A Theory of Dynamic Preferences

Oliver Richardson oli@cs.cornell.edu

July 17, 2019

1 A Broad Picture of Decision Making

If you already have a (A) model of the world which is detailed enough to describe all possible things that can happen, and in addition you have both (B) beliefs about how the world is likely to evolve in response to any of your actions, and (C) preferences over all possible sequences of events that could happen, then classical decision theories such as Savage's [Sav72] or Jeffrey's [Jef90] gives satisfying answers to (D) decision problems — at least, given enough computational power.

Unfortunately, all three of these components are difficult to get. A perfect world representation (A) seems impossible, but by providing models at a higher abstraction, we can bypass this and proceed with the caveat that decisions can only be as good as the model. Given (A) and some not entirely broken guess at (B), Bayesian updating tells you how to iteratively refine (B) from observations. Representation learning provides some answers about how to refine (A) and (B) at the same time if you already have a task (C) in mind. Model-free reinforcement learning skips the decision theory bit entirely, and uses a reward function (C) to infer actions (D) directly, implicitly refining both (A) and (B) based on experiences.

The work done in search of acquiring preferences (C) preferences has been much less thorough, in part because preferences have always been the grounding used to evaluate the other components of the decision making process. However with so many advances in the other areas of decision making, acquiring the right preferences has recently become a bottleneck, and extremely important:

- Recommender systems, such as the YouTube sidebar, targeted advertising, internet radio, and even search algorithms, have the job of learning users' preferences (both individual and collective) about content. This makes good models of preference dynamics useful directly for predictive power in humans.
- These learning systems and others have objective functions which they optimize—preferences of their own—but getting these objective functions "wrong" can result in systems that are totally useless, such as game playing agents that incur no loss by pausing the game, ad delivery systems that plaster the page in ads to maximize ad clicks. Worse, they can appear right at first, but have nasty side effects: unfairly reinforcing inequality, serving news that reinforces confirmation biases, and so forth.
- Both markets and democracies can also be thought of as optimization procedures for specific objectives (defined by the currency and voting systems, respectively). Both have well-documented failure cases, and in response, many have come up with alternative ways of regulating trade and counting votes. But how do we determine if they're any good? We would need a model of preference updates.
- Many now worry that an artificial general intelligence (AGI), smart enough to out-maneuver humanity, but with undesirable preferences ("unaligned") is becoming increasingly likely. Many serious AI practitioners take this concern seriously: DeepMind, OpenAI, MIRI, and FHI all have AI safety divisions for this reason. OpenAI and Google Brain's seminal paper [Amo+16] details many of the particular concerns and the current ideas we have for solving them.

I think it is important to mention that, while these are all issues that have only recently become problematic, the general problem of having to figure out what to care about has been a (and arguably the only) fundamental problem in philosophy for a long time: ethics is a field dedicated to deciding what we should care about;

meta-ethics a field about how to make those decisions. Sociological accounts of cultural revolution and art history both are also both theories of preference change in specific domains.

Even the specific worries about mis-specifying objectives to powerful optimizing procedures have ancient cultural roots: there are endless stories and jokes of leprechauns, genies, and monkey paws granting wishes that turned out to have unintended side effects.

This is all to say: the problem at hand is enormously important, and people have already put an enormous of work into it. And yet, none of few formal theories of preference acquisition and updating has received much attention, and none of them are in dialog with ML theory. It is crazy to expect that the single contribution I make here could single-handedly render all of this other work obsolete, but it is rather intended that this will provide a unified language and formalism that others can use to represent their own models and solutions, some general results which could be used as support for positions in any of these fields, and also some that are more specific to the setting of machine preference updates.

Responses to some preliminary concerns

Nobody doubts that getting the "right" preferences are important, but there are a number of reasons that people might see this as a garbage research agenda; before continuing, I'd like to quickly address a few of these:

There's no such thing as "right". The first issue has to do with the quotation marks that I've put around the morally loaded words I've used up until now. Moral relativism, perhaps in combination with an intuition of preferences that comes from classical economic theory, would suggest that every preference is equally valid. This viewpoint also meshes well with the distinction emotion / reason dichotomy people like to draw, and the orthogonality thesis [Bos12]. In this view, the merit of something can only be evaluated with respect to your own preferences, and [**todo**: *preference updating merits*]

This all just boils down to one more objective you've put on top. [**todo**: *implicit objective may be better, if it has desirable properties*]

Relevant Work in Representing, Acquiring, and Updating Preferences

Bayesian Networks are a particularly relevant technique for representing (2), and represent a factorization of a joint probability distribution across a number of variables N as the product of conditional probability distributions associated with the edges of a graph whose nodes are N.

CP Nets [Bou+04] are one attempt to factor preferences analogously to Bayesian Networks, but they suffer from a number of issues, which we will explore in section **??**.

Drawing the analogy further, observations

[**todo**: *finish background: Inverse Reinforcement Learning, CP Nets, Girard, Fenrong, Skill Babling, Policy Space Topology, Taxonomy from 2009 Pref Change book*]

2 What I want to do

I'm trying to get a theory of preference dynamics.

Currently preferences are thought of as static objects, fixed as part of the structure and identity of an agent, independent of beliefs, complete, and over some fixed domain. This is clearly not at all how human preferences work, and I posit that it's not the right way to think of preference for synthetic agents either.

Perhaps among other things, desiderata for a model of preference dynamics should:

1. Provide an answer to the 'value loading' problem: show how you can learn "reasonable" preferences by interacting with the world
2. Reduce to static models for some parameter settings
3. Behave reasonably when combined with changes of perspective
4. Be resistant to standard challenges to irrationality, such as dutch booking
5. 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 maths), 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 exam scores becomes valuable, independent of whether you learn math[1]). Related to Goodhart and Cambell's laws.
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.
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. Technically, we want to capture the fact that additional experiences increase preference complexity.
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.
5.
6. **Novelty Effects (anti-adaptive preferences).**

3 Informal Examples of Preference Dynamics

3.1 Coffee vs Beer

Example 3.1. 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 3.2. 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. △

[1] 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.

Example 3.3. 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 3.4. 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 3.5. 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 3.6. 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 3.7. 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 3.8. 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 3.9. 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. △

3.2 Other Examples

Example 3.10. 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 3.11. 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:

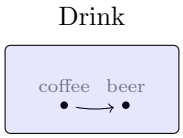
1. You have no reason (or have since forgotten) why you hate flying to HK, and discover no new reasons to hate the journey: you get a reasonable amount of work done and the entertainment is just as good as you would have done for yourself. As your preference for seeing your family is in conflict with 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.

△

4 Failure of Classical Approaches to Capture These Examples

4.1 Coffee and Beer

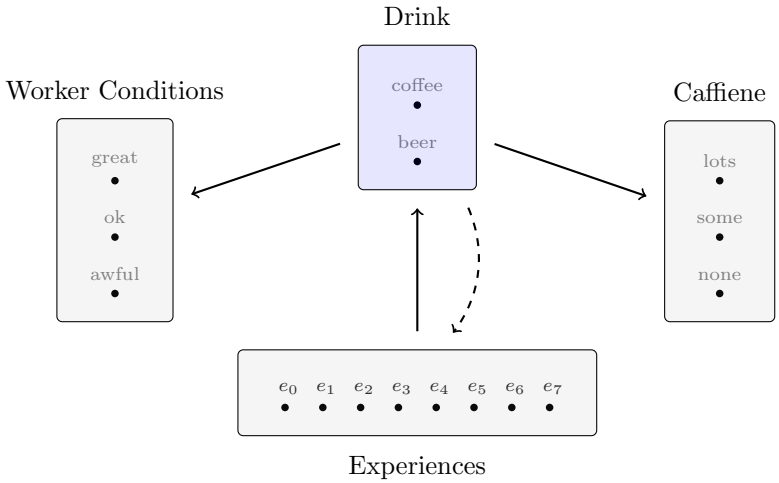
Suppose you are trying to chose between coffee and beer. Had you been forced to make the decision instantly, you may have already had an answer that you could spit out: maybe you have a cached preference for beer from the last time you made this decision, or perhaps you already think of yourself as the kind of person who prefers beer over coffee. At this point, we could stop modeling, and merely state that this is one of those axiomatic things we require to make decisions. This corresponds to the following diagram, which requires no computation:



On the other hand, given enough time to think, you may remember you had reasons for choosing beer the last time. Some considerations might include:

1. The coffee is not fair trade, and contributes to poor working conditions in Columbia
2. Your previous experiences with coffee have been better than your previous experiences with beer
3. Coffee has more caffeine than beer

Note that each of these considerations consists of two parts: another thing you assign value to (the pleasure of your previous experiences, exploited workers, caffeine), plus some belief about how a decision in one domain will impact the others (what impact does coffee have on the world, what do I remember about my experiences, what do I know about the caffeine content). We can picture this with the following diagram:



According to the diagram, we endorse the following decision making process: start with your ambient (prior, in analogy to Bayesian updating) preference for the decision at hand: which drink to order. Then, consider other domains you have ambient preferences attached to, the way in which this decision impacts them, and nudge your current preference in the appropriate way. While the two things on the edges look like they have some causal flavor, the one on the bottom is different: there's no way that having a drink can cause you to have one of your previous experiences. Instead, your experiences have the property of **[todo:]**

Now that we have added more information to the model, let's consider what the standard modeling approach would be in this situation: we would need (1) utilities $U : \mathcal{W} \to \mathbb{R}$ over worlds, which in this case would be modeled as an assignment $\mathcal{W} = (D \times W \times C \times E)$ to each of the four variables (or selection of a point from each of the domains in the picture above), and also (2) a conditional probability distribution over worlds $\Pr(\mathcal{W}|D)$, for each choice of drink. At this point we compute expected utilities,

$$\mathbb{E}\left(U \;\middle|\; \text{coffee}\right) = \sum_{w \in \mathcal{W}} U(w) \Pr(w|\text{coffee}) \qquad \mathbb{E}\left(U \;\middle|\; \text{beer}\right) = \sum_{w \in \mathcal{W}} U(w) \Pr(w|\text{tea})$$

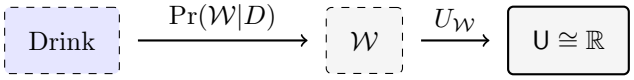
and compare the results. Doing this computation directly has a number of draw-backs:

- There's not a clear way of making use of partial computations: if you're halfway through your expected utility computations, you still haven't made a decision.
- It requires a joint distribution over all variables you care about: in our example, we would have needed to know how our experiences related to worker conditions in Columbia, how both of these relate to your caffeination levels.

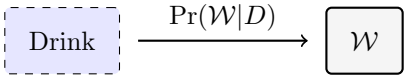
- From the modeler's perspective, this whole setup changes drastically if it turns out we at the last minute remember that social signaling is relevant to the decision. Such a change causes even the collection of worlds to be different, and it's not immediately obvious if we can recycle our old utility function and probability distributions.

It may seem at this point like we have merely introspected a hacky algorithm which does not suffer from these issues by design, ignored the rationality guarantees provided by expected utility, and hidden the fact that the expected utility computations can often be approximated quickly with some straightforward tricks. From the classical point of view, all we needed to do was assume that U is additively separable, that the independence assumptions that accord with the interpretation of the above diagram as a Bayesian Net hold, and just compute the correct answer directly. From this perspective, it looks as though our procedure is a special case of expected utility but implemented in a riskier, incremental way.

On the other hand, expected utility can be seen as a special, particularly brittle and centralized, case of our updating method, in which the only thing that matters is a special domain $\mathbb{R}$, there is always exactly one path that can be used to compute it, which involves integrating over all possible worlds, and we don't save any computations. In particular, by using a dashed boundary to denote a domain which does **not** accrue value, this corresponds to running our updating on the following diagram, and just computing until it converges:

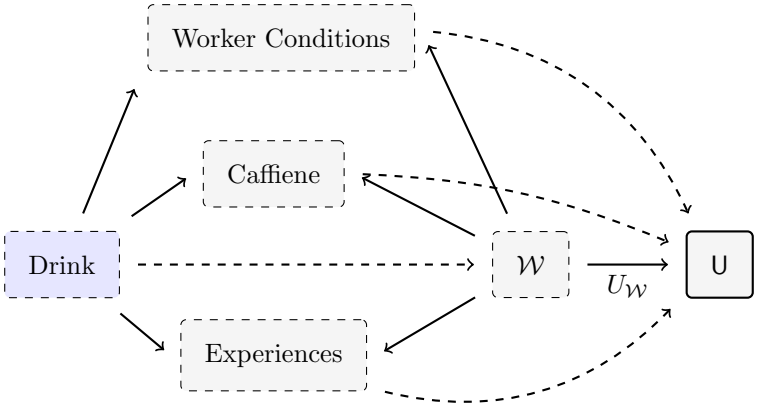


Note also that if the preferences on $\mathcal{W}$ did not come from a utility function, but rather you simply had preferences over possible worlds, we have an even simpler picture:



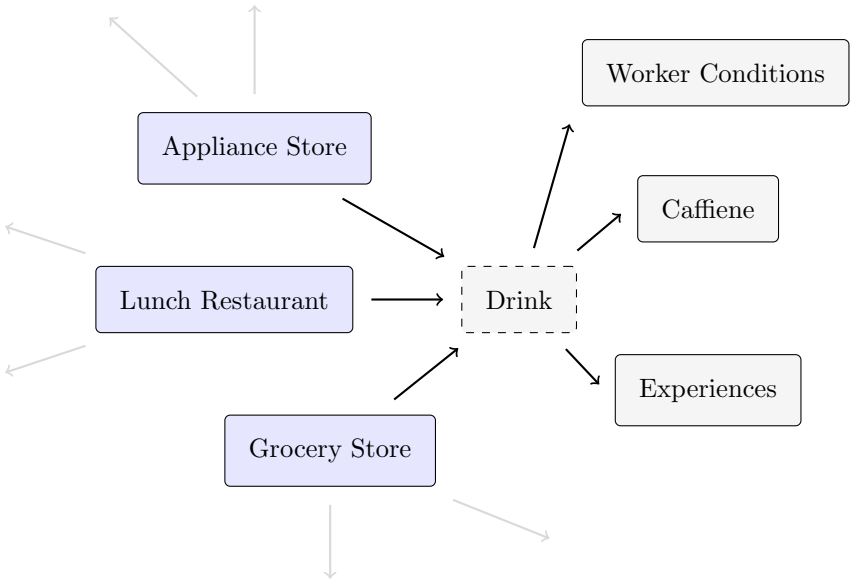
This is a quick graphical illustration of why preferences representable by utility functions are a strict sub-class of preferences over worlds. Also, it suggests that a utility function is mathematically a "belief-like" object, tying two domains together, which is why it makes sense to define products between it and probability distributions.

Moreover, if the only thing we cared about was the special utility box anyway, both the additive separability and conditional independence assumptions have a graphical interpretation if we represent them in this manner. Consider the diagram below:



The line from Drink to $\mathcal{W}$ has to be inferred somehow, if the only way we can compute value is by going to U via $U_{\mathcal{W}}$. Doing this by assuming that all of the variables are independent is the only option we really have available, which fully utilizes all of the information we have in the diagram.

The decision of what drink to have does not exist in a vacuum. Suppose you also have to decide what restaurant to go to get lunch, and also where to buy appliances, and groceries. Each of these depend on what drinks you like, among other things. This might look something like the diagram below.



Having to re-compute your preference for drinks every time you have to make a decision is expensive, and while it is always possible that your drink preference would have changed in the mean time if you had re-computed it, this is probably not a very effective use of your computation, which could be spent sorting through the many other relevant features of grocery stores and restaurants.

Once again, this can be thought of in two ways: the classical one is to state that

5 Some Intuition about the Framework

We think of preferences over most things as actually being cached computations of how to optimize other, previously held preferences, filtered through some beliefs about the world. For instance,

- A taste for ice cream can be thought of as a cached version of the evolutionary preference for staying alive, seen through the lens of food choice: perhaps a cached preference computation from times when calories were scarce.
- A preference for green apples over red ones might be the cached computation from the last time you were faced with a similar choice, incorporating factors such as usefulness in the kitchen, sourness, color preference, and so forth. This preference is not an intrinsic feature of your agency, but rather a cached view of other preferences through the apple lens.
- An affinity towards the political ideal freedom can be thought of as a cached view of your preferences over governments you've read about or experienced, seen through the lens of one particular feature: how much freedom the government offers.
- Vegetarianism can be thought of as a cached preference for environmentalism and reduction of animal suffering, filtered through the lens of food choice.
- Habits can be thought of as cached versions of your preferences (nutritional, entertainment, intellectual, structural, etc.), filtered through the lens of stimulus-response functions

With this in mind, our picture of preferences encodes a bunch of redundant information, in many different settings, and the picture is informed by the ways that the world can cause one setting to have a bearing on other related ones.

Note that it is possible to de-couple a preference from the original reasons for forming it: the first and last examples above are good illustrations. A sweet-tooth is no longer the best way to ensure evolutionary success now due to environmental distributional shift, but we retain this preference anyway. In the last example, people who became vegetarian purely for environmental and animal rights reasons, will do better on both fronts eating meat for a day in exchange for a friend abstaining for two — and yet people are often hesitant to make this trade, because the preference acquires.

If we were both cognitively unbounded and certain about what we cared about, we would not need to keep around preference domains for anything else for very long. We could just re-compute everything we needed from scratch from only our single true preference domain, which was precise enough to exactly capture the one thing we care about — this corresponds to the classical view of static preferences.

6 Preliminaries and Foundational Theory

6.1 Preferences as Binary Relations

In its most basic form, a preference over a set A of alternatives, i.e., the set of things we can chose between, is a binary relation $\preccurlyeq$ on A that is transitive and reflexive, where $a \preccurlyeq b$ is interpreted as "Given a choice between a and b, I would choose b". To make them self-consistent, people additionally impose axioms:

$$\forall x \in A.\ x \preccurlyeq x \qquad \text{(Reflexivity)}$$
$$\forall x, y, z \in A.\ (x \preccurlyeq y)\ \wedge\ (y \preccurlyeq z) \Rightarrow\ (x \preccurlyeq z) \qquad \text{(Transitivity)}$$
$$\forall x, y \in A.\ (x \preccurlyeq y)\ \vee\ (y \preccurlyeq x) \qquad \text{(Completeness)}$$

It is also common to speak of indifference between two alternatives ($x \sim y$), and strict preference of y to x ($x \prec y$), which we define as:

$$\text{Indiference:} \qquad x \sim y\ := (x \preccurlyeq y) \wedge (y \preccurlyeq x)$$
$$\text{Strict Preference:} \qquad x \prec y\ := (x \preccurlyeq y) \wedge \neg(y \preccurlyeq x)$$

6.2 Preference Matrices

Because a binary relation $\preccurlyeq$ on A is a function $\mathbf{A}_{\preccurlyeq} : A \times A \to \mathbb{B}$, taking a pair of alternatives, and returning a boolean, we can think of it as a boolean matrix indexed by values of A, and re-interpret the axioms above as matrix properties. This will allow us to get fuzzy preferences for continuous updates, and allow us to express things as "the desire to make preferences more transitive" without requiring that they always are in this state.

Example 6.1. If $A = \{a_0, a_1, a_2\}$, and $A \prec B \prec C$, then using the basis $[a_0, a_1, a_2]$, we have:

$$\mathbf{A}_{\preccurlyeq} = \begin{array}{c} \\ a_0 \\ a_1 \\ a_2 \end{array} \begin{array}{c} \begin{array}{ccc} a_0 & a_1 & a_2 \end{array} \\ \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \end{array}$$

△

Example 6.2. On the other hand, with again $A = \{a_0, a_1, a_2\}$, if we were indifferent between all of the alternatives, we would have

$$\mathbf{A}_{\preccurlyeq} = \begin{array}{c} \\ a_0 \\ a_1 \\ a_2 \end{array} \begin{array}{c} \begin{array}{ccc} a_0 & a_1 & a_2 \end{array} \\ \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \end{array}$$

△

This representation, suggests some additional algebraic structure. The Boolean algebra $\mathbb{B}$ forms a semi-ring where addition is $\vee$ and multiplication is $\wedge$, and as a consequence, the $\mathbb{B}$-matrices over a finite set A, denoted

$M_A(\mathbb{B})$, also form a semi-ring, where for $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{B})$ we define:

$$(\mathbf{A} + \mathbf{B})_{i,j} := \mathbf{A}_{i,j} + \mathbf{B}_{i,j}$$
$$(\mathbf{AB})_{i,j} := \sum_{k \in A} \mathbf{A}_{i,k}\mathbf{B}_{k,j}$$

We can also think of the boolean algebra order-theoretically: we can define an order from addition, so that $a \leq b := (a + b = b)$, for $a, b \in \mathbb{B}$. Analogously, we define the matrix order on $M_A(\mathbb{B})$ point-wise:

$$(\mathbf{A} \leq \mathbf{B}) := \forall i, j \in A.\ \mathbf{A}_{i,j} \leq \mathbf{B}_{i,j}$$

Fact 6.1. *If $\preccurlyeq$ is a binary relation on A, then $\preccurlyeq$ is complete if and only if $\mathbf{A}^T + \mathbf{A} = \mathbf{1}$, where $\mathbf{1}_{i,j} = 1$*

Fact 6.2. *If $\preccurlyeq$ is a binary relation on A, then $\preccurlyeq$ is reflexive if and only if $\mathbf{I} \leq \mathbf{A}$, where*

$$\mathbf{I}_{i,j} := \begin{cases} 1 & i = j \\ 0 & \textit{otherwise} \end{cases}$$

Proposition 6.3. *If $\preccurlyeq$ is a binary relation on A, then $\preccurlyeq$ is transitive if and only if $(\mathbf{A}_{\preccurlyeq})^2 \leq (\mathbf{A}_{\preccurlyeq})$*

Proof. To make room for index subscripts, we will write $\mathbf{A}^{\preccurlyeq}$ instead of $\mathbf{A}_{\preccurlyeq}$. Unwinding the definitions:

$$\begin{aligned}
& (\mathbf{A}_{\preccurlyeq})^2 \leq (\mathbf{A}_{\preccurlyeq}) \\
\iff \quad & \forall i, j \in A. \quad \left(\sum_{k \in A} \mathbf{A}^{\preccurlyeq}_{i,k}\mathbf{A}^{\preccurlyeq}_{k,j}\right) \leq \mathbf{A}^{\preccurlyeq}_{i,j} \\
\iff \quad & \forall i, j \in A. \quad \left(\sum_{k \in A} \mathbf{A}^{\preccurlyeq}_{i,k}\mathbf{A}^{\preccurlyeq}_{k,j}\right) + \mathbf{A}^{\preccurlyeq}_{i,j} = \mathbf{A}^{\preccurlyeq}_{i,j} \\
\iff \quad & \forall i, j \in A. \quad \left((i \preccurlyeq j) \vee \bigvee_{k \in A} (i \preccurlyeq j) \wedge (j \preccurlyeq k)\right) \Leftrightarrow (i \preccurlyeq j)
\end{aligned}$$

For any $\varphi \vee \psi \Leftrightarrow \varphi$ is equivalent to $\varphi \vee \psi \Rightarrow \varphi$, since the reverse direction always holds, which is in turn equivalent to $\psi \Rightarrow \varphi$. As a result, our expression becomes

$$\begin{aligned}
\iff \quad & \forall i, j \in A. \quad \left(\bigvee_{k \in A} (i \preccurlyeq j) \wedge (j \preccurlyeq k)\right) \Rightarrow (i \preccurlyeq j) \\
\iff \quad & \forall i, j \in A. \quad \neg\left(\bigvee_{k \in A} (i \preccurlyeq j) \wedge (j \preccurlyeq k)\right) \vee (i \preccurlyeq j) \\
\iff \quad & \forall i, j \in A. \quad \left(\bigwedge_{k \in A} \neg\Big[(i \preccurlyeq j) \wedge (j \preccurlyeq k)\Big]\right) \vee (i \preccurlyeq j) \\
\iff \quad & \forall i, j \in A. \quad \bigwedge_{k \in A} \Big(\neg\Big[(i \preccurlyeq j) \wedge (j \preccurlyeq k)\Big] \vee (i \preccurlyeq j)\Big) \\
\iff \quad & \forall i, j, k \in A. \quad \Big[(i \preccurlyeq j) \wedge (j \preccurlyeq k)\Big] \Rightarrow (i \preccurlyeq j) \\
\iff \quad & \preccurlyeq \text{ is transitive}
\end{aligned}$$

□

We now have a test for transitivity, that suggests that iterated products are the way to compute transitive closure. In fact, the matrix star operator from iterated application is indeed what we're looking for, but to prove this first we need a lemma:

Lemma 6.4. *If $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are square matrices over a (pre[2])-semiring $(S, +, \times)$, where an ordering $\leq$ is defined from $+$ as above, and $\mathbf{A} \leq \mathbf{B}$, then $\mathbf{AC} \leq \mathbf{BC}$, and $\mathbf{CA} \leq \mathbf{CB}$.*

Proof. As $\mathbf{A} \leq \mathbf{B}$, which is shorthand for $\mathbf{A} + \mathbf{B} = \mathbf{B}$, we have:

$$\mathbf{AC} + \mathbf{BC} = \sum_{s \in S} \mathbf{A}_{i,s}\mathbf{C}_{s,j} + \sum_{s \in S} \mathbf{B}_{i,s}\mathbf{C}_{s,j} = \sum_{s \in S} (\mathbf{A}_{i,s} + \mathbf{B}_{i,s})\mathbf{C}_{s,j} = \sum_{s \in S} \mathbf{B}_{i,s}\mathbf{C}_{s,j} = \mathbf{BC}$$

And therefore $\mathbf{AC} \leq \mathbf{BC}$. A similar proof works on the left with left distributivity. Note that we don't even need to expand the indices, as the matrices themselves form a semi-ring, so this is just weakly disguised distributivity. □

Now, the promised result:

Proposition 6.5. *If $\preccurlyeq$ is a relation on A, whose matrix is $\mathbf{A}$, then $\mathbf{A}^* := \mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \mathbf{A}^3 + \ldots$, if it exists, is the reflexive transitive closure of $\preccurlyeq$.*

Proof. To prove this, we must show that $\mathbf{A}^*$ is transitive, includes $\mathbf{A}$, and is included in any other reflexive transitive relation which has both of these properties. As $+$ is element-wise $\vee$ in this case, if $\mathbf{A}_{i,j} = 1$, then $\mathbf{A}^*_{i,j} = \mathbf{I}_{i,j} \vee \mathbf{A}_{i,j} \vee \cdots = 1$, so $\mathbf{A} \subseteq \mathbf{A}^*$. The transitivity argument follows from distributivity and idempotence of $+$:

$$\begin{aligned}
\left(\sum_{n=0}^{\infty} \mathbf{A}^n\right)^2 &= \left(\sum_{n=0}^{\infty} \mathbf{A}^n\right) + \mathbf{A}\left(\sum_{n=0}^{\infty} \mathbf{A}^n\right) + \mathbf{A}^2\left(\sum_{n=0}^{\infty} \mathbf{A}^n\right) + \cdots \\
&= \mathbf{I} + \mathbf{A} + (\mathbf{A}^2 + \mathbf{A}^2) + (\mathbf{A}^3 + \mathbf{A}^3 + \mathbf{A}^3) + \cdots \\
&= \mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \mathbf{A}^3 + \cdots \\
&= \mathbf{A}^*
\end{aligned}$$

This clearly works in the finite case, and can be inductively extended to countably infinite sets for boolean algebra, as $\wedge$ distributes even over infinite disjunctions. Finally, if $\mathbf{A} \leq \mathbf{B}$, and $\mathbf{B}$ is transitive, then we will prove by induction on k that $\sum_{k=0}^{n} \mathbf{A}^k \leq \mathbf{B}$. First note that since $\mathbf{B}$ is transitive, we have $\mathbf{B}^2 \leq \mathbf{B}$, and so it follows inductively that $\mathbf{B}^k \leq \mathbf{B}$ for all $k \in \mathbb{N}$.

Now base case, for $n = 0$ (i.e., $\mathbf{I} \leq \mathbf{B}$) is immediate because B is assumed to be reflexive. Now suppose this holds for a particular n. Then,

$$\sum_{k=0}^{n+1} \mathbf{A}^k = \mathbf{A}^{n+1} + \sum_{k=0}^{n} \mathbf{A}^k \leq \mathbf{B} + \mathbf{A}^{n+1} \leq \mathbf{B} + \mathbf{B}^{n+1} = \mathbf{B}(\mathbf{I} + \mathbf{B}^n) \leq \mathbf{B}(\mathbf{IB}) = \mathbf{B}^2 \leq \mathbf{B}$$

Passing to the limit, we therefore have $\mathbf{A}^* \leq B$, as desired. □

Remark. Because 1 is an anihilator for $\vee$, closures always exist (if A is finite) for $M_A(\mathbb{B})$ [Gri18], making it a Kleene Algebra.

Remark. We can also think of a preference domain $(A,\ \mathbf{A}\colon A \times A \to S)$ over a semiring S as an S-enriched category. In particular, orders are thin categories. As we will see, the links can be thought of as profunctors on these categories, and many categorical constructions have useful analogs.

[2] a semiring that might not have identities

6.2.1 Interpretations

Relaxations of this theory to fuzzy semirings, combined with information from other domains, will be the formal basis of our preference updating theory. As sanity checks, we can now go through some other common interpretations of semiring matrices, particularly over boolean algebras, and verify that the preference interpretation makes sense:

1. If we think of a set of propositions S as a preference domain, where we prefer "truer" statements, i.e., $s \preccurlyeq t$ if $s \Rightarrow t \in S$, then $(\mathbf{S}^*_{\preccurlyeq})$ is the inferred preference on

6.3 Non-Boolean Preference Matrices

One reason to react against purely order-based preference models is that they throw away a lot of important information. For instance, perhaps prefer vanilla ice cream to chocolate, and chocolate ice cream to 47 years in prison. One of these choices is a lot easier than the other. Moreover, this is not purely an internal feeling: putting on our frequentist hats for a moment, repeated exposure to this choice, would likely result in some probability mass on selections of both vanilla and chocolate ice cream, but no takers for the prison.

This is one of many benefits utility has over preference orderings. Recall that a utility function on a set of alternatives A is a function $u : A \to \mathbb{R}$, the value of which one tries to maximize with choices. However, assigning utilities to things is generally considered less general (despite the above concern) than the order-theoretic approach, because not all valid preferences can be described using utility functions: for this to be the case, preferences must be both transitive and complete. In the presence of uncertainty (i.e., choices take the form of acts: maps from states to outcomes), Savage [Sav72] shows that a number of additional restrictions are necessary to represent a preference order with a probability distribution over states and a utility function on outcomes.

With real-valued preference matrices, we can get the representational flexibility of both approaches: any preference relation $\preccurlyeq$ on a set A can be represented just as we did in examples 6.1 and 6.2, by setting

$$\mathbf{A}^{\preccurlyeq}_{i,j} := \begin{cases} 1 & \text{if } i \preccurlyeq j \\ 0 & \text{otherwise} \end{cases}$$

On the other hand, we can also represent a utility function $u : A \to \mathbb{R}$, as a matrix, in several useful ways:

$$\begin{aligned} \left[\mathbf{A}^{+u}\right]_{i,j} &:= u(j) - u(i) \\ \left[\mathbf{A}^{\times u}\right]_{i,j} &:= \left[\exp \mathbf{A}^{+u}\right]_{i,j} = \exp(u(j) - u(i)) \\ \left[\mathbf{A}^{\sim u}\right]_{i,j} &:= \left[\frac{1}{\mathbf{1} + \exp(-\mathbf{A}^{+u})}\right]_{i,j} = \frac{1}{1 + \exp(u(i) - u(j))} \end{aligned}$$

The different representations will allow us to deal with utilities as matrix factorizations, by representing combinations multiplicatively rather than aditively, as in proposition 6.8, and allow us to make use of existing theory in [EV04]. But first, some more basic, sanity results. We would like the different embeddings to generate the same orders:

Proposition 6.6. *If X is a set, and $\preccurlyeq$ is an order, corresponding to according to a utility function $u : X \to \mathbb{R}$, i.e., $x \preccurlyeq y \Leftrightarrow u(x) \leq u(y)$, then*

$$\mathbf{X}^{\preccurlyeq} = H(\mathbf{X}^{+u}) = H(\mathbf{X}^{\times u} - \mathbf{1}) = H(\mathbf{X}^{\sim u} - \tfrac{1}{2}\mathbf{1})$$

where H is the right-biased pointwise heaviside function,

$$H(\mathbf{T})_{i,j} := \mathbb{1}_{\left[\mathbf{T}_{i,j} \geq 0\right]} = \begin{cases} 0 & \text{if } \mathbf{T}_{i,j} < 0 \\ 1 & \text{if } \mathbf{T}_{i,j} \geq 0 \end{cases}$$

Proof. Each can easily be reduced to $\mathbf{X}^{\preccurlyeq}$, as each of $\mathbf{X}^{+u}$, $\mathbf{X}^{\times u}$, and $\mathbf{X}^{\sim u}$ are all strictly monotonic functions of the difference in utility at each coordinate:

$$H(\mathbf{X}^{+u})_{i,j} = \mathbb{1}_{\left[\mathbf{X}^{+u}_{i,j} \geq 0\right]} = \mathbb{1}_{\left[u(j)-u(i)\geq 0\right]} = \mathbb{1}_{\left[u(j)\geq u(i)\right]} = \mathbb{1}_{\left[i \preccurlyeq j\right]} = \mathbf{X}^{\preccurlyeq}$$

$$H(\mathbf{X}^{\times u} - \mathbf{1})_{i,j} = \mathbb{1}_{\left[\mathbf{X}^{\times u}_{i,j} \geq 1\right]} = \mathbb{1}_{\left[\exp(u(j)-u(i))\geq 1\right]} = \mathbb{1}_{\left[u(j)\geq u(i)\right]} = \mathbb{1}_{\left[i \preccurlyeq j\right]} = \mathbf{X}^{\preccurlyeq}$$

$$H(\mathbf{X}^{\sim u} - \tfrac{1}{2}\mathbf{1})_{i,j} = \mathbb{1}_{\left[\mathbf{X}^{\sim u}_{i,j} \geq \frac{1}{2}\right]} = \mathbb{1}_{\left[\frac{1}{1+e^{u_i-u_j}} \geq \frac{1}{2}\right]} = \mathbb{1}_{\left[1+e^{u_i-u_j}\leq 2\right]} = \mathbb{1}_{\left[e^{u_i-u_j}\leq 1\right]} = \mathbb{1}_{\left[u_i \leq u_j\right]} = \mathbb{1}_{\left[i \preccurlyeq j\right]} = \mathbf{X}^{\preccurlyeq}$$

□

In the case of $\mathbf{X}^{\sim u}$, we have a much stronger limit coherence result, which does not require rounding or shifting:

Proposition 6.7. *For any set X with utility $u : X \to \mathbb{R}$, whose induced ordering on X is $\preccurlyeq$, we have*

$$\lim_{t\to\infty} \mathbf{X}^{\sim tu} = \mathbf{X}^{\preccurlyeq},$$

where $tu := x \mapsto t \cdot u(x)$ is a scaled version of u, and therefore induces the same ordering.

Proposition 6.8. *A non-negative preference matrix $\mathbf{A} : X \times X \to \mathbb{R}_{\geq 0}$ is a rank-1 matrix with $\mathbf{A} \odot \mathbf{A}^T = \mathbf{1}$ (where $\odot$ is the element-wise, or Hammard product), if and only if $\mathbf{A}$ is multiplicatively representable by a utility function, i.e., there exits $u : X \to \bar{\mathbb{R}}$ such that $\mathbf{A} = \mathbf{X}^{\times u}$*

Proof. We'll start with the easier, backwards direction. If $\mathbf{A} = \mathbf{X}^{\times u}$, then

$$\mathbf{A}_{i,j} = \left[\mathbf{X}^{\times u}\right]_{i,j} = \exp(u(j) - u(i)) = \left(e^{u(j)}\right)\left(e^{-u(i)}\right)$$

Therefore we can write $\mathbf{A} = x^T y$, where x and y are $\mathbb{R}$-vectors over X, defined as $x_i = e^{-u(i)}$ and $y_j = e^{u(j)}$, so $\mathbf{A}$ is rank-1. Also, we have

$$\left[\mathbf{A} \odot \mathbf{A}^T\right]_{i,j} = \mathbf{A}_{i,j}\mathbf{A}_{j,i} = \frac{e^{u_j}}{e^{u_i}}\frac{e^{u_i}}{e^{u_j}} = 1$$

On the other hand, now suppose that $\mathbf{A}$ is rank-1, with $\mathbf{A} \odot \mathbf{A}^T = \mathbf{1}$. Because it is rank-1, there exist real valued vectors x and y over X such that $\mathbf{A} = x^T y$, and so the second condition becomes: $x^T y \odot y^T x = \mathbf{1}$, or equivalently,

$$\forall i, j \in X. \quad x_i y_j y_i x_j = 1 \qquad \Longrightarrow \qquad x_j y_i = \frac{1}{y_j x_i}$$

We now define $u(k) := \frac{1}{2}(\log y_k - \log x_k) = \log\sqrt{y_k/x_k}$, so that

$$\left[\mathbf{X}^{\times u}\right]_{i,j} = \exp\left(\log\left(\sqrt{y_j/x_j}\right) - \log\left(\sqrt{y_i/x_i}\right)\right) = \exp\log\sqrt{\frac{y_j x_i}{x_j y_i}} = \sqrt{(y_j x_i)(y_j x_i)} = x_i y_j = \mathbf{A}_{i,j}$$

□

Corollary 6.8.1. *A preference matrix $\mathbf{A} : X \times X \to \mathbb{R}$ is additively representable by a utility function (i.e., $\exists u : X \to \mathbb{R}$ such that $\mathbf{A} = \mathbf{X}^{+u}$) if and only if $\mathbf{A}$ is skew symmetric and $\exp(\mathbf{A})$ is rank-1.*

6.3.1 Non-standard Semi-rings and Low-Rank Decompositions

Peron-Frobenius Vector.

Subtropical-Eigenvector. [EV04][ED10]

6.3.2 Parameterized Weights

6.4 Domain Links

Preference domains are only half of the picture: we also want to link them together, and describe how preferences in one domain can impact ones in other domains. To do this, we will need some mathematical formalism to describe them. For now[3], we will formalize these as conditional probability distributions. A link L from one preference domain to another is just a function

$$L : A \to \Delta B$$

which yields a probability distribution ΔB on B for every element of A. This is thought of as the impact of the choice of alternatives in domain A on the domain B. If both A and B are discrete, these conditional distributions can be represented as stochastic matrices[4] (matrices $\mathbf{L} : A \times B \to [0,1]$ such that $\forall a \in A. \sum_{b \in B} \mathbf{L}_{a,b} = 1$); for continuous domains, we can represent them as Markov kernels, which have analogous properties for the more complicated setting of measurable spaces.

6.5 Dynamics

In addition to keeping track of preferences, the whole point is to provide a prescription for how they should change over time. There are several driving forces we could consider: we have already vaguely suggested iterative matrix powers and low-rank approximation as methods of compressing data and pushing towards transitivity (it would take only a single matrix factorization to find a loss-minimizing solution), and it's straightforward to use these results to create differential pressure for complete, reflexive, and anti-symmetric preferences.

All of these are relatively well-known results, and apply only to a single domain; for us, the driving force will come from the interaction between them. We conjecture this updating method is strictly general:

Conjecture 6.9 (distributed transitivity)**.** *If* D *is a preference domain, then the fixed point of the model*

$$\mathcal{D} = \left\{ \{d, d', \bullet\}, \mathbf{D}\big|_{d,d'} \right\}_{d \neq d' \in D}, \qquad \mathcal{B}\Big((i,j),(i',j')\Big) = \begin{array}{c} \begin{matrix} i' & j' & \bullet \end{matrix} \\ \begin{bmatrix} \delta_{i,i'} & \delta_{i,j'} & 1-\delta_{i,i'}-\delta_{i,j'} \\ \delta_{j,i'} & \delta_{j,j'} & 1-\delta_{i,i'}-\delta_{i,j'} \\ 1-\delta_{i,i'}-\delta_{j,i'} & 1-\delta_{i,j'}-\delta_{j,j'} & \end{bmatrix} \end{array} \begin{matrix} i \\ j \\ \bullet \end{matrix}$$

corresponds to the transitive closure of D*.*

Proof. [todo:] □

As a first pass, note that if we had a degenerate distribution (a function) $f : X \to Y$, and we had preferences on Y but not X, we could define $x_1 \preccurlyeq_X x_2 := f(x_1) \preccurlyeq_Y f(x_2)$. It is an important feature that this process of building up preferences backwards, in the reverse direction of the links, is the natural way to form them. In fact, to go the other direction, you need to apply Bayes' rule or compute a pseudo-inverse to get something that can be used the link going the other way.

Example 6.3. If the preferences on Y are represented by utilities $u : Y \to \mathbb{R}$, then the preferences on X induced by a functionX f and Y are given by $u \circ f$.

[3]For example, I get the sense we could also have used joint probabilities and undirected models to get a Markov Network, or even a causal model, where the structural equations are edges, and the nodes are variables. However, this would require different notions than the ones described here.

[4]If the edges were undirected and we had started with a Markov random field, these would be matrices of the same shape, representing joint probability distributions

If we only had a function $Y \to X$, and not one from $X \to Y$, then there would be no natural composition of the parts at our disposal to form a utility function on X. △

Proposition 6.10. *If* A, B *are preference domains,* $\mathbf{B} = \mathbf{B}^{+u}$ *is additively represented by* u*, and* $\mathbf{P} : A \to \Delta B$ *is a stochastic matrix, then the preferences* $\mathbf{A} := \mathbf{P}\mathbf{B}\mathbf{P}^T$ *suggested by* $\mathbf{B}$ *filtered through* $\mathbf{P}$ *is additively represented by expected utility* $\mathbb{E}u$*.*

Proof.

$$\begin{aligned}
\mathbf{A}_{il} &= \sum_{j,k \in B} \mathbf{P}_{i,j}\mathbf{B}_{j,k}\mathbf{P}_{l,k} \\
&= \sum_{j,k \in B} \mathbf{P}_{i,j}\mathbf{P}_{l,k}(u_k - u_j) \\
&= \sum_{j \in B} \mathbf{P}_{i,j} \left[\sum_{k \in B} \mathbf{P}_{l,k}(u_k - u_j) \right] \\
&= \sum_{j \in B} \mathbf{P}_{i,j} \left[\sum_{k \in B} \mathbf{P}_{l,k}u_k - \sum_{k \in B} \mathbf{P}_{l,k}u_j \right] \\
&= \sum_{j \in B} \mathbf{P}_{i,j} \underbrace{\left[\sum_{k \in B} \mathbf{P}_{l,k}u_k \right]}_{\text{indep. of } j} - \sum_{j \in B} \mathbf{P}_{i,j}u_j \quad \underbrace{\sum_{k \in B} \mathbf{P}_{l,k}}_{\text{column sums to 1}} \\
&= \left[\sum_{k \in B} \mathbf{P}_{l,k}u_k \right] \underbrace{\sum_{j \in B} \mathbf{P}_{i,j}}_{\text{column sums to 1}} - \sum_{j \in B} \mathbf{P}_{i,j}u_j \\
&= \left[\sum_{k \in B} \mathbf{P}_{l,k}u_k \right] - \sum_{k \in B} \mathbf{P}_{i,k}u_k \\
&= \mathop{\mathbb{E}}_{k \sim P(l)} u(k) - \mathop{\mathbb{E}}_{k \sim P(i)} u(k)
\end{aligned}$$

□

Proposition 6.11. *If* A, B *are preference domains that an agent believes to be conditionally independent, then* $\mathrm{Img}_{\mathsf{A}}[\mathsf{B}] = \mathbf{0} = \mathrm{Img}_{\mathsf{B}}[\mathsf{A}]$

Proof. If A and B are conditionally independent, and $\mathbf{P} = \Pr(B \mid A)$ is the stochastic matrix representing the conditional probability, then for some non-negative probability distribution $\vec{p}$ over B, we have:

$$\mathbf{P} = \begin{bmatrix} & \vdots & \\ - & \vec{p} & - \\ - & \vec{p} & - \\ & \vdots & \end{bmatrix} \qquad \mathbf{P}_{i,j} = \vec{p}_j$$

As a result, we calculate:

$$\begin{aligned}\mathrm{Img}_{\mathsf{A}}[\mathsf{B}]_{i,l} &= \mathbf{P}\mathbf{B}\mathbf{P}^T \\ &= \sum_{j,k \in B} \mathbf{P}_{i,j}\mathbf{B}_{j,k}\mathbf{P}_{l,k} \\ &= \sum_{j,k \in B} (\vec{p}_j)\mathbf{B}_{j,k}(\vec{p}_k)\end{aligned}$$

[**todo**: *this is only true for symmetric matrices; argue that this is enough*] □

7 Present Formalism

7.1 Representation

Definition 7.1. A *preference domain* $\mathsf{D} = (D, \vartheta, S, \mathbf{m}_{\mathsf{D}})$ is a non-empty set of alternatives D, some parameter space ϑ, a semiring S, and a method $\mathbf{m}_{\mathsf{D}}$ of getting a preference matrix $\mathbf{D} : D \times D \to S$ from any setting of parameters $\theta \in \vartheta$.

$$\mathbf{m}_{\mathsf{D}} : \vartheta \to M_D(S) \quad \text{or equivalently} \quad \mathbf{m}_{\mathsf{D}} : \vartheta \to D \times D \to S$$

For ease of notation, if D is a preference domain, define $\mathbf{D} := \mathbf{m}_{\mathsf{d}}(\theta)$. To describe the system statically, it is enough to give the set D and the matrix $\mathbf{D}$, but by allowing this more general form, we can deal with compressed representations in a more satisfying way, making gradient updates possible without requiring that agents carry these complete matrices around.

Definition 7.2. A *preference instance* is a preference domain $\mathsf{D} = (D, \vartheta, S, \mathbf{m}_{\mathsf{D}})$ together with a particular value of $\theta \in \vartheta$.

Definition 7.3. A *preference link* $L[\mathsf{A} \to \mathsf{B}]$ between the preference domains $\mathsf{A} = (A, \vartheta_A, S_A, \mathbf{m}_{\mathsf{A}})$ and $\mathsf{B} = (B, \vartheta_B, S_B, \mathbf{m}_{\mathsf{B}})$ is parameter space ϑ and a function

$$L : \vartheta \to (B \times B \to S_B) \to (A \times A \to S_A)$$

Note again that the direction of the link $A \to B$, which is interpreted as the direction of causal consideration in a decision making process, is opposite from the direction of its effect on the matrices; if A impacts B, then what you think of B should influence how you deal with A. We also have to be able to instantiate these links, so we have

Definition 7.4. A *link instance* is a link $L[\mathsf{A} \to \mathsf{B}]$, again together with a particular value $\theta \in \vartheta$. We write $L(\mathbf{B})$ to denote the application of the link L to the parameter θ and matrix $\mathbf{B}$ (yielding a preference matrix $\mathbf{A} \in M_A(S_A)$). Let $\mathrm{LinkInst}[\mathsf{A} \to \mathsf{B}]$ denote the space of link instances from A to B.

At each point in time t, a representation of the agent's state $(\mathcal{D}_t, \mathcal{B}_t)$ consists of

- a collection of preference instances $\mathcal{D}_t$,
- a collection of link instances $\mathcal{B}_t$, each of which is between two preference domains in $\mathcal{D}_t$

$$\mathcal{B}_t \subseteq \prod_{\mathsf{A},\mathsf{B} \in \mathcal{D}_t} \mathrm{LinkInst}[\mathsf{A} \to \mathsf{B}]$$

Remark. To define an agent's state, it is necessary to specify preference and link instances, but the parameter space is only necessary for the update mechanism, described in section 7.2.

7.2 Dynamics

For each link, whose conditional distribution we imagine being parameterized by some latent variables θ, we define the inconsistency at a domain D as the difference between the computed preference matrix $\mathbf{D}$, and the sum of all of the images of contributions from any given link.

$$\zeta_{\mathsf{D}}(\Theta) = \left\| \mathbf{m}_{\mathsf{D}}(\Theta) - \sum_{B[\mathsf{D}\to\mathsf{X}]} B\big[\mathbf{m}_{\mathsf{X}}(\Theta)\big] \right\|_F$$

where Θ contains the parameters required to generate each of these: the link B, and the preferences $\mathbf{X}$ and $\mathbf{D}$. [5] We can also sum across the entire graph to get a measure for the whole model:

$$\zeta(\mathcal{D}, \mathcal{B}) = \sum_{D\in\mathcal{D}} \zeta_\Theta(D)$$

The update rule is now just gradient descent on the parameters Θ with respect to the global inconsistency:

$$\frac{\partial\Theta}{\partial t} = -\eta\nabla_\Theta \sum_{D\in\mathcal{D}} \zeta_\Theta(D)$$

where η is the standard rate parameter for gradient descent.

7.3 A More Concrete Instance: Real Domains with Conditional Links

We will focus on real-valued preference domains, where the links are conditional probability distributions; in this case, every domain $(D, \vartheta, S, \mathbf{m}_{\mathsf{D}}) \in \mathcal{D}_t$ shares a semiring S, which will be either all real numbers $\mathbb{R}$, positive real numbers $\mathbb{R}^+$, or the interval $[0, 1]$, depending on the use case [**todo**: *expand*].

There are many reasonable choices for the parameter space. In the simplest case, we can set $\vartheta = M_D(S)$, so that $\mathbf{m}_{\mathsf{D}}(\theta) := \theta$ can be the identity, and gradient decent corresponds directly to matrix derivatives. On the other hand, if we want to enforce consistency, and object to the idea that each pairwise comparison be considered independently, we can set $\vartheta = \mathbb{R}^D$ to be a utility function on D, so that $\mathbf{m}_{\mathsf{D}}(\theta) := \mathbf{D}^{+\theta}$ (or alternatively $\mathbf{m}_{\mathsf{D}}(\theta) = \mathbf{D}^{\times\theta}$ or $\mathbf{m}_{\mathsf{D}}(\theta) := \mathbf{D}^{\sim\theta}$) is just its matrix representation as discussed in section 6.3.

The link instances, we will think of being conditional probability distributions — that is,

$$\text{LinkInst}[\mathsf{A}\to\mathsf{B}] := A \to \Delta B$$

which is to say, a probability distribution ΔB over the alternatives in B associated each alternative in A. We will write $\Pr(Y|X)$ to denote this conditional distribution $B \in \mathcal{B}$ over Y whose co-domain is X. We can use the conditional probability distribution P (thought of as a stochastic matrix) to generate the link function

$$\begin{aligned} L : (B\times B \to \mathbb{R}) &\to (A \times A \to \mathbb{R}) \\ \mathbf{B} &\mapsto P^T\mathbf{B}P \end{aligned}$$

[5] In the previous version, we defined the consistency of a link $B : X \to \Delta Y$ as

$$\zeta\big(B[X\to Y]\big) = \sigma\left(\left(1 - \frac{|W_X - W_Y|}{W_X + W_Y}\right) \sum_{x,x'\in X}\sum_{y,y'\in Y} X(x,x')Y(y,y')B(x)(y)B(x')(y')\right)$$

where $D(d,d') := \begin{cases} 1 & d \prec_D d' \\ -1 & d \succ_D d' \\ 0 & \text{otherwise} \end{cases}$. [**todo**: *typeset proof of relationship*]

Writing this out more carefully, this is

$$\mathbf{A}_{i,j} := \sum_{b,b' \in B} \Pr(b \mid i)\ \mathbf{B}_{b,b'}\ \Pr(b' \mid j)$$

We define the link instances first, because they are the more important piece of the puzzle. The parameterization for the links corresponds to the entrenchment knobs we want to expose for belief revision, and so when we ultimately tie this into belief revision, gradient descent on parameters will allow us to keep both preferences and reject the belief that they are related. For the examples here, we are primarily concerned with demonstrating preference rather than belief updates, so we will leave this unformalized so we can manipulate link updates with words.

Remark. If $\mathcal{D}$ consists of only a dingle domain D, with the identity distribution $\mathcal{B} = B(x, y) = \delta_{x,y} : D \to D$, then $\mathcal{D}$ is consistent, i.e., $\zeta(\mathcal{D}, \mathcal{B}) = 0$. Similarly, if there are no links, (or only identity links), then the model is consistent, because in each case the sum is vacuously zero.

8 Revisited Examples

For the examples that follow, we will assume that the parameterization ϑ of preference domains is just the set of matrices themselves, unless otherwise stated, and that there is no parameterization for beliefs[6]. As discussed above, we will assume that the semi-ring underlying the preference matrices is the real numbers.

8.1 Framing Problems

Recall example 3.10:

> **Example 8.1.** 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. △

There are two preferences you have, over different alternatives: one, over tax policies when the default number of children is zero, and one where the default number of children is two. These two choices are represented as preference domains, D_0 and D_2, respectively; the assertion that the rich should not receive a bigger tax break from the government, but also should pay more as a surcharge, are the small arrows:

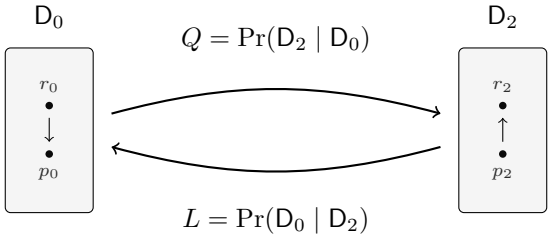
[6] By this, we mean that the parameter set is a final object $\vartheta = \{*\}$, the product of zero variables; there are no parameters to do gradient descent on, and so the belief does not change.

Suppose that our model agent had already made up their mind about these two choices before considering how the choices relate to one another: there are no belief arrows connecting the domains, as above. More concretely still, suppose that the following are the preference matrices on the two domains:

$$\mathbf{D}_0 = \begin{array}{c} \begin{matrix} r_0 & p_0 \end{matrix} \\ \begin{bmatrix} 0 & 3 \\ -3 & 0 \end{bmatrix} \end{array} \begin{matrix} r_0 \\ p_0 \end{matrix} \qquad \mathbf{D}_2 = \begin{array}{c} \begin{matrix} r_2 & p_2 \end{matrix} \\ \begin{bmatrix} 0 & -2 \\ 2 & 0 \end{bmatrix} \end{array} \begin{matrix} r_2 \\ p_2 \end{matrix}$$

This choice of matrices corresponds to an additive representation of a utility difference; the only choice we have made here is to effectively give D_0 a weight of 3 and D_2 a weight of 2. This corresponds to a greater certainty of preferences in the domain D_0 — we make this assumption because in the real tax code that we live with, the default number of children is zero.

Now, we begin to think through the relationship between the two choices. Suddenly we have belief links between the two domains: we begin to suspect that a choice in one domain might impact the other one, represented like this in our diagram:



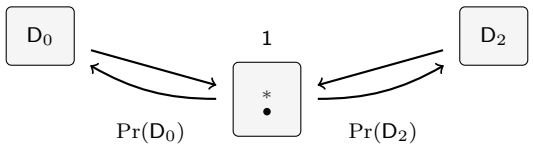
For instance, if our agent explicitly believes that the choices are independent, then by proposition 6.11, they have no effect on one another through the dynamics, which is equivalent to the case where there are no explicit connections at all[7].

Now, after thinking about this a bit harder, we're starting to think that r_0 and r_2 are actually the same option. Suppose we now have 0.9 credence that r_0 would be selected over p_0, if we were to choose r_2 over p_2 (once again, we're not modeling belief revision, so how this happens is not relevant). Looking now at the image of D_0 on D_2 (or vice versa, it doesn't matter), we have

$$L = \Pr(D_0 \mid D_2) = \begin{array}{c} \begin{matrix} r_0 & p_0 \end{matrix} \\ \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix} \end{array} \begin{matrix} r_2 \\ p_2 \end{matrix} \quad \Longrightarrow \quad L(\mathbf{D}_0) = L\mathbf{D}_0L^T = \begin{array}{c} \begin{matrix} r_2 & p_2 \end{matrix} \\ \begin{bmatrix} 0 & 2.4 \\ -2.4 & 0 \end{bmatrix} \end{array} \begin{matrix} r_2 \\ p_2 \end{matrix}$$

[**todo**: *derive gradient, show logistically represented utility function with gradient descent allows for stronger entrenchment via vanishing gradients*]

[7]Note that we can draw this out as a factorization through the (categorically final) singleton preference domain:



8.2

Consider example **??**. At time $t = 0$, we have $\mathsf{Drink} \in \mathcal{D}_t$; with an initial preference for beer over coffee represented by the matrix

$$\mathbf{D} = \begin{array}{c} \begin{matrix} \text{coffee} & \text{beer} \end{matrix} \\ \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \end{array} \begin{matrix} \text{coffee} \\ \text{beer} \end{matrix}$$

This preference is additively represented by a utility function. This encoding is the simplest, which is why we start with it, but it is not sufficient to distinguish between weakly held strong preferences, and strongly held weak preferences.

At time $t = 1$, we remember that caffeine is relevant to the decision, which corresponds to adding a new preference domain considering a choice of caffeine (which we think of as strongly positive, twice as much as our preference for drinks), and a belief that coffee has more caffeine than beer. These are represented, respectively, as

$$\mathbf{C} = \begin{array}{c} \begin{matrix} \text{caff} & \text{no caff} \end{matrix} \\ \begin{bmatrix} 0 & -2 \\ 2 & 0 \end{bmatrix} \end{array} \begin{matrix} \text{caff} \\ \text{no caff} \end{matrix} \qquad \text{and} \qquad L = \Pr(C \mid D) = \begin{array}{c} \begin{matrix} \text{caff} & \text{no caff} \end{matrix} \\ \begin{bmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{bmatrix} \end{array} \begin{matrix} \text{coffee} \\ \text{beer} \end{matrix}$$

As a result, we have

$$L(\mathbf{C}) = L^T \mathbf{C} L = \begin{array}{c} \begin{matrix} \text{coffee} & \text{beer} \end{matrix} \\ \begin{bmatrix} 0 & -1.7 \\ 1.7 & 0 \end{bmatrix} \end{array} \begin{matrix} \text{coffee} \\ \text{beer} \end{matrix}$$

which we interpret as the preference for drinks that we *should* have, looking at it exclusively through the lens of caffeine. Note that 1.7 is not as large as 2, because there is some uncertainty, but points in the opposite direction from our original preference on D. Before we continue with the updating rule, we want to distinguish this case from the framing problem: why should your preference for caffeine not change much? Presumably it is anchored by some reason you want caffeine; let's say this is because we want to get work done — which is to say, we have another domain W, connected like this:

Without modeling the exact dynamics of what happens with the work domain, it will in turn be linked to other things, perhaps with cycles, and this portion of the graph will have a great deal of weight — we will simplify this by just assuming that W does not change, and that the link $L[\mathsf{C} \to \mathsf{W}]$ results in our original preference for caffeine, with small deviations from $\mathbf{C}$ resulting in a much higher impact on the global inconsistency than any of the parts of the picture we've already modeled.

Now, we can apply our update rule:

$$\zeta_{\mathsf{D}}(\mathbf{D}) = \left\| \mathbf{m}_{\mathsf{D}}(\Theta) - \sum_{B[\mathsf{D}\to\mathsf{X}]} B\big[\mathbf{m}_{\mathsf{X}}(\Theta)\big] \right\|_F = \big\| \mathbf{D} - L(\mathbf{C}) \big\|_F$$

$$= \left\| \begin{array}{c} \begin{matrix} \text{coffee} & \text{beer} \end{matrix} \\ \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \end{array} \begin{matrix} \text{coffee} \\ \text{beer} \end{matrix} - \begin{array}{c} \begin{matrix} \text{coffee} & \text{beer} \end{matrix} \\ \begin{bmatrix} 0 & -1.7 \\ 1.7 & 0 \end{bmatrix} \end{array} \begin{matrix} \text{coffee} \\ \text{beer} \end{matrix} \right\|_F$$

[**todo**: *find a satisfying way of writing down the gradient so you can see the C-terms balancing with the W-terms, without just an unreasonable wall of new meaningless symbols from W that I don't want to focus on*]

9 Applications

9.1 Serving Content

Since humans' preferences are not always dynamic, it is a mistake to design recommender systems as though they were. The problem of designing a recommender system is almost always framed as the task of deciding what content to display to what user at a single point in time, based on information about the content and the user, which hides potentially dynamic elements of preferences. Such a model can still accurately predict that a user's preferences have changed by looking at updated user data and considering the new user an entirely different person, but the approach of fixing this issue by trying to get more recent feature data, necessarily lags behind that source of data. While in principle it is may be possible to predict where a person's tastes are likely to go in the future, this kind of thinking mirrors only the recent past.

Similarly, issues with recommender systems zeroing in on a particular niche preference set, and serving content only similar to the content the user has seen before. However, sometimes the similarity of the content is actually a detriment. For instance, if you've already watched an instructional video on integration by parts, you are less likely to click on other such videos, rather than more likely — despite (and indeed because of) similarities between the two videos.

9.2 Specification of Robot Preferences

The problem with simply giving an agent an objective to optimize, is that people are really bad at understanding what optimizing any given objective actually means until they see it. In practice, people mis-specify objectives all the time, and are often mistakenly under the impression that they've incentivized the agent to do something slightly different. There are even numerous fairy tales and parables cautioning against this, which can be summed up with the aphorism "be careful what you wish for"

People don't have to be all that careful what they wish for around their friends and family, however, because they do not express their desires in such clear terms as objective functions, and the wishes they express are not taken literally, but rather in context with the other preferences this person has expressed, and the social context. A theory of preference dynamics would allow for both partial specification and overspecification, which may be a step towards corrigibility.

[**todo**: *simulate this. Try to develop reasonable preferences from unreasonable ones.*]

10 Relation to Other Models

[**todo**: *Joint parents vs aggregate; why it's ok to sacrifice expressive power when domains can be combined*]

10.1 CP Net Problems

The CP semantics are in many ways appealing: the dominance of a over $\bar{a}$ independent of context seems to be a really nice way of capturing the utility brought by a independent of the context of everything else. However, CP Nets are not well-behaved under changes of perspective, and lead to strange results if you think in terms of expected utility. The primary issue is that all of the variables are assumed to be independent, so that the space $\mathcal{W}$ of possible worlds is just the product of all of the individual variables:

$$\mathcal{W} \cong \prod_{X:\mathcal{X}} \Omega_X$$

Depending on how you think about it, there's always an injection one direction— a world results in a setting of all of the variables:

$$\mathcal{W} \to \prod_{X:\mathcal{X}} \Omega_X$$

but even this may not be in keeping with the way we might want to use these models: some variables might not be relevant or well-defined in all contexts. For instance, people use choice of food as an example variable all the time: the variable represents my selection of dinner at restaurant X. But such a variable does not mean anything if I decide not to eat, or if I go to restaurant Y which serves different items. One way to fix this might be to add a special "not applicable" value in the range of each variable like this, but now we've exacerbated our first problem even further: what does the world where ALL variables are set to "not applicable" look like? This failure to have a natural correspondence causes a number of problems:

10.1.1 Incompatibility with Expected Utility

[**todo**: *show:conditional preferences' failure to represent trade-offs.*] [**todo**: *strange representation asymmetry: redefine one to be flipped, get different representative power*]

10.2 Savage Emulation

[**todo**: *Write up P1-P3 in terms of this framework; explore the rest of the postulates*]

References

[Amo+16] Dario Amodei et al. "Concrete problems in AI safety". In: *arXiv preprint arXiv:1606.06565* (2016).

[Bos12] Nick Bostrom. "The superintelligent will: Motivation and instrumental rationality in advanced artificial agents". In: *Minds and Machines* 22.2 (2012), pp. 71–85.

[Bou+04] Craig Boutilier et al. "CP-nets: A tool for representing and reasoning withconditional ceteris paribus preference statements". In: *Journal of artificial intelligence research* 21 (2004), pp. 135–191.

[ED10] Ludwig Elsner and P van den Driessche. "Max-algebra and pairwise comparison matrices, II". In: *Linear Algebra and its Applications* 432.4 (2010), pp. 927–935.

[EV04] Ludwig Elsner and Pauline Van Den Driessche. "Max-algebra and pairwise comparison matrices". In: *Linear Algebra and its Applications* 385 (2004), pp. 47–62.

[Gri18] Tim Griffin. *Lecture notes in Algebraic Path Problems*. Jan. 2018. URL: https://www.cl.cam.ac.uk/teaching/1718/L11/L11_2017_lecture_3_2up.pdf.

[Jef90] Richard C Jeffrey. *The logic of decision*. University of Chicago Press, 1990.

[Sav72] Leonard J Savage. *The foundations of statistics*. Courier Corporation, 1972.