as distributed across a network of concepts, whose nodes are preferences on small sets of alternatives, and whose edges are beliefs about the impact of one variable on another.
2. Dynamics: A formulation of how to revise the representation in light of new information or additional computation. We will show how can be dealt with through inconsistency, illustrate some options for regularization, and ultimately factor out these as general meta-preferences.

Contents

Part I

Motivations



1 Issues with Standard Issue Decision Theory

1.1 Some Reasons not to Calculate Expected Utility

1. **No utility function**. Maybe your preferences are not representable in this way — maybe they violate the vNM axioms, or you just don't have anything that resembles an ordered set. Maybe your intrinsic utility is stochastic.
2. **No complete prior.** It is possible that you're not a Bayesian, and don't have (access to) a probability distribution on all salient features of your world[1].
3. **Thinking takes time.** (this constrains the complexity of your utility and probability information, or at least requires them to be in a form that admits fast computations)
4. **Internal Actions.** In addition to the external actions, there are also invisible mental ones: nobody has complete control over what they do in all respects at all times. Emotionally preparing yourself to talk to your boss, warming up your fingers to play piano, and intentionally committing a phone number to memory are all examples of mental actions, which are rarely modeled. More generally, for any interface with the world that a model prescribes, there will be extra
5. **Different world representation.** Models, both in probabalistic reasoning and decision theory, usually start by defining the set of possible states of the system. For a modal logician, this is the underlying set of a Kripke structure; for probability theorists, it's a set of outcomes.This is fine if agents are aware of the structure you've set up, but simply don't know what world they're in — but nothing precludes them from conceptualizing things in entirely different terms.

   When the agent fails to obey expected utility for this reason, there are three ways to assign blame: it could be the agent's fault for not truly understanding what is possible, the modeler's fault for failing to capture the the agent's picture of the world.
6. **Things Change over Time.**

Despite the fact that there has been work backing off of each of these assumptions, utility maximization is still ubiquitous (and taken for granted) in most primary applications, including consumption models and the

[1] In fact, once we add enough representational complexity to our world, this is impossible to have; see section B.1

construction of artificial agents. In section **??** I argue that this is the source of some cultural problems. But part of the reason that many of the other techniques has not gained traction is that none is as clean and as useful as expected utility maximization, and to the best of my knowledge people try to isolate and tackle these issues individually. My framework provides a way of representing many of these issues effectively and cleanly, while also reducing to the expected utility computations when none of these issues apply.

1.2 Fixed World Representation

The set of things which are even considered possible is contentious enough that we don't want to build it (and actually the correct one) canonically into an agent's picture of the world. In some sense only the current world is possible, as far as anyone can be sure. Moreover, no matter how you chose to represent the world, an agent's picture of what's possible can change over time, which means the modeler needs to write down beforehand everything that could possibly happen. We talk more about this in section 13.3

1.2.1 World Explosion

2 The Big Picture: Why This is Important

1. Subjective representations of the world
2. Joint dynamics
3.

2.1 Fixing the Ethos of Agency

The distinction between the things that you care about (e.g., utilities, goals, preferences, objectives), and the tools you have for understanding and affecting the world (e.g., beliefs, reasons, plans, optimization, inference) plays a large role in technical accounts of agency[2]. Generally the two are considered separately, and also people tend to fix preferences[3]. This is nice because the fixed preferences provide a clean way to evaluate the quality of decisions, and criteria for success are a very important part of science. It is clear why modeling things in this way caught on so quickly and durably — it is simple, prescribes tractable computations, is expressive enough to capture any behavior over fixed time periods, if you make modeling choices in the right ways.

This heavy dependence on modeling assumptions is an underappreciated shortcoming, and freedom to chose the representation gives modelers a lot of power to over-fit to their desired outcomes. I would cut this?

In real life, human preferences and beliefs often depend on one another, and in more complex settings the line is much more blurred:

- When you play a video game, are you trying just so you can win? If so, why do you care about winning? Decision theorists often take this as primitive, but it can also be considered in a larger picture, which decision theorists also purport to model. It buys you very little socially, and costs time and money. Or does the process of optimizing for this arbitrary goal itself have value? Does this make entertaining yourself fundamentally irrational? I'm missing the "story" of this example. How do we get an interdependence between belief and action here?

???

[2] Beliefs should be considered optimization power: they you believe that action A will have effect 1, an action B will have effect 2, and a preference between the 1 and 2

[3] In the context of optimization, this corresponds to the practice of writing down an objective and finding extrema; in the context of decision theory, it is finding the best actions to satisfy some preferences

- It is common for humans to have beliefs about preferences: ('I think I like cheese', 'I believe freedom is bad'), and also preferences over beliefs ('Believing true things is good', 'I want to know calculus') — but the clear separation in standard decision theory makes this difficult to understand.

  This is not a belief.

  These can be nested quite deeply without anyone thinking too hard: If I ask if you like cheddar, you now believe that I want to know whether you like cheese — note that this is a belief about a desire for a piece of knowledge about a preference, spanning multiple people.

- On a small scale, we think of a loss function as an optimization objective, and the algorithm (say, stochastic gradient descent) as optimization power. One can use both of these together to form a neural network, which is thought of as being just a tool, with no

  I don't see what this has to do with the interdependence of belief and preference.

What's a "complete agent"

- Complete agents can be used in either way: people, who have desires and experience emotions, can be used as optimization power with an external objective (e.g., hiring employees), but can also as the ultimate source of value (e.g., running an organization because it will help people)

  I still don't understand what it means for something to be "used as optimization power"

This is an argument against the orthogonality hypothesis [**orthogonality**], and similar lines of thought. It seems clear that if we could combine any picture of truth, with any model of value, and any optimization procedure, which look remotely like those that humans have, the interactions between them would certainly be woven together, and their dynamics would be intertwined.

This is why we model the dynamics of beliefs and preferences together, and why in some cases we will be able to represent a belief as a preference or vice versa[4]. The way we've laid things out, the interaction between the two resembles a duality.

Please cut this.

2.2 Better Models of Humans, whose Preferences Change

When you're born you have no conception of what foods or professions or ideals are good; all you have is your own evolutionarily programmed pleasure and pain. Not only does this feeling get more sophisticated in a way that seems to depend on the environment, but also later on in life, people willingly sacrifice their experiential pleasure for other things they care about. All of this stands even if we didn't see change in the more obvious toy cases: foods, activities, and luxury goods. Such changes do in fact occur, and do not detract from a person's ability to be conceptualized as a coherent agent; often they contribute to a person's character instead.

The standard models do not account for any of this, and for many good reasons — but dynamic preferences are necessary to capture a great deal of human behavior, and are much more important in a world where value is informational and changes quickly. Things like memes, fashion, games, conversations, lofty ideals, and art—things that standard economic theory has had a lot of difficulty ascribing value to—certainly play a bigger economic role than they have in the past, and arguably move faster as well.

2.2.1

What does this mean?

This is why we model changes in preferences, and acquisition from only a very limited set of base values. We have some reasons to suspect that in most cases preferences generated this way will be similar, which if true would (1) provide an additional explanation (beyond empathy and genetics) for why humans end up sharing a lot of values, and (2) allay some concerns about misaligned AIs that are constructed in this way

[4] the categorical interpretation, where each node is a category, the big graph forms a 2-category, and beliefs are value-preserving functors, supports the observation that the distinction between preferences and beliefs isn't a fundamental feature the objects themselves but rather how they're used— here the arrow category here gives us preferences over beliefs. Similarly, quotient objects represented as arrows, just as utilities can be identified with beliefs about utilities.

Text

2.3 Computational Tractability

The standard picture of decision theory requires keeping around preferences and beliefs about all possible things that are relevant to any decisions you make. [todo: *Is there an edit here to be made about small vs large worlds?*] To fully describe a general agent's values, then, you need to specify a preference ordering over all possible histories of settings of observable features. This is wildly intractable, and also annoyingly depends ?? on what "possible" means, which is part of the agent's internal beliefs[5]. This is why people in practice restrict the scope of an agent to only a few modeler-chosen variables, assume that they only encounter one kind of decision, specify objectives in a compressed syntactic form.

But even these more tractable restricted approximations we use do not degrade gracefully: in order to make any decisions at all you have to do expected utility calculations (which could be very expensive depending on how clear the impact of your decision on the world is), and there's certainly no clear way of making use of partial computations under time pressure.

By keeping parts of preferences around in many different forms, attached to different contexts, not only can we immediately re-use them for recent decisions we've made in a pinch, but also reduce the complexity of adding new nodes. This is because we can split the computation into meaningful chunks (one for each preference domain (node) we can connect a decision to).

[todo: *mention related work on tractability: BN's, Markov Networks, Markov Networks*]

2.4 The Value Of Inconsistency.

There are still many tasks human-coded expert systems excel at, which are difficult for trained statistical systems. However, they are slowly losing their ground due to a debilitating shared flaw: any input that the designer has not anticipated causes the entire calculation to be wrong in un-salvageable ways. From the perspective of the expert system, though, there is nothing wrong.

?? What program?

The program is specified by *behavior*, It's coded without an outside view of the explicit purpose: there's no possibility that the designer was *wrong* in the algorithm specification, and there's no reasonable inference to make instead even we acknowledged it was possible she was.

The only way you can know if you're wrong is if there is enough redundancy to highlight inconsistency.

Incorporating probabilities and specifying objectives instead of algorithms has helped this problem enormously: by using a ton of redundant, noisy, examples of the algorithm working correctly, we can be reasonably resistant to malformed input. Unfortunately, these systems are not perfect either: they are often biased and sometimes cheat. We face a different kind of brittleness here: what if we forget about a second order effect and slightly mis-specify an objective function? What if we run it on a different input distribution? ML systems won't crash, but they will confidently display wrong answers. The system doesn't think there's any possibility that its *objective is wrong*, and even if it acknowledged this possibility, it's not clear there's anything to be done at this point.

In analogy to the solution that statistical ML provided to the problem of brittle classical AI, the solution would seem to be a noisy, redundant encoding of the objective function in conjunction with a meta-objective: in our case, consistency. The diffusion of preferences (which we will also refer to as value capture) can also be thought of as a source of uncertainty about your values even if you didn't start with any: if there are multiple objectives consistent with everything you've seen, you acquire a preference for both of them.

Of course, humans are also often uncertain about their values (and experience value capture) in addition to their beliefs and the appropriate choice of actions.

[5] Or alternatively, if you're going to do the work beforehand as a modeler, has to change when humanity discovers new features of the universe

[Theoretical Unity] Reductions to other theories and algorithms

[todo: *Edit this*] The generality is a genuine mathematical motivation for this theory: it can be viewed as a number of different things. I will briefly mention some of them that I'm excited about here, and then focus on simple things for the rest of the document.

- Our model reduces to standard expected utility calculations in the degenerate case where the agent has preferences over possible histories of worlds, the representation of a world never changes, and the agent is cognitively unbounded
- It has a categorical interpretation which has some features I'd like to explore: in particular, meta-preferences seem to be related to higher order structure, a total utility function is like a limit, a total probability function is like a co-limit, preference-preserving beliefs are functorial, and the nodes already form a 2-category, whose objects are themselves enriched flat categories.
- There also a natural interpretation of this as an artificial neural network. If the underlying graph is a DAG with one sink, then it is a feed-forward network for calculating expected utility. In most interesting cases, it will have recurrent connections and no special output.
- For agents that have priors, the beliefs encoded in this way can be converted to a Baysian Network with a simple transformation
- The preference propagation is a bit like broadcasting on a network (of physical machines) to compute path lengths. This can be done Dijkstra / Distributed Bellman Ford [6] and can be computed by taking iterative powers of a giant matrix over the tropical semi-ring. Funnily enough, this is exactly how you compute the transitive closure of preference matrices on a small scale (except with a different semi-ring), which lends more credibility to the higher order categorical interpretation.

You're welcome to do this, but I would make it low priority …

I'm not sure what you're talking about here.

The possibility of bridging these many disparate fields makes the abstraction feel much more universal, and the ability to think about preference change in any of these terms could make it easy to identify analogs of non-trivial facts in other fields.

3 Desiderata

Perhaps among other things, any model of preference dynamics that is to be used to build synthetic agents which have control over important decisions, should have the following properties:

D1. Provide an answer to the 'value loading' problem: show how you can learn "reasonable" preferences by interacting with the world
D2. Reduce to static models for some parameter settings
D3. Behave reasonably when combined with changes of perspective
D4. Be resistant to standard challenges to irrationality, such as dutch book arguments
D5. Have weak safety guarantees: an agent should not eagerly adopt preferences which are totally in conflict with its current ones

Because the view of preferences we adopt here is different from the standard ones in economics (in particular, it lends itself naturally to incorporating boundedness), we have a hope of explaining some behavior which was previously classified irrational, as an optimal in some sense. The following psychological effects lend themselves to explanation:

1. **Value Capture.** You really care about X (say, learning math), which is vague and hard to measure, so you come up with some metric Y (say, scores on undergrad math exams). Over time, optimizing for Y will cause you to optimize less effectively for X and assign intrinsic value to Y (getting good

[6] depending on whether we take right or left powers of matices

exam scores becomes valuable, independent of whether you learn math[7]). Related to Goodhart and Cambell's laws.

I don't know these laws. The story here, I presume, is that you want to model a certain type of preference change.

2. **Framing Effects.** It is well-documented that the style of presentation, even for logically equivalent scenarios, can have a significant impact on a person's choice. But if we formalize these as the same outcome, there's no way for utility-maximizing agent to behave this way.

That's true, of course. But what in your framework allows us to view this as different outcomes.

3. **Connoisseur Effects.** Someone who listens to a lot of rap music has more nuanced, complicated preferences on rap music than someone who has not heard as much. Similarly, people work to develop palates for wine, and "all Indian food tastes the same" is an insult, indicating a shallow experience with the cuisine. There are two technical aspects: first, additional experiences increase preference complexity.

See below.

4. **Adaptive Preferences.** Even things we find objectionable are normalized over time, and people often change to prefer things they're used to, even if they are initially opposed. This is sometimes thought of as a prioritization of safety, and is maybe best thought of as a thought-saving feature: the things you're used to have gotten you this far already.

I wouldn't bring in safety. Moving up a level, you still need to make it clear why your framework will be better at capturing this than other.

5. **Novelty Effects (anti-adaptive preferences).**

"Increase in preference complexity" just seems to me an outcome of using more features to describe the situation. This can be done using the standard framework. What isn't clear to me is why your framework makes it easier to do this.

[7]Social signaling plays into this, but this occurs also in cases where people try to hide the signal: constant grade checking, pokemon go addiction, etc.

Part II

Probabilistic Constraint Graphs

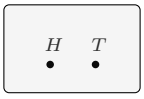
The basic tool that we will use to solve these problems is a graphical representation of (a piece of) an agent's mental space, which we call a marginal constraint graph. Throughout this section, we will introduce them from many different perspectives, but They look a bit like Bayesian Networks, but are more general: rather than representing a single distribution, they represent (soft) constraints on distributions.

- If the constraints pick out a single distribution, the PCG is a graphical factorization of that distribution.
- If the PCG is under-constrained, there are many distributions consistent with it. We can use this to emulate weaker notions of uncertainty than probability distributions themselves. However, not all such distributions are equal: some may be simpler than others, or better match the symmetries of the constraints; we use entropy to distinguish between them.
- An over-constrained PCG is has internal inconsistencies, the resolution of which drives belief and preference change. Just as in the under-constrained case, not all resolutions are equal: some are more or less constrained

4 Uncertainty Examples [todo: *fold into other sections*]

4.1 Coin Tosses

Example 4.1. You're about to flip a coin. The result of this flip is the only thing you're thinking about, but you know nothing about the outcome, other than that it could land heads or tails. This (lack of) knowledge is the single node in an PCG:

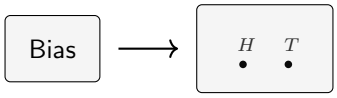


Coin

One might think this is vacuous, but in fact this conveys two important things: that the agent is aware of and conceptualizes the coin flip (with the two outcomes), and a total uncertainty about what values they actually take and how they relate to anything else.

The fact that this represents total uncertainty depends on the semantics, which you haven't yet discussed.

Example 4.2. Now you start reasoning about this coin, and decide that the outcome of the flip is determined by some intrinsic feature of the coin: its bias; you'd be willing to assign a probability to the coin flip if you knew the bias. We illustrate this with the following diagram:



Coin

where Bias is some value in $[0, 1]$ and the arrow P is given by

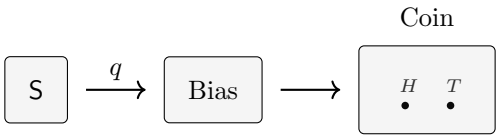
$$P : \mathsf{Bias} \to \Delta\{H, T\}, \qquad P(b) = \begin{cases} H & \text{with probability } b \\ T & \text{with probability } 1 - b \end{cases}$$

Only if you have an underlying state space; remember, a random variable is a function from states to the reals

Bias is now an infinite random variable, and functions as a higher-order probability, which has been "internalized" to the picture. Doing this again may seem vacuous, as we have not gained any information

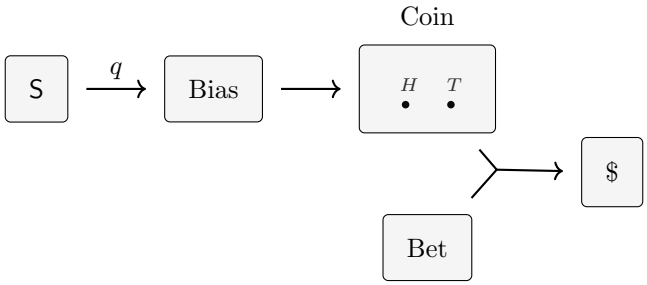
about whether the coin will land heads or tails, but once again we get information about what the agent thinks is relevant. Putting this node in the space can make otherwise innocuous changes suddenly meaningful — for instance, if Bias had only a finite selection of possibilities, a granularity perhaps imposed by the agent's way of representing numbers, then the agent has given themselves information about the coin through arbitrary choices. △

Example 4.3. We will take our coin example a bit further. Already in the previous case, the set of biases is consistent with some set of distributions. Suppose that our agent continues to think and realizes they consider some set of distributions over bias possible. Now that they are aware of it, we give it a name (S), and put it into the picture:



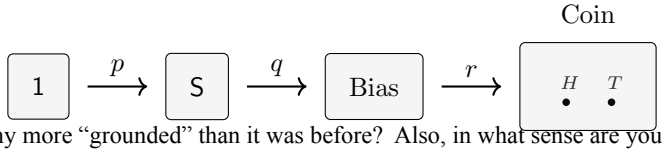
The arrow q represents the conditional distribution μ_s on Bias, for $s \in S$. This example illustrates that not only can these PCGs represent sets of distributions with some external interpretation, but this can also be internalized to the PCG itself, and so the kind of inference that can be done externally with lower probabilities, for instance, can be done just by the the prescribed decision theory. ??? I don't understand what you're saying here

For example, if $S = \{\mu_B : \mu_B(H) > 0.5\}$, i.e., you are certain that the probability of heads is at least 0.5 (maybe you know the company which makes it, which only weights the head side, but can do any amount), and you are trying to decide whether to bet that the coin lands heads or tails, you now have the following picture:



From this, even though I have no prior on S, I know that no matter what distribution on bias is actualized, an even bet for heads nets more money than an even bet for tails. △

Example 4.4. Finally, we can also add weights, which we can use to reason about higher order probability, although we need to be careful about normalization and our arrows will either need to be interpreted as non-standard measures, or we have to make some extra assumptions in the general case. Intuitively, the picture looks like this:



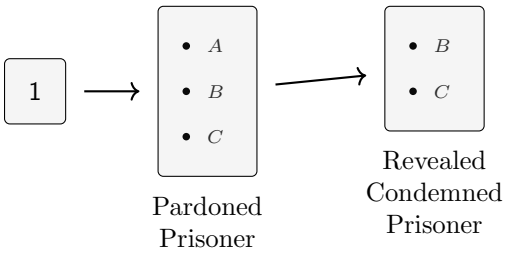
In what sense is this any more "grounded" than it was before? Also, in what sense are you adding "weights" here?

Now we finally have grounded the chain so that it has the unique node 1 as a root of a tree, and so by Proposition 8.1, there is exactly one distribution consistent with it. We can see that the higher order probability is still represented, and modifications to the arrows which represent more abstract concepts do change the rest of the problem setup — but at the end, we can use composition (which here is acting as the bind of our probability monad) to squash all of the information a probability. ??? △

I have no idea what this means, and don't want to know now. *Please* remove all category-theoretic language. Trust me; adding it will only hurt the paper at this point.

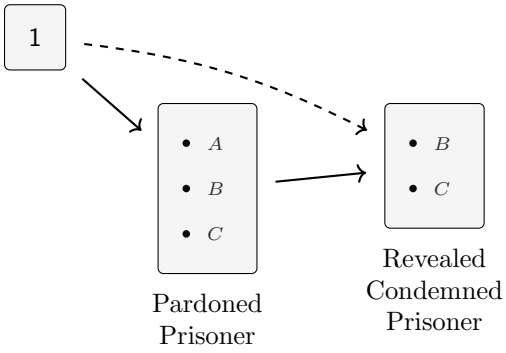
4.2 Pardoned Prisoner

Example 4.5. We take this example as well from the Reasoning about Uncertainty book. Suppose there are three prisoners, A, B, and C. One of them has been chosen at random to be pardoned. Prisoner A asks the executioner which of the other two was executed. Just before she gets her answer, the agent's PCG looks like this:



Note that in order to tie together the two nodes with an arrow, we need to either know the probability of each prisoner being pardoned for each choice of revealed condemned prisoner, or more naturally, the probability that the executioner reveals each prisoner, given the identity of the pardoned one — that is, we need to model the protocol of the executioner. If we assume the executioner is truthful, we still need to know what happens when A was the pardoned prisoner and the executioner has a choice. Suppose that in this case, B is always revealed. This results, through composition, in a prior distribution over the revealed prisoner, with $\frac{2}{3}$ on B and $\frac{1}{3}$ on C.

Once we make our observation, we get a new arrow; suppose we hear that B was revealed.



There is now an inconsistency: the two paths from 1 give different distributions on the revealed prisoner. There are many ways to ameliorate this:

1. The agent can ignore what the executioner told her, because it conflicts with your a-priori distribution on the probability of what the executioner would say
2. The agent can revise their view of the executioner's protocol upon hearing B, and now believe that the executioner would have told her this no matter who was pardoned.
3. The agent can change their belief about how likely each prisoner is to have been pardoned. In this case, she can conclude that her chances of dying are now 1 in 2. Of course, if the executioner had said C, she would have had to conclude she was guaranteed to die.

△

Example 4.6 (Prisoners and subdistributions)**.** Maybe we actually do not know the guard's protocol. We can still interpret the arrow as a sub-Markov transition / sub-stochastic matrix, where now we do not know what happens when A is the pardoned prisoner. In this case, we still can compose the arrows, and obtain a sub-distribution $\Pr_*(B) = \frac{1}{3}$, $\Pr_*(C) = \frac{1}{3}$. Upon hearing the name of the revealed prisoner, we cannot update, and still have the same sub-distribution. △

???

???

---

5 Definitions and Semantics

Definition (PCG, semi-formal)**.** A marginal constraint graph $(\mathcal{N}, \mathcal{L})$ is a collection of variables $\mathcal{N}$, 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 $A \to \Delta[B \cup \{\bullet\}]$[8] representing beliefs about how a setting of A to $a \neq \bullet$ will impact the value of B.

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 (which is not fixed) into disjoint (but possibly not exhaustive) events. Disjointness here is very weak and can be forced by adding tagging outcomes with additional data, analogous to a disjoint union.

5.1 Partial Interpretations

By analogy to the distinction between a qualitative and quantitative Bayesian Network, we can make a distinction between a sources of information in in an PCG. We can talk about the shape of the PCG — topology of the graph itself, before any numbers have been included— in addition to its interpretation in a probabilistic setting. One can imagine also interpreting them with other models of uncertainty. I will also remark that the effects here are captured succinctly and with clearer relations to other interpretations (e.g., of databases, semantics, and diagrams), by viewing a PCG as a functor; see section 13.2 for details.

5.1.1 Qualitative PCGs

The connectivity a BN (as well as of a Markov Network or factor graph) has something to say about independence — that if two nodes are not connected, then they have to be independent, if we fix some separating variables. This is a very strong statement. Intuitively when building a graph like this, if I have not connected two nodes, it is because I don't really know how they interact, not because I'm certain that the probability distribution separates when we fix some middle node along a path.

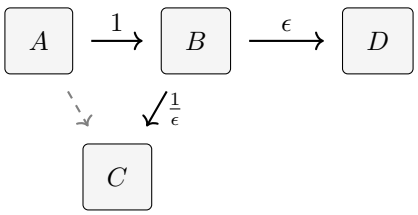
For us, an arrow $A \to B$ indicates that you'd be willing to bet on B (and believe you'd be rational in doing so) if you knew the value of A. Unlike the other graphical models, such a structure does not rule out any joint distribution on the nodes, which makes them less useful for factoring a unique distribution.

Still, the connectivity alone is meaningful, as it determines:

- How over / under-constrained the agent's knowledge is
- What variables the agent thinks are most robustly predictive of which others
- Which beliefs are most redundantly held

If we're also given access to the entrenchment weights on the arrows, we can use the connectivity to determine which beliefs are likely to change when observations are made. For instance, if observations are made on C via Jeffrey's rule for each value of A, we can infer an edge $A \to C$ in the picture below:

[8] 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 8.2 for details



As a result of this observation, we can infer from these and the connectivity that the belief $A \to B$ will change, rather than the one from $B \to C$. The connectivity is important; note that the belief $B \to D$ will not change even though it is much more weakly held.

5.1.2 Quantitative PCGs

Filling in all of the numbers, we get a quantitative PCG the PCG gives us the object we've been talking about until now, with families of distributions on each edge.

5.1.3 Partially Quantitative PCGs

We can also interpret arrows part way. By giving ourselves access to the amount of uncertainty in the arrows (e.g., the entropy, or a whether or not a given arrow is deterministic), we can draw additional conclusions.

We can also interpret only a subset of the arrows, which places (soft) constraints on the remaining arrows.

5.2 Semantics

The semantics of a quantitative PCG

5.3 Formalism

We can now give a more formal definition:

Definition 5.1. A probabilistic constraint graph (PCG) is a tuple

$$\left(\mathcal{N} : \mathbf{FinSet}, \quad \mathcal{L} : 2^{\mathcal{N}\times\mathcal{N}}, \quad \langle \mathcal{S}, \Sigma\rangle : \mathcal{N} \to \mathbf{MeasSet}, \quad \mathbf{p} : \prod_{(A,B):\mathcal{L}} \Big[\mathcal{S}_A \times \Sigma_B \to [0,1]\Big]\right)$$

where $p(L)$ is a Markov kernel, i.e., for every $L[A,B] : \mathcal{L}$, and $a \in \mathcal{S}_A$, $\mathbf{p}_L[a \mid \cdot\ ]$ is a probability distribution on $(\mathcal{S}_B, \Sigma_B)$, and $\mathbf{p}_L[\ \cdot \mid B]$ is Σ_A-measurable for every $B \in \Sigma_B$.

6 Recovering the Standard Picture

6.1 Probability Spaces

6.2 Random Variables

6.3 Preferences, and Utility

Suppose the only thing I conceptualize is a single variable A. A preference on the alternatives $a \in \mathcal{S}(A)$ often takes the form of an order. If it is a total order (and $|A| \leq |\mathbb{R}|$), then it can be represented as a utility

function. We can capture this explicitly with our model: think of utility as a separate preference domain $\mathsf{U} \cong (\mathbb{R}, \leq)$, whose elements are the real numbers, with a preference given by the usual ordering. Now a choice of A has an impact on what happens in U, and if we know that each choice of A gives us a deterministic utility, then this impact is actually just a function, and in particular can be represented as a degenerate probability distribution $\Pr(\mathsf{U} \mid A)$, which is just a function from A to $\mathbb{R}$.

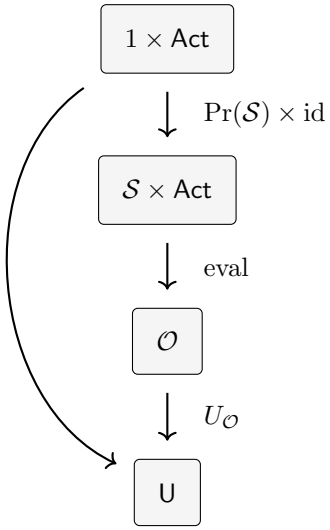
$$\boxed{\mathsf{A}} \xrightarrow{U_A : A \to \mathbb{R}} \boxed{\mathbb{R}, \leq}$$

In this sense, we can think of utility functions as a way of representing preferences on A as a belief about how A impacts a standard universal preference domain of "goodness". Similarly, if we have multiple variables, we could always just take their product all of their variables, and consider a conditional probability distribution from this to the utility preference domain U.

$$\boxed{\mathsf{A}_1 \times \mathsf{A}_2 \times \cdots} \xrightarrow{U} \boxed{\mathbb{R}, \leq}$$

6.4 Savage

Savage's theory can also be represented in this context. If you have a preference order on Act, which is the set of all maps from states $\mathcal{S}$ to outcomes $\mathcal{O}$, which obeys Savage's postulates P1-P7, then it can be factored into a probability distribution $\Pr(\mathcal{S})$ and a utility function $U_{\mathcal{O}}$, as shown on the right.



7 Relation to Bayesian Networks

7.1 Differences

These are diagrams which look and behave like Bayesian Networks in many cases, but are more general. Here are some structural differences:

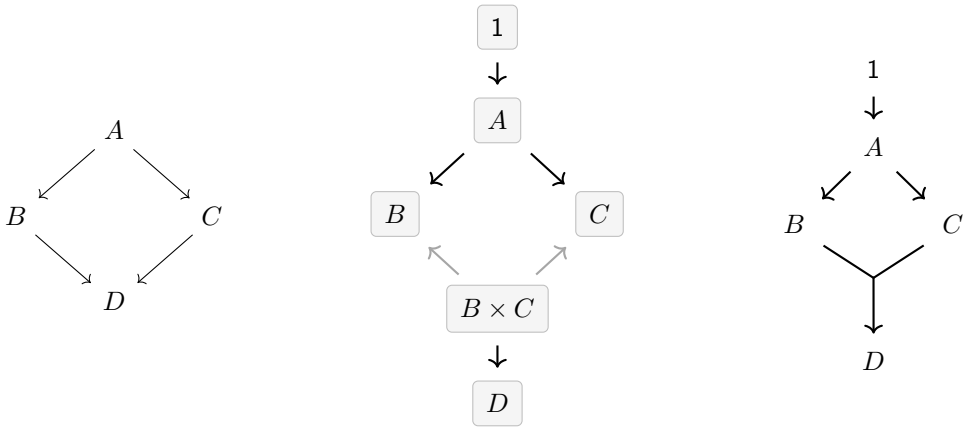
- We interpret each arrow as a conditional probability table in isolation, rather than considering all arrows into a node together as a larger conditional probability from the cartesian product. This gives us a notion of composition, which we will defend later.
- Similarly (this is the zeroth order case of the above), we do not assign distributions to nodes without parents.
- We make no explicit assumptions about independence (except insofar as the arrows encode this). We achieve separation in a weaker way, by giving preference to maximum entropy distributions.

7.2 Converting BNs to PCGs

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 $\text{Par}(N)$, and a conditional probability distribution $\Pr(N \mid \text{Par}(N))$, which is a distribution over the values of N for each setting of every variable in $\text{Par}(N)$. 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.

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 $\Pr(N \mid *) = \Pr(N)$.

This sound much more complicated than it is. Consider the example below, where the left is a BN, the center is the corresponding PCG, and the right is a compact visual representation of the PCG in the case where $B \times C$ has no other edges attached to it directly.



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 an PCG $(\mathcal{A}, \mathcal{L})$ 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 model to be related, without ever conditioning on ancestors, which include products[9]. 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. We also get:

[9] 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.

Theorem 7.1. *The center of the PCG 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.*

We will clarify this definition and explain the connection to thermodynamics more carefully (section **??**) 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.

7.3 A More Formal Treatment

Definition 7.1. A Baysian network (BN) is a tuple

$$\mathcal{B} = \left(\mathcal{N} : \mathbf{FinSet}, \quad \mathrm{Par} : \mathcal{N} \to 2^{\mathcal{N}}, \quad \mathcal{S} : \mathcal{N} \to \mathbf{FinSet}, \quad \Pr : \prod_{N:\mathcal{N}} \left[\mathcal{S}_N \times \left(\prod_{P:\mathrm{Par}(N)} \mathcal{S}_P \right) \to [0,1] \right] \right)$$

such that

- the graph $\bigcup_{N,P \in \mathrm{Par}(N)} (N, P)$ is acyclic, i.e., there exists no cycle of nodes $N_0, N_1, \cdots, N_k = N_0$ in $\mathcal{N}^k$ such that $N_{i+1} \in \mathrm{Par}(N_i)$ for each $i \in \{0, 1, \cdots, k\}$.
- For all $N \in \mathcal{N}$, $\Pr(N)$ is a probability distribution on $\mathcal{S}_N$, i.e.,

$$\forall N \in \mathcal{N}.\ \forall \vec{p} \in \prod_{P:\mathrm{Par}(N)} \mathcal{S}_P.\quad \sum_{n \in \mathcal{S}_N} \Pr_N(\vec{p}, n) = 1$$

8 Exploration of PCG Properties

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

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

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.

8.1.1 Should paths be equal?

This design decision also has the effect of proving multiple different ways to calculate things, which leads to the somewhat counter-intuitive fact that that not all diagrams commute (this just means that the path to calculate a quantity does not matter), even ignoring preferences entirely. Often this can create inconsistency which drives preference change; other times there is no direct consistency cost (though there is still an entropic

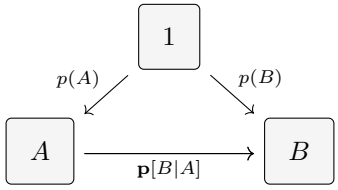
one; see section **??**) to pay. We will begin by giving some intuition about why one might expect paths to be equal.

1. Markov processes, which are the more common place stochastic matrices appear, only have one path, so everything is path independent. Also, most anything can be modeled as a Markov process, if you put enough information into your state. Unfortunately, to do this, we require control over the state representation, and we don't want to contaminate our variables by stuffing a bunch of information into them for the sole purpose of reconstructing all of our knowledge from anywhere.
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 (see proposition 8.2).
4. If everything were purely probabilistic, all diagrams would commute (see proposition 8.1).
5. Even in our setting, in many cases this must be the case, and in fact several axioms of probability can be expressed as requirements that large classes of them must.

Example 8.1 (Marginalization)**.** Recall how a probability can be obtained by marginalization:

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

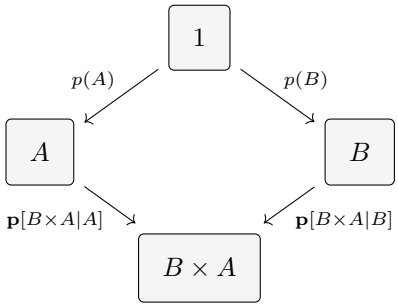
below is an illustration of this fact:



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, one more example:

Example 8.2 (Dinky Bayes Rule)**.** We can also represent a dinky version of Bayes' rule, $p(a \mid b)p(b) = p(b \mid a)p(a)$ as an assertion that two paths from:



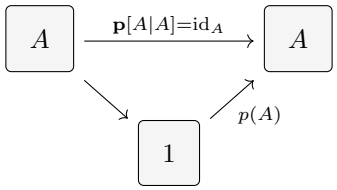
This can be seen as two applications of marginalization, one for each half. On the left, we have

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

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

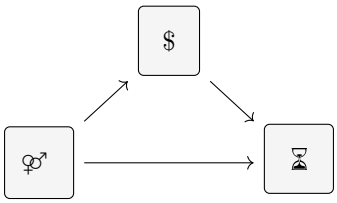


△

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 "the composition of paths depends only on their endpoints". Still, there are some blatantly inconsistent pictures one could draw — anything that violates Bayes' rule or marginalization, for example (see section 16 for more).

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.

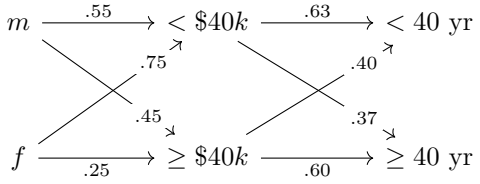
Example 8.4. Men earn more than women, and people who earn more are generally older, but women live longer than men, so the top composition in the picture below



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) ⚥ → $ → ⌛:

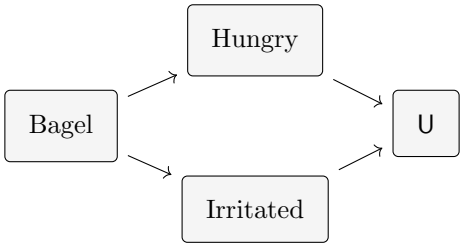
	<40 yr	≥40 yr
♂	.53	.47
♀	.57	.43

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.

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

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 8.5. 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 8.5 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 the absence of additional information, the composition of two links is in some sense the best available estimate, in that it is compatible with any distribution in the center of the PCG, and therefore (as measured by relative entropy) is at the centroid of all other possible links.
- 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, or getting the same advice separately from two experts.
- In many other cases, the composition is guaranteed to equal or approximate the true probability — for instance, if you have bounds on the entropies of your links, or have some independence assumptions somewhere.

8.1.2 Some Results

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 8.1. *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 8.1, 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 8.2. *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 $h \circ f = g$.*

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

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

□

,

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 8.4 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 8.3. *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.

8.2 Sub-stochastic Transitions and Conditionals

Sometimes an otherwise very useful variable might not apply in a small percentage of cases; in this case, we want a way of putting all of the extra probability mass in a "something else happened" bucket, giving us effectively a sub-stochastic matrix, or a a lower probability on singletons. For instance, 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 "not applicable / something else" bucket.

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— this is problematic, as a big part of the reason we've been using links to avoid assigning probabilities to everything. 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 is not applicable. Why might it not be applicable? 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?".

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

$$p(B \mid A) = \begin{matrix} \\ a \\ \bar{a} \end{matrix}\!\!\begin{matrix} \begin{matrix} b & \bar{b} \end{matrix} \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{matrix} \quad \text{and} \quad p(B \mid A) = \begin{matrix} \\ a \\ \bar{a} \end{matrix}\!\!\begin{matrix} \begin{matrix} b & \bar{b} \end{matrix} \\ \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{matrix}$$

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.

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

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. A probability distribution alone, and by extension a standard Bayesian network cannot do this — because we require the look-up tables to exactly match all possible values, we can't drop any without totally giving up on any world which looks like that.

8.2.1 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*]

9 Arguments for PCGs

9.1 The Possibility of Types and Embedded Logic

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. Even leaving that aside, there are some obvious ways one might want to combine existing domains; the primary one we are concerned with here is products.

In terms of beliefs, you might already have beliefs about how likely two events A and B are to occur, 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:

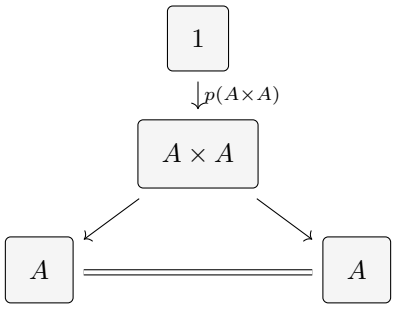
$$A \xleftarrow[\pi_A]{} A \times B \xrightarrow[\pi_B]{} B$$

It may be worth noting that we can always construct $\langle f, g \rangle$, but uniqueness in general does not hold[10]. This is important because without any additional assumptions, knowing a distribution on A and a distribution on B does not allow you to infer a distribution on $A \times B$; there may be correlations. However, this is still a unique maximum entropy one, which assumes that the two are independent.

9.1.1 Products Variables vs Unions of Variable Sets

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.



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 8.2 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.

There is also a good argument that both should be allowed if the agent can unify variables, which we can leave for a future discussion.

[10] so it is not (yet) a categorical product

9.1.2 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)*]

9.2 Human Beliefs as Marginals

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.

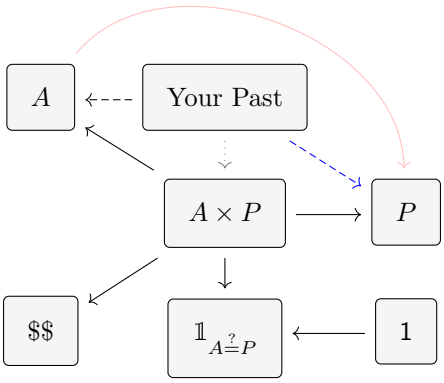
Example 9.1. After reading a number of empirical studies, you come to believe that smokers have a 70% chance of developing cancer, compared to 20% for non-smokers. At the same time, you believe that those who use tanning beds have a 80% chance of developing cancer, compared to 18 % for those who do not use them. You have no information about how the two interact. △

Example 9.2. 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 9.3 (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 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 $\to A \times P$.



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 $A \times P \to \mathbb{1}_{A \overset{?}{=} P}$). On the other hand, you also have an outside view of how likely the predictor is to be right (the arrow $1 \to \mathbb{1}_{A \overset{?}{=} P}$). A belief that your actions influence the prediction would correspond to the pink arrow above.

We have not resolved the paradox; we have merely represented the beliefs in the setup, which make very little sense as a BN. The point is that to encode interesting scenarios with potential conflict, it makes sense to use these graphs, even though there's not a clear way of encoding all of the information in the scenario with a BN or factor graph.. △

9.3 Human Preferences as Marginals

Beliefs are part of the decision theory picture, at least classically; we also need to deal with utilities. Rather than treat utilities as a special function from $\mathcal{W}$ to $\mathbb{R}$, we will introduce a new utility domain U, which is subject to all of the same constraints as we've seen in previous sections. As a special case, when one explicitly has the product of all variables (save for this new one) as a special variable $\mathcal{W} \in \mathcal{N}$, and its value totally determines U with no uncertainty, we are back to the classical case. However, in general, we will allow any variable to have an arrow to U, which is not deterministic. These could be calculated as expected utilities through $\mathcal{W}$ and the utility function if we had one, but for us, these will be primitive.

As such, we do not make a distinction between a 'pure' utility and an 'expected' one; much of this section is devoted to making the argument that this makes more psychological sense. Clearly from a mathematical perspective the two behave differently: the pure utilities can be used to make sure the whole picture is consistent in the classical setting, and there is one global utility across time which changes in beliefs do not alter, even if it is filtered differently through different lenses at different points in time. I view both of these properties as a poor fit for modeling humans.

Example 9.4. 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.

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. △

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 sensation) should be treated in context.

Example 9.5. 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? If you only discovered that this were secretly true recently? Would you change your preference for free society?

The point is that the preferences that an agent forms are constrained by experiences and the other beliefs they have, and are therefore subject to revision. Moreover, this still happens with important things and abstract concepts, in addition to concrete sensations. △

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; at best, we have some notion of good (maybe even more than one) that we can estimate from other quantities, and use to make decisions. We can get a more satisfying context dependence by associating them to separate domains which apply only at certain times.

Even just collapsing the distribution of "good" to its mean to make a utility function is not clearly a reasonable thing to do— the optimal risk calculus when dealing with a deterministic utility function, for instance, is locally clear, because you know how good worlds are. The possibility of feeling bad even though the state of the world is exactly what you were aiming for, almost never accounted for, and yet ennui and purposelessness are very real.

Before we get back to looking at the world as a collection of conditional distributions, I want to offer one final example that suggests that utilities in humans are more like constraints than additive components of some utility function.

Let X be a variable, and suppose the utility of x_1 is extremely high. If we also have a view of how good other unrelated things are in a different domain, then we can conclude that x_1 must be uncommon. This kind of reasoning might feel backwards — we're using our preferences, which are supposed to be internal, to infer something about the world — but if our preferences interact with our beliefs, this makes a lot more sense. It seems people actually do this: as things become more common, they cannot be valued as highly. An explanation for the hedonic treadmill falls right out of expressing preferences through multiple constraints, rather than a single utility function.

10 Co-algebraic Structure

Please don't spend time on Sections 10-13 for now. This is not where your energy should be going!

So far, we have not made a big deal out of our extra element • that we have included into every domain, but the fact that we have exempted ourselves from giving it a distribution turns out to dramatically increase the representational capacity of the model. Of course, we will be able to say less about this larger class of models, but many parts of the story will translate cleanly.

10.1 Bundling and Unbundling

To begin, by relaxing the restriction on our links from stochastic to sub-stochastic matrices, we have gifted ourselves the ability to un-bundle and re-bundle variables back together. One should think of this as a way of building "left-handed" or inductive constructions, rather than "right-handed" or co-inductive ones. This will allow us to represent game trees and automata, as well as

[todo: *DIAGRAM 1*]

10.2 Branching

11 Reduction: Fragment of Factor Graphs

12 Reduction: Game Trees and Automata

13 Category Theory

We can also define these structures more compactly (and in my view, cleanly) with a categorical notation, perhaps shining some light on how the different definitions fit together, justify some design decisions, and import a giant mathematical toolbox which makes some features completely obvious. If you're sold already, you can skip next bit (section 13.1) in which I attempt to give a more complete justification, and if you are not willing to parse abstract nonsense, you can skip this section entirely.

13.1 Why Use Category Theory?

13.2 Categorical Definition

We'll start with the fragment of

Definition 13.1. A *categorical PCG* is a diagram in **Mark** of shape $\mathcal{A}$ — that is, a functor $\mathcal{A} \to \mathbf{Mark}$, where $\mathcal{A}$, thought of as attention, is (the category generated by) the graph whose nodes and edges are relevant features of the problem setup.

This makes sense, as these topological graphs

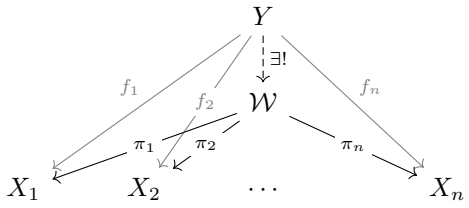
13.3 Limits

We can use this definition to

Definition 13.2. A *world object* of an PCG $G : \mathcal{A} \to \mathbf{Mark}$ is a weak limit of G — that is, a cone over G, for which there exists a (possibly non-unique) arrow

Example 13.1. If $\mathcal{A}$ is a discrete category, with n elements, then PCGs of shape $\mathcal{A} = \{\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n\}$ is just a set of n names (with identities), to be interpreted by M as random variables (with their identities). If

$M : \mathcal{A} \to \mathbf{Mark}$ is a (weak) final cone, as illustrated below.



We claim that $\mathcal{W}$ is just the standard product of measurable spaces, and each π is a projection. The reason for this can be seen information theoretically — clearly the product of all of the variables with projections is a cone over the M, since there are no non-trivial arrows in $\mathcal{A}$ so no equations must be satisfied.

The only sticking point is the, possibility that it's not △

This has some huge advantages over defining a world up front. First of all, we don't need to describe the set of all possible worlds globally and in a way that people can agree on.

We do not get the problem in 1.2.1.

Example 13.2. [todo: *projections*] △

Example 13.3. [todo: *filtered limit*] △

Conjecture 13.1.

13.4 Bundling and the Category of Elements

13.5 Meta-links and Higher Structure

13.6 Denotational Semantics

Part III

Values on Probabilistic Constraint Graphs

The first strand of this project is a general representation of many ways that have been historically used to represent values: namely, utilities, goals, desires, preferences, choice functions, and natural extensions of these suggested by this representation, some of which are also well-studied.

14 Setup

The general idea is that to attach some additional preference data to each alternative in a domain

14.1 Definitions

15 Reductions

The standard tool to represent values (and the ones that people are most familiar with) are orders, i.e., flat categories.

The simplest non-trivial order is:

$$\mathbb{B} = \boxed{\overset{0}{\bullet} \prec \overset{1}{\bullet}}$$

15.1 Desires

I want φ, is

One standard logical approach would be to give a Kripke model, so that a statment like $s \vDash W_\alpha \varphi$, for instance, would be true when agent α wants φ in state s. In such a model, there is already machinery for talking about the set of all possible states (call it $\mathcal{W}$), and so effectively we have specified a function $W_\alpha \varphi : \mathcal{W} \to \mathbb{B}$, which assigns 1 to worlds where α wants φ and 0 to the others.

This is the behavior we would like to capture, but rather than use the set of possible worlds $\mathcal{W}$, we will adopt

15.2 Utilities

A utility function $u : \Omega \to \mathbb{R}$ is a $\mathbb{R}$-value on the set of outcomes Ω.

Effectively,

15.3 Preferences

We can also consider orders which are not total; a discrete order encodes an inability to compare any of the options, lattices encode a possible inability to compare individual options, but provide a way to formalize the combination the best and worst features of two alternatives.

An order is a function $\leq : \Omega \times \Omega \to \mathbb{2}$

Part IV
Dynamics

In addition to the representation itself, we want to prescribe . In the case where an agent is Bayesian

16 Consistency

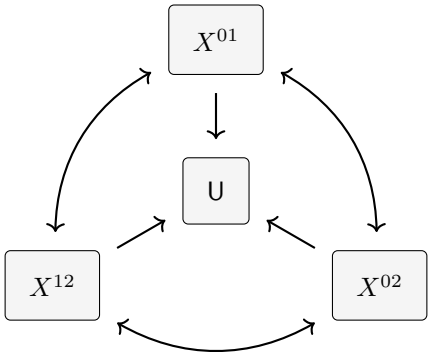
We will start with a simple predicate definition of PCG consistency, of preference consistency, and give some examples before stating a less-brittle continuous generalization that will allow us to deal with inconsistent models. Then we will discuss an alternate definition, which we call link inconsistency, which appeared in previous formulations, and discuss the connection.

Definition 16.1 (consistency)**.** A PCG $(\mathcal{A}, \mathcal{L})$ 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 16.2. A PCG $(\mathcal{A}, \mathcal{L})$, 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, $\big(\mathcal{A} \setminus U, \quad \mathcal{L} \setminus \{L : \square \to U\}\big)$ is consistent.

16.1 Intransitivity of Preferences as Inconsistency

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 $\bullet$ 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:



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

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

And more concretely still:

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

The other three are the transposes of these. Note that without the utility node, the PCG 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,

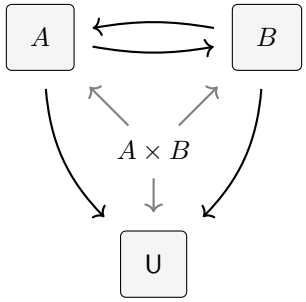
$$\begin{aligned} u_0 := \int_{u:\mathbb{R}} u \cdot p(u \mid x_0)\ \mathrm{d}\mu = \mathbb{E}U_{X^{01}}(x_0) \;&\geq\; \mathbb{E}U_{X^{01}}(x_1) = \int_{u:\mathbb{R}} u \cdot p(u \mid x_1)\ \mathrm{d}\mu \\ = \mathbb{E}U_{X^{12}}(x_1) \;&\geq\; \mathbb{E}U_{X^{12}}(x_2) = \int_{u:\mathbb{R}} u \cdot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

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.

16.2 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

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

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:

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

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 9.3), we can quickly get preferences that conflict purely due to beliefs.

16.3 Link Consistency

Many of our motivating examples can be thought of as a "behavioral inconsistency": you value one thing, but your actions are inconsistent with it. Maybe you value sleep, are aware that to get sleep you need to go to bed early, and yet still do something you know is less valuable instead, such as checking social media. Maybe you

Any example where you have to domains like this. [todo:]

16.4 Continuous Global Consistency

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.

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

17 Reduction: Belief Updating

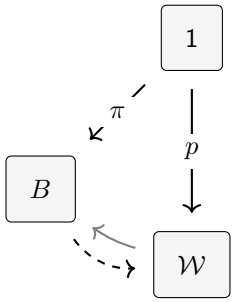
Belief revision, both through Bayes' and Jeffrey's rules, can be thought of as the addition of a new marginal to a distribution, and then a resolution of inconsistency. In Dietrich, List, Bradly [DLB16], a belief revision is an update $p \mapsto p_I$ of a belief state p (in their formulation, these are still distributions) to a new one consistent with the input I.

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

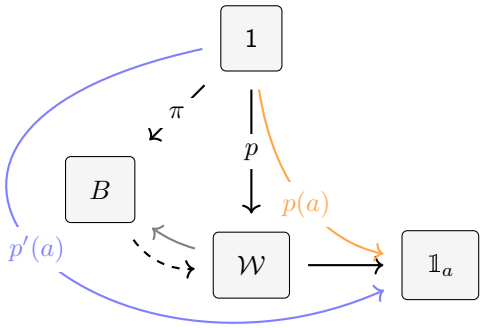
where the variable is binary and the observation is certain. Jeffrey's rule prescribes a posterior probability p' by:

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

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 [DLB16], 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.

18 Abstraction

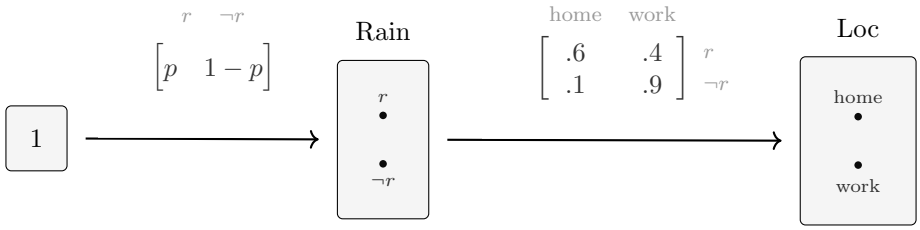
In addition to nodes which represent specific, observable variables that we want to ensure everyone is aware of, we are in the business of allowing agents to construct nodes of their own — examples we've seen so far include product nodes, utility nodes, and other features of the world which become suddenly salient.

Here we want to deal with abstraction nodes, which often arise by approximate factorization of arrows into more manageable pieces.

18.1 Compression

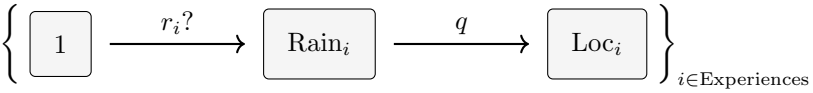
18.2 Divorcing the Specific and General Cases

Example 18.1. Suppose you are more likely to go into work if there's no rain, and you have some probability p of there being rain. A simple linear model might look like this:

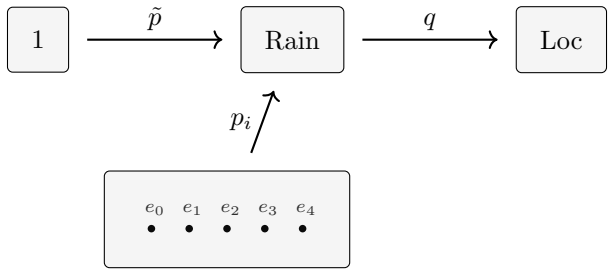


Now, suppose it happens that it's actually raining right now — what happens to the model? The standard answer is to condition, and allow the new, specific information to overwrite the general model. Doing this is a perfectly reasonable thing, but cements a meaning of the node "Rain" that we may not have intended — it now means "it is raining today", rather than in general. This is not a problem for a human modeler, who presumably has fit the model to data somehow, and plans on copying it for each use. There are (at least) two downsides to this. First, we've really just abdicated responsibility to a trustworthy human who will keep track of the copying for us, and handle any sort of abstraction for us. Secondly, it is not clear how and when to update the model based on new experiences.

In the meta-theory, this gives the human in charge an indexed family of variables:



Of course, we can also internalize this to the agent's picture (and given that our formalism is entirely built around conditioning on variables, it seems silly not to) — a picture that might look like this:



Experiences

We store concrete values of whether or not it rained (with possible uncertainty due to lack of memory) in the "Experiences" node. Note that the picture we've drawn below is not expressible with any standard graphical model, because has an arrow merge, which we interpret as a constraint, and so we get non-standard behavior — our model can be inconsistent, for instance, if the general probability of rain $\tilde{p} = 0$ when it has rained.

We gain the additional benefit of

We can also play tricks by bundling and un-bundling this experience variable:

△

---

19 Value Updating

19.1 Informal Examples

19.1.1 Coffee and Beer

Example 19.1. At one time you were a baby, who had never encountered coffee or beer. After numerous good experiences hanging out with friends getting drunk, and several bad experiences being too jittery to talk to people after having coffee, you form a preference for beer over coffee. △

Example 19.2. Suppose you already have a preference for beer over coffee. You now re-examine your preference; you realize that you had not considered the difference in caffeine between the two drinks. You believe that coffee has more, and you're looking to stay awake. As you contemplate and verify that you do indeed believe these things about coffee, you come to weight coffee more heavily. At any point the server can interrupt you and ask for your order and you can tell them what your current preference is. △

Example 19.3. You initially have no preference between beer and coffee. Through a process unknown to you, interesting people are more likely to sit down and talk to you if you have coffee than if you have beer. Your experiences shape your preferences between the two drinks; also, you learn to enjoy the taste of coffee more.

After having formed the preference for coffee, you move to a new place where all of the interesting hip people drink beer, and coffee drinkers sit alone with their headphones. Despite having originally chosen coffee (unintentionally) to make friends, you now value its taste. △

Example 19.4. You had a preference for coffee over beer. You then take some medication which makes you forget things and slightly scrambles your preferences. You now think you prefer beer to coffee. However, as you think harder, you remember that you've always been sick when you've tried beer, and that coffee gives

you energy (which you like). Weighing the evidence, you update and recover your original preference for coffee.

At the same time, if you had instead forgotten some experiences with beer, or whether or not you like being on stimulants, you could use your intact preference for coffee over beer to recover this instead. △

Example 19.5. You have a preference for beer over coffee. You move to Japan. Despite the phonetic similarities, you do not realize that is beer and is coffee. Neither is served out of a container you are used to; for some reason you develop a preference for over .

When the correspondence comes to your attention, you are in internal conflict because you cannot hold both of these preferences and the belief that they correspond all at the same time. Challenging the belief, you ask if maybe this is different beer and coffee from what you're used to, but in fact they show you that they import from your home town, which instead solidifies your belief that the correspondence is correct. Your preference between both beer and coffee, and between and soften as a result, and agree on a more neutral position. To do this, you weigh the more distant but substantial experiences of having beer and coffee in the US, against your confounded and shorter, but more recent experiences in Japan. △

Example 19.6. You live on the border between Utah and Wyoming; you are trying to decide whether to buy groceries in the bigger Utah town (which among other things sells higher quality coffee) or drive across the border, where you can buy beer with a higher alcohol content. As part of this decision, you consider your preference between coffee and beer, which you already had. You do not think about the exploitation of workers or your reminisce about either drink, because you already have this preference available; you decided on a drink yesterday. Also, when you have to go shopping again at short notice next week, you already have a cached preference for what store you prefer.

At the same time, you also have to make decisions about what appliances to buy, and where to go out for eat. Your preferences between beer and coffee also help here, and you still do not think about labor issues while making this decision. △

Example 19.7. You like coffee more than beer, beer more than juice, and juice more than coffee. You're willing to pay $1 to exchange in each case, and you pay a 20th century dutch bookie to make the exchange. Separately, you have a desire not to lose money. Realizing the pattern, you form a desire to have transitive preferences. Taken together your three preferences and your desire for transitive preferences are in conflict. You update your preferences, but the transitivity does not budge much because it's anchored by the prospect of losing unbounded money. You become less sure of what you actually prefer until the three options until your preferences are transitive. △

Example 19.8. You love coffee. So much so, that it's a problem. You've gone to therapy, and nobody seems to be able to cure you of this. However, you also don't like staying up late, an effect which coffee has on you. You experiment, and discover the unfortunate fact that there is no getting around staying up late if you have coffee. While this makes you slightly less excited about choosing coffee over beer, it's pretty negligible, you decide to embrace the effects, and learn to love staying up late. △

Example 19.9. You like coffee more than beer. Maybe even by a lot. Somebody offers you a choice between a coffee that will re-wire your brain to make you want to kill your family, and a beer. You don't take the coffee; why would you? You report them to the police instead. △

19.1.2 Other Examples

Example 19.10. You believe that the rich should not get a larger federal tax exemption (than the poor do) for having children. At the same time, you believe that if the default number of children were 2, and people paid a surcharge to the government for foregoing them, that the rich should pay a larger surcharge (than the poor do) for this.

You come to realize that these preferences are logically in conflict; you then update your beliefs about both of them, causing yourself to be less sure, until the more entrenched one wins, and your viewpoint on the other issue has swapped. △

Example 19.11. Your family lives in Hong Kong. You hate flying; you also dislike the city. But you love your family, and this is an immensely positive thing which out-weighs flying. As a result, you would fly to Hong Kong if (and only if) you could visit your family; the experiences are perfectly correlated and causally linked.

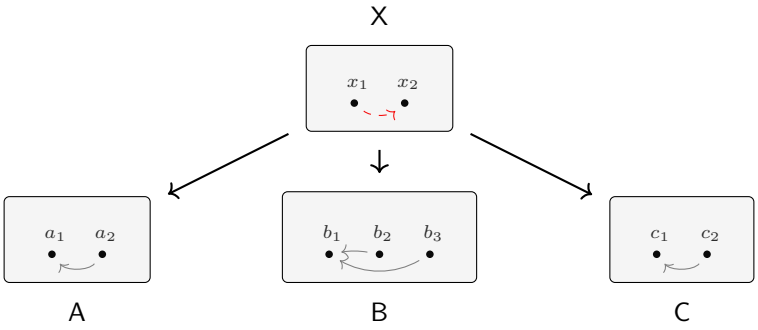
Consider two cases:

1. You have no reason (or have since forgotten) why you hate flying to HK, and discover no new reasons to hate the journey: you get a reasonable amount of work done and the entertainment is just as good as you would have done for yourself. As your preference for seeing your family is in conflict with your dis-preference for flying to HK, but the former is much stronger than the latter (which is unfounded) and so eventually you actually start to like flying to HK.
2. On the other hand, suppose there are good reasons for disliking the flight: it is long, expensive, there's no leg room, people are assholes, and it comes at an opportunity cost. You reach a stable equilibrium where you hate the flight while simultaneously loving your family, even though the two events happen together.

△

19.2 Deductive Preference Formation

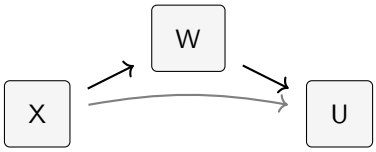
Suppose you're trying to decide what your preference for X is: a standard procedure that humans rely on is the compilation of a (weighted) list of pros and cons. In other words, you examine the impact of a decision in X on other things you care about, say variables $A, B, C \ldots$. Maybe x_1 makes a_1 more likely, which is a good thing, also b_3 and c_2 which are both bad. On the other hand, maybe x_2 doesn't impact A as much, but makes b_1 and c_1 more likely, which is ideal, making x_2 better than x_1. Now, we have external reason to prefer x_2 to x_1 — which means that there's a preference conflict if we were initially indifferent between the alternatives in X! By revising our preferences in X to reflect $x_2 \succcurlyeq x_1$, we can remove the conflict.



Note what we have done: we have a link for how a choice in X impacts downstream things, and so existing preferences flow backwards against the flow of causality, to form a preference on X. This is the standard, deductive way of making decisions.

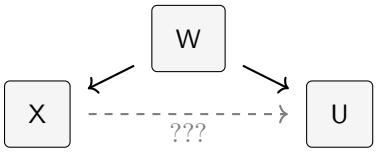
To see more clearly why preferences form *backwards* along links, consider the diagram before, except where we represent preferences are represented by utility functions. If we have a conditional probability $\Pr(W \mid X)$ and a utility function $U_W = \Pr(U \mid W)$, then we can simply compose them to get a utility on X. Note that there is not an easy way to combine a utility function on W and a conditional probability that goes the other direction (see the diagram below).

Easy to compose (deductive formation):



Existing preference $W \to U$ can be extended naturally to X by precomposition, backwards along $X \to W$

Hard to compose (inductive formation):



Getting a utility on X via W requires the inversion of an arrow, or some other additional information

19.3 Learning from Experience

[todo: *This section still needs updating*]

We can use the same technique of reducing conflict to model preferences learned from data. Let Exp be a preference domain representing all of your experiences (which you have preferences/utilities over), and X be some domain consisting of all possible combinations of features (e.g., what food you ate, who you were with, whether you got work done, etc.). Suppose further that each experience had a utility, and also a setting of the features (we can model even the features being fuzzy by making use of the probability distribution half of our links— maybe you forgot what kind of ice cream you were eating).

19.3.1 Classical Models

The standard picture of preferences does not prescribe change, and so learning is generally outside of the scope. Still, there's another reading which makes sense more-or-less classically:

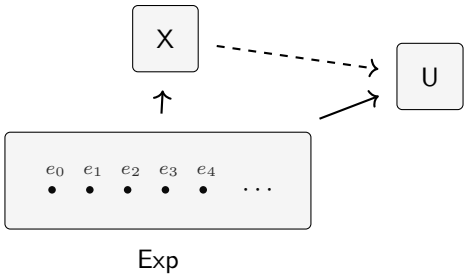
You have some fixed preferences, you just don't know what they are. As you get more samples, you uncover what your preferences must have been all along. The features do not provide any value in and of themselves— they only provide some information about the total state of the world. The experiences are true samples of your fixed utility function.

This can be made to work, but is not a particularly satisfying account:

- Economics outsources the procedure for how to do this to a completely different discipline.
- Humans seem to develop preferences because of experiences, in addition to merely uncovering them — whether or not you have food poisoning the first 10 times you eat mangos, for instance, will probably color your perception of the taste even if you don't get poisoning after this.
- The classical picture also posits that people make decisions according to their preferences, but this is not possible if you don't know what they are — and if it were the case that you were unknowingly making decisions this way, then you could easily access any preferences just by making hypothetical decisions.

19.3.2 Our Description

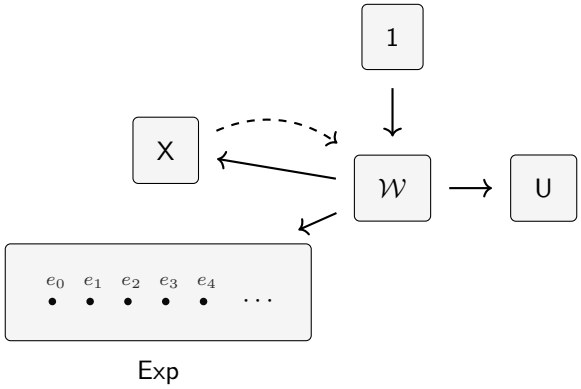
We can succinctly capture the entire setup in this diagram:



To avoid conflict, we would need it to be the case that $\mathsf{Exp} \to \mathsf{X} \to \mathsf{U}$ is the same as $\mathsf{Exp} \to \mathsf{U}$. There are many methods of doing this — we can guess and check, do gradient descent on parameters, set up a fancy network encoding some similarity metrics and priors, ... but in each case, we're just reducing inconsistency, same as before.

It may be worth noting that kernel methods do this in a particular way: they compute a compressed approximation of the inversion of the conditional distribution $\mathsf{Exp} \to \mathsf{X}$. Rather than storing the features for each experience, we can describe features as weighted sums of the experiences they most resemble. Normalizing this weighted sum, we get an arrow $\mathsf{X} \to \mathsf{Exp}$. Having done this approximation, we can then compute utilities for X by precomposition as in the previous section. Also, computing the full pseudoinverse of $\Pr(\mathsf{X} \mid \mathsf{Exp})$, regarded as a matrix, is just a computationally expensive way of computing a least squares fit on U.

The classical picture described above corresponds to this diagram:

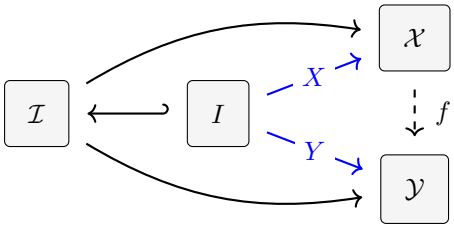


Here $\mathcal{W}$ is the set of all possible worlds, which has a probability measure and utility on it (to the right and above). Really that's all we need, but the world determines which experience you're having and what features as well. You can use the prior over worlds and Bayes' rule to invert the arrow $\mathcal{W} \to \mathcal{X}$ if you really want it, but in any case all you're doing is updating a belief about how X impacts the world so you can compute expected utility better—-you probably don't actually have preferences over X, unless you've been guaranteed that your utility function additively separates over some component of it.

19.3.3 General supervised learning as reduction of inconsistency

The picture we've drawn is actually an instance of a fully generic as a description of a supervised learning problem. In the standard formulation, we're given samples $(X, Y) = \{(x_i, y_i)\}_i$, $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the feature space, $\mathcal{Y}$ is the label space, there is some true oracle function $\hat{f} : \mathcal{X} \to \Delta(\mathcal{Y})$, and therefore (X, Y) is the training set, with $Y \sim \hat{f}(X)$. The problem is then to infer a function $f \approxeq \hat{f} : \mathcal{X} \to \Delta(\mathcal{Y})$. The naive solution is to minimize the training error on the samples, but this leads to overfitting, partly because this implicitly involves thinking of a sample (x, y) as representing "the true value of $\mathcal{Y}$ when $\mathcal{X} = x$ is y" — when in reality there is some noise: both x and y do not represent a probability distribution of uncertainty, but rather concrete sampled values.

We can describe the same scenario equivalently (and arguably more clearly) in our model. If I is the index set that i comes from, then we've been given a function $X : I \to \mathcal{X}$, $Y : I \to \mathcal{Y}$, and learning function $f : \mathcal{X} \to \Delta(\mathcal{Y})$ such that there's minimal conflict corresponds to minimizing training error (inner triangle in the diagram below)



On the other hand, if $I \subseteq \mathcal{I}$, where $\mathcal{I}$ is the set of all possible possible states of the relevant world, then we can think of minimizing generalization error as reducing the inconsistency of the outer diagram[11]: we want it to be the case that f is approximately correct in all cases, not just the ones indexed by our sample I. Bayesian methods make the additional assumption that we have a prior $1 \to \mathcal{I}$, which can be used with Bayes' rule to invert arrows.

This also makes the source of overfitting mentioned above clearer: the training error will also be minimal when we find the best fit, if we were provided X and Y as conditional distributions rather than samples.

20 Thermodynamic Analogy

Appendix

A Category Theoretic Preliminaries

[todo: *Minimal Category Theory needed to understand: categories, functors, diagrams, cones, limits*]

A.1 Markov Category and Giry Monad

A.2 sub-Markov Category

B More Arguments against the Standard Model

B.1 Can't Have Priors on Everything

[11] which is equivalent to minimizing the inconsistency of the outer triangle, because the top and bottom ones both commute because $I \hookrightarrow \mathcal{I}$ is an inclusion