Infra-Bayesianism Unwrapped

by **Adam Shimi** 20th Jan 2021

24

Distillation & Pedagogy | Logical Uncertainty | Decision Theory | Infra-Bayesianism | AI | Frontpage

Introduction

Infra-Bayesianism° is a recent theoretical framework in AI Alignment, coming from Vanessa Kosoy and Diffractor (Alexander Appel). It provides the groundwork for a learning theory of RL in the non-realizable case (when the hypothesis is not included in the hypothesis space), which ensures that updates don't throw away useful information, and which also satisfies important decision-theoretic properties playing a role in Newcomb-like problems°.

Unfortunately, this sequence of posts is really dense, and uses a lot of advanced maths in a very "textbook" approach; it's thus hard to understand fully. This comes not from the lack of intuition (Diffractor and Vanessa are both very good at providing intuitions for every idea), but from the sheer complexity of the theory, as well as implicit (or quickly mentioned) links with previous research.

Thus my goal in this post is to give enough details for connecting the intuitions provided and the actual results, as well as the place of Infra-Bayesianism within the literature. I will not explain every proof and every result (if only because I'm not sure of my understanding of all of them). But I hope by the end of this post, you have a clearer map of Infra-Bayesianism, one good enough to dig into the posts themselves.

This post is splitted into three section

- Section 1 explores the context of Infra-Bayesianism, the problem it attempts to solve, and some of the relevant literature. You can see it as unwrapping this section° in Introduction to The Infra-Bayesianism Sequence°.
- Section 2 gives a bird's-eye view of Infra-Bayesianism; a map to navigate through the sequence.
- Section 3 follows one path through this map: the one focused on decision-theoretic properties for Newcomb-like problems.

Reading advice: The reader is king/queen, but I still believe that there are two main ways to read this post to take something out of it:

- Read it quickly, in one go to get a general idea of Infra-Bayesianism and a very high-level map.
- Read it in detail, for a time between 2 hours and a whole afternoon, to get every detail explained here. If you do so, I really believe that you'll have a quite detailed map of Infra-Bayesianism, enough to explore the sequence by yourself without getting lost in the mathematical jungle.

Thanks to Vanessa and Diffractor for going above and beyond in answering all my questions and providing feedback on this post. Thanks to Jérémy Perret for feedback on this post.

Section 1: Why is Infra-Bayesianism Important?

Cruxes

Before going into the problem tackled by Infra-Bayesianism, I want to give some context in which to judge the value of this research.

AI Alignment is not yet a unified field; among other things, this means that a lot of researchers disagree on what one should work on, what constitutes a good solution, what is useful and what isn't. So the first thing I look for when encountering a new piece of AI Alignment research is its set of underlying assumptions. In rationalist parlance, we would say the cruxes.

Infra-Bayesianism, just like most of Vanessa's research, relies on three main cruxes made explicit in her research agenda°:

- **(AI Alignment requires more than experimental guarantees)** I would say this one is almost a consensus among AI Alignment researchers. The fact that so many things can go wrong, at scales where we'll be unable to act, or even notice the issue, means that just running experiments and checking that everything is ok isn't enough. Moreover, in some cases if the experiment fails, it's game over. So we want explanations and models for why the AI we build will indeed be aligned.
- **(Such guarantees must come from a mathematical theory of AI Alignment)** This one, on the other hand, is a major source of disagreement. Most researchers agree that

having mathematical results for the kind of guarantees we care about would be great -- some are simply pessimistic that such results exist. See for example Rohin's position in this discussion°.

- **(The best candidate for such a theory is a theory of RL)** Lastly, this crux is a bit more difficult to put into context, because it relies on the previous, controversial crux. That being said, many of the people unconvinced by the previous crux seem to focus their research effort into prosaic AGI, that is on DeepRL. So working on RL appears rather accepted too.

In summary, Infra-Bayesianism makes sense as a component of a theory of RL, with the goal of proving formal guarantees on alignment and safety. Even if you disagree with some of these cruxes, I feel that being aware of them will help you understand these posts better.

Non-realizability: the heart of the matter

The main idea motivating Infra-Bayesianism is the issue of non-realizability. Realizability is a common assumption on learning tasks, where the thing we are trying to learn (the function to approximate for example) is part of the hypothesis space considered. Recall that because of fundamental results like the no-free-lunch theorems, learning algorithms cannot consider all possible hypotheses equally -- they must have inductive biases which reduce the hypothesis space and order its elements. Thus even in standard ML, realizability is a pretty strong assumption.

And when you go from learning a known function (like XOR) to learning a complex feature of the real world, then another problem emerges, related to embedded agency: the learning agent is embedded into the world it wants to model, and so is smaller in an intuitive sense. Thus assuming that the hypothesis space considered by the learning algorithm (which is in some sense represented inside the algorithm) contains the function learned (the real world) becomes really improbable, at least from a computational complexity perspective.

One important detail that I missed at first when thinking about non-realizability is that the issue comes from assuming that the true hypothesis is one of the efficiently computable hypotheses which form your hypothesis space. So we're still in the non-realizable setting when you might know the true hypothesis, but it's either uncomputable or prohibitively expensive.

Going back to Infra-Bayesianism, non-realizability is a necessity for any practical mathematical theory of RL. But as explained in the first post° of the sequence, there is not that many results on learning theory for RL:

For offline and online learning there are classical results in the non-realizable setting, in particular VC theory naturally extends to the non-realizable setting. However, for reinforcement learning there are few analogous results. Even for passive Bayesian inference, the best non-realizable result found in our literature search is Shalizi's which relies on ergodicity assumptions about the true environment. Since reinforcement learning is the relevant setting for AGI and alignment theory, this poses a problem.

If you're like me, you get the previous paragraph, with the possible exception of the part about "ergodicity assumptions". Such assumptions, roughly speaking, mean that the distribution of the stochastic process (here the real world) eventually stabilizes to a fixed distribution. Which will probably happen, around the heat-death of the universe. So it's still a very oversimplified assumption, that Infra-Bayesiansm removes.

Now, the AI Alignment literature contains a well-known example of a non-realizable approach: Logical Induction. The quick summary is that Logical Induction deals with predicting logical consequences of known facts that are not yet accessible due to computational limits, in ways that ensure mistakes cannot be exploited for an infinite amount of "money" (in a market setting where predictions decide the "prices"). Logical inductors (algorithms solving Logical Induction) deals with a non-realizable setting because the guarantee they provide (non-exploitation) doesn't depend on the "true" probability distribution. Equivalently, logical inductors attempt to approximate a probability distribution over logical sentences that is uncomputable, and that has no computable approximation in full.

Building on Logical Induction (and a parallel line of research, which includes the idea of Defensive Forecasting), a previous paper by Vanessa titled Forecasting Using Incomplete Models, extended these ideas to more general, abstract and continuous settings (instead of just logic). The paper still deals with non-realizability, despite having guarantees that depend on the true hypothesis. This is because the guarantees have premises about whether the true hypothesis is inside an efficiently computable set of hypotheses (a convex set), instead of requiring that the true hypothesis is itself efficiently computable. So instead of having a handful of hypotheses we can compute and saying "it's one of them", Forecasting Using Incomplete Models uses efficiently computable properties of hypotheses, and say that if the true hypothesis satisfies one of these properties, then an efficiently computable hypothesis with the same guarantees will be learned.

This idea of sets of probability distributions also appears in previous takes on imprecise probabilities, notably in Walley's Statistical Reasoning with Imprecise Probabilities and Peng's Nonlinear Expectations and Stochastic Calculus under Uncertainty. That being said, Vanessa and Diffractor heard about these only after finishing most of the research on Infra-Bayesianism. These previous works on imprecise probabilities also don't deal with the decision theory aspects of Infra-Bayesianism.

Lastly, all the ideas presented for prediction above, from logical induction to imprecise probabilities, provide guarantees about the precision of prediction. But for a theory of RL, what we want are guarantees about expected utility. This leads directly to Infra-Bayesianism.

Section 2: What is Infra-Bayesianism, and What can it do?

Bird's-eye View of Infra-Bayesianism

The main object of Infra-Bayesianism is the infradistribution (Definition 7° in Basic Inframeasure Theory°): a set of "pimped up" probability distributions called sa-measures. These sa-measures capture information like the weight of the corresponding distribution in the infradistribution and the off-history utility, which prove crucial for decision theoretic reasoning further down the line (in Belief Functions and Decision Theory°). Infradistributions themselves satisfy many conditions (recapped here° in Basic Inframeasure Theory°), which serves to ensure it's the kind of computable property of environments/distributions that we want for our incomplete models.

Basic Inframeasure Theory°, the first technical post in the sequence, defines everything mentioned previously from the ground up. It also brushes up on the measure theory and functional analysis used in the results, as well as show more advanced results like a notion of update (Definition 11°) that takes into account what each sa-measure predicted, the corresponding Bayes Theorem for infradistributions (Theorem 6°), a duality result which allow manipulation of infradistributions as concave, monotone, and uniformly continuous functionals (Theorem 4°), and a lot of others useful theoretical constructions and properties (see for example the section Additional Constructions°).

The next post, Belief Functions and Decision Theory°, focuses on using Infra-Bayesianism in a decision theoretic and learning theoretic setting. At least the decision theoretic part is the subject of Section 3 in the present post, but before that, we need to go into more details about some basic parts of inframeasure theory.

(Between the first draft of this post and the final version, Vanessa and Diffractor published a new post called Less Basic Inframeasure Theory°. Its focus on advanced results means I won't discuss it further in this post)

Maxmin Expected Utility: Knightian Uncertainty and Murphy

Recall that we want to build a theory of RL. This takes the form of guarantees on the expected utility. There's only one problem: we don't have a distribution over environments on which to take the expectation!

As defined above, an infradistribution is a **set** of probability distributions (technically sa-measures, but that's not important here). We thus find ourselves in the setting of Knightian uncertainty: we only know the possible "worlds", not their respective probability. This fits with the fact that in the real world, we don't have access to clean probabilities between the different environments we consider.

As theoretical computer scientists, Vanessa and Diffractor are fundamentally pessimistic: they want worst-case guarantees. Within a probabilistic setting, even our crowd of paranoid theoretical computer scientists will get behind a guarantee with a good enough probability. But recall that we have Knightian uncertainty! So we don't have a quantitative measure of our uncertainty.

Therefore, the only way to have a meaningful guarantee is to assume an adversarial setting: Murphy, as he's named in the sequence, chooses the worst environment possible for us. And we want a policy that maximizes the expected utility within the worst possible environment. That is, we take the maxmin expected utility over all environments considered.

To summarize, we want to derive guarantees about the maxmin expected utility of the policy learned.

From Probability Distributions to Sa-Measures

So we want guarantees on maxmin expected utility for our given infradistributions. The last detail that's missing concerns the elements of infradistributions: sa-measures. What are they? Why do we need them?

The answer to both questions comes from considering updates. Intuitively, we want to use an infradistribution just like a prior over environments. Following the analogy, we might wonder how to update after an action is taken and a new observation comes in. For a prior, you do a simple bayesian update of the distribution. But what do you do for an infradistribution?

Since it is basically a set of distributions, the obvious idea is to update every distribution (every environment in the set) independently. This has two big problems: loss of information and dynamic inconsistency

Relative probabilities of different environments

In a normal Bayesian update, if an environment predicted the current observation with a higher probability than another environment, then you would update your distribution in favor of the former environment. But our naive updates for infradistributions fails on this count: both environments would be updated by themselves, and then put in a set.
Infra-Bayesianism's solution for that is to consider environments as scaled distributions instead. The scaling factor plays the role of the probability in a distribution, but without some of the more stringent constraints.

Now, these scaled measures don't have a name, because they're not the final form of environments in Infra-Bayesianism.

Dynamic Consistency

Even with scaled measures, there is still an issue: dynamic inconsistency. Put simply, dynamic inconsistency is when the action made after some history is not the one that would have been decided by the optimal policy from the start.

For those of you that know a lot of decision theory, this is related to the idea of commitment, and how they can ensure good decision theoretic properties.

For others, like me, the main hurdle for understanding dynamic consistency is to see how deciding the best action at each step could be suboptimal, if you can be predicted well enough. And the example that drives that home for me is Parfit's hitchhiker.

> You're stranded in the desert, and a car stops near you. The driver can get you to the next city, as long as you promise to give him a reward when you reach civilization. Also very important, the driver is pretty good at reading other human beings.

Now, if you're the kind of person that makes the optimal decision at each step, you're the kind of person that would promise to give a reward, and then not give it when you reach your destination. But the driver can see that, and thus leaves you in the desert. In that case, it would have been optimal to commit to give the reward and not defect at your destination.

Another scenario, slightly less obvious, is a setting where Murphy can choose between two different environments, such that the maxmin expected utility of choosing the optimal choice at each step is lower than for another policy. Vanessa and Diffractor give such an example in the section Motivating sa-measures° of Introduction To The Infra-Bayesianism Sequence°.

The trick is that you need to keep in mind what expected utility you would have if you were not in the history you're seeing. That's because at each step, you want to take the action that maximizes the minimal expected utility over the whole environment, not just the environment starting where you are.

Vanessa and Diffractor call this the "off-history" utility, which they combine with the scaled measure to get an a-measure (Definition 3° in Basic Inframeasure Theory°). There's a last step, that lets the measure be negative as long as the off-history utility term is bigger than the absolute value of any negative measure: this is an sa-measure (Definition 2° in Basic Inframeasure Theory°). But that's mostly relevant for the math, less for the intuitions.

So to get dynamic consistency, one needs to replace distributions in the sets with a-measures or sa-measures, and then maintain the right information appropriately. This is why the definition of infradistributions uses them.

Interestingly, doing so is coherent with Updateless Decision Theory°, the main proposal for a decision theory that deals with Newcomb-like or Parfit's hitchhiker types of problems. Note that we didn't build any of the concepts in order to get back UDT. It's simply a consequence of wanting to maxmin expected utility in this context.

UDT also helps with understanding the points of updates despite dynamic consistency: instead of asking for a commitment at the beginning of time for anything that might happen, dynamically consistent updates allows decisions to be computed online while still being coherent with the ideal precommitted decision. (It doesn't solve the problem of computing the utility off-history, though)

Section 3: One Path Through Infra-Bayesianism, Newcomb-like Problems and Decision Theory

Lastly, I want to focus on one of the many paths through Infra-Bayesianism. Why this one? Because I feel it is the most concrete I could find, and it points towards non obvious links (for me at least) about decision theory.

This path starts in the third post of the sequence, Belief Function and Decision Theory°

Beliefs Functions and their Unexpected Consequences

Beliefs functions (Definition 11° in Belief Function and Decision Theory°) are functions which take as input a partial policy (according to Definition 4°), and return a set of a-measures (according to the definitions in Basic Inframeasure Theory° mentioned above) on the outcome set of this partial policy (according to Definition 8°).

We have already seen a-measures in the previous sections: they are built from a scaled distribution (here over outcomes) and a scalar term that tracks the off-history utility (to maintain dynamical consistency). For the rest of the new terms, here are the simple explanations.

- An o-history h (Definition 3°) is a sequence of alternating observations and actions, that ends with an observation. These are the input to policies, which then return the next action to take.
- A partial policy π_{pa} (Definition 4°) is a partial function from o-histories to actions, such that π_{pa} is defined coherently with the prefixes of histories on which it is defined: if there is an o-history h such that $\pi_{pa}(h)$ is well defined, then for every prefix of h of the form $h'a$, we have $\pi_{pa}(h') = a$. Basically, a partial policy is defined on increasing o-histories, until it's not defined anymore.
- The outcome set $F(\pi_{pa})$ (Definition 8°), is the set of o-histories that are not in the domain of π_{pa}, but which have all of their prefixes in it, and the output of π_{pa} are the coherent actions for these prefixes. These are the o-histories where π_{pa} stops (or its infinite histories)

To summarize, a belief function takes a policy, which gives new actions from histories ending in observations, and returns a property on distributions over the final histories of this policy. This generalizes a function that takes a policy and returns a distribution over histories.

Now, a confusing part of the Belief Function and Decision Theory° post is that it doesn't explicitly tell you that this set of a-measures over outcomes actually forms an infradistribution, which is the main mathematical object of Infra-Bayesianism. And to be exact, the outputs of a belief function are guaranteed to be infradistributions only if the belief function satisfies the condition listed here°. Some of these conditions follow directly from the corresponding conditions for infradistributions; others depend on the Nirvana trick, that we will delve into later; still others are not that important for understanding the gist of Infra-Bayesianism.

So at this point in the post, we can go back to Basic Inframeasure Theory° and look at the formal definition of infradistributions. Indeed, such a definition is fundamental for using belief functions as analogous to environments (functions sending policies to a distribution over histories).

Easier Infradistributions: the Finite Case

The general case presented in Basic Inframeasure Theory° considers measures, a-measures and sa-measures defined over potentially infinite sets (the outcome set might be infinite, for example if the policy is defined for every o-history). This requires assumptions on the structure of the set (compactness for example), and forces the use of complex properties of the space of measures (being a Banach space among other things), which ultimately warrants the use of functional analysis, the extension of linear algebra to infinite dimensional spaces.

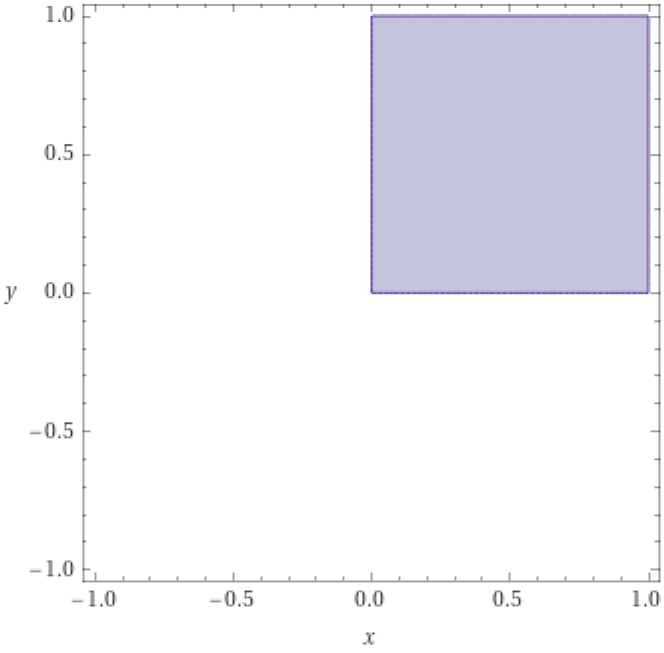
Personally, I'm not well read enough in measure theory and functional analysis to follow everything without going back and forth between twenty Wikipedia pages, and even then I had trouble keeping with the high level abstractions.

Fortunately, there is a way to simplify tremendously the objects with which we work: assume the finiteness of the set on which measures are defined. This can be done naturally in the case of outcome sets, by considering $X_n = F_n(\pi_{pa})$, the set of outcomes of length $\leq n$.

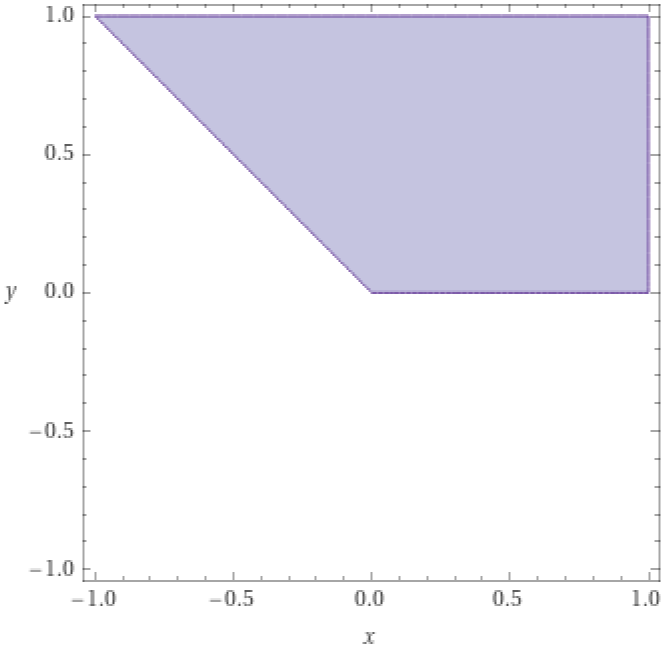
In that context, a measure over X_n is equivalent to a function from a finite domain to $\mathbb{R}^+$; which is equivalent to a point in $\left(\mathbb{R}^+\right)^{|X_n|}$. So the space of measures over X_n is just the Euclidean space of dimension $|X_n|$. We're back into linear algebra!

Now geometrical intuition can come to our help. Take Definition 2° of an sa-measure: it is just a point of $\left(\mathbb{R}^+\right)^{|X_n|+1}$ such that the sum of the negative numbers among its first $|X_n|$ components is less in absolute value than the last component. And an a-measure (from Definition 3°) is an sa-measure where every component is non-negative. The sets $\mathcal{M}^{sa}(X_n)$ and $\mathcal{M}^a(X_n)$ are then respectively the sets of all sa-measures and the sets all a-measures.

We can even visualize them pretty easily (with $|X_n| = 1$):



$M^a(X_n)$



$M^{sa}(X_n)$

There's one more definition to go through before attacking infradistributions: the definition of the **expectation of some continuous function from X_n to $\mathbb{R}$ by a set of sa-measures B**. This is described by Vanessa and Diffractor as the behavior of f (continuous from X_n to $[0,1]$) according to B. Definition 4° gives $\mathbb{E}_B(f)$ as the infinimum of $m(f)+b$ for $(m,b)\in B$. And in our finite case, $m(f)=\sum_{x\in X_n} m(x)f(x)$. So $\mathbb{E}_B(f)$ can be rewritten as the infinimum of $\sum_{x\in X_n} m(x)f(x)+b$ for $(m,b)\in B$.

Intuitively, $\mathbb{E}_B(f)$ represents the worst expected utility possible over B, where f is the utility function. This fits with our previous discussion of Knightian Uncertainty and Murphy, because we assume that the environment picked (the sa-measure) is the worst possible for us. That is, the one with the worst expected utility.

Geometrically in our finite setting, this is the smallest dot product of a point in $\bar{B}$ with the point of $\left(\mathbb{R}^+\right)^{|X_n|+1}$ which has for its first $|X_n|$ components the values of f for the corresponding element of X_n, and for its last component 1.

We can finally go to infradistributions: an infradistribution B is just a set of sa-measures satisfying some conditions. I'll now go through them, and try to provide as much intuition as possible.

- **(Condition 1, Nonemptiness)**: $B\neq\emptyset$ This is without a doubt the most complex condition here, but I have faith that you can make sense of it by yourself.
- **(Condition 2, Closure)** $B=\bar{B}$ This condition says that B contains its limit points. Another way to see it is that if you have a set B and you want to make it into an infradistribution, you need to take the closure of B -- add the limit points of B to the set. Why can we do that? Because the definition of expectation uses an infinimum over B, which is basically a minimum over $\bar{B}$. So the expectation, which captures the behavior of utility functions on our set, already takes into account the limit points of B. Adding them will thus maintain all expectations, and not change anything about the behavior of the set.
  Why do it then? There are two ways to see it. First, adding the limit points makes the study of B easier, because closed sets are nicer than generic sets. Among other things, the expectation is now easier to compute, because it doesn't involve taking limits. And second, if B and $\bar{B}$ are distinct, but have the same behavior according to expectations, then they should be collapsed together in some way. (This is Desideratum 2° from Introduction To The Infra-Bayesianism Sequence°)

- **(Condition 3, Convexity)** $B = convexHull(B)$ This is the same kind of condition that the previous one, but instead of adding the limit points, we add the convex combinations of points in B: $(m', b') = \lambda(m_1, b_1) + (1-\lambda)(m_2, b_2)$, for $\lambda \in [0,1]$ and $(m_1, b_1), (m_2, b_2) \in \mathcal{M}^{sa}(X_n)$. We can add such points because $m'(f) + b' = \lambda(m_1(f) + b_1) + (1-\lambda)(m_2(f) + b_2)$. One of the two components must be smaller or equal to than the other; without loss of generality, let's say it's $m_1(f) + b_1$. Then $\lambda(m_1(f) + b_1) + (1-\lambda)(m_2(f) + b_2)$ $\geq \lambda(m_1(f) + b_1) + (1-\lambda)(m_1(f) + b_1) = m_1(f) + b_1$.
  Hence (m', b') is not changing the expectation for any f.
- **(Condition 4, Upper-Completion)** $B = B + \mathcal{M}^{sa}(X_n)$ Once again, a condition adds points to the set. This one adds all points formed by the sum of an element of B and an element of $\mathcal{M}^{sa}(X_n)$. The reason why it's possible is even more intuitive here: we're adding points that have strictly more measure and expected utility, so they don't influence the minimum in the expected value definition, and thus don't change the behavior of the set.
- **(Condition 5, Minimal-positivity)** $B^{min} \subseteq \mathcal{M}^a(X_n)$ If your set is formed by summing some points with all of $\mathcal{M}^{sa}(X_n)$, what is a minimal set of points that would generate your set through this sum? These are the minimal points of B, noted B^{min}. In fact, there is only one such set, because a minimal point cannot be generated from any other point of the set summed with a point of $\mathcal{M}^{sa}(X_n)$.
  This condition requires that such minimal points have no negative measure. So there is no element of X_n for which they return a negative number. This is a slightly less straightforward condition to motivate, because it stems from the maths. Basically, a positive measure is just a scaled probability distribution, so it behaves nicely for a lot of purposes. Whereas not all signed measures can be rescaled to probability distribution. So Infra-Bayesianism uses "negative probabilities" for some computations and reasoning (notably to have a stronger upper-closure property), but the end results are really scaled probabilities. This is why requiring minimal points to be a-measures makes sense.
- **(Condition 6a, Minimal-boundedness)** $\exists C$ a compact set such that $B^{min} \subseteq C$ In all honesty, I don't know exactly where this condition matters. The original post says that compactness is used in the proofs, and it is a pretty useful mathematical assumption in general. But I don't know exactly where it pans out, and I'm not convinced that you need to know it for getting the gist of Infra-Bayesianism.
  I'll thus ask you to accept this condition as a mathematical simplification without much philosophical meaning.

(The actual condition used in the definition of infradistribution is 6b: $f \mapsto \mathbb{E}_B(f)$ is uniformly continuous. But it is similarly a mathematical condition without much philosophical weight).

- **(Condition 7, Normalization)** $\mathbb{E}_B(1) = 1 \wedge \mathbb{E}_B(0) = 0$ This last condition on the other hand has more to tell. Recall that $\mathbb{E}_B(f)$ captures the expected utility over B, using f as a utility function. Since by hypothesis the utility of a state is in $[0, 1]$, the function 0 with 0 utility at every state represents the worst-case utility, and the function 1 with utility 1 at every state represents the best-case utility.
  The conditions then simply say that if no state is worth any utility, the expected utility is 0 ; and if all states have maximal utility, then the expected utility is maximal too, at 1. So our expected utility lies between 0 and 1, and are normalized.

Armed with these conditions, we now understand Definition 7° of $\square X_n$, the set of infradistributions: it contains all the set of sa-measures that satisfy the conditions above.

Another Perspective on Infradistribution: Duality

There is another way to think about infradistributions: as functionals (in this case, applications from functions to $\mathbb{R}$) with specific properties. This duality is crucial in many proofs and to build a better intuition of Infra-Bayesianism.

Given an infradistribution as a set B, how do we get its dual version? Easy: it's the function h defined by $h(f) = \mathbb{E}_B(f)$. So the expectation with regard to our set B is the other way to see and define the infradistribution. Theorem 4° states this correspondence, as well as the properties that h gets from being defined in this way through B:

> **Theorem 4, LF-duality, Sets to Functionals:** *If B is an infradistribution/bounded infradistribution, then $h : f \mapsto \mathbb{B}(f)$ is concave, monotone, uniformly continuous/Lipschitz over $C(X, [0, 1])$, $h(0) = 0$,$h(1) = 1$, and $range(f) \not\subseteq [0, 1] \implies h(f) = -\infty$.*

Let's look at the properties of h.

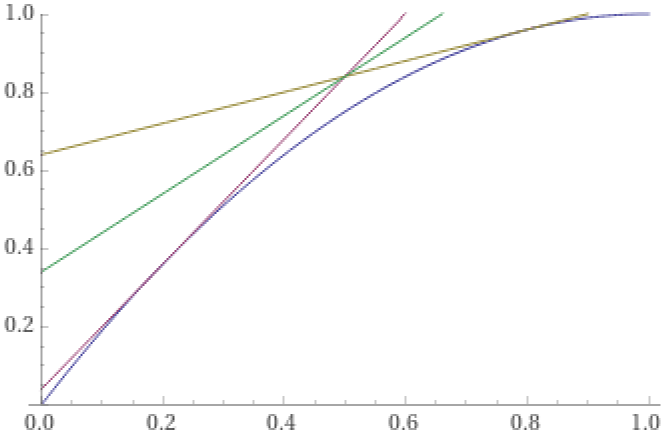
- **(Concavity)** h is a concave function -- it's shaped like a hill when seen in few dimensions. This comes simply from the definition of the expectation and some elementary algebraic manipulations. Notably, it doesn't depend on properties of B.

- **(Monotony)** If $f \leq g$ then $h(f) \leq h(g)$. First recall that the usual order in function space with a partial order as codomain is: $f \leq g \iff \forall x : f(x) \leq g(x)$. So intuitively, this means that if the utility by g is greater or equal than the one by f for every outcome in X_n, then the expected utility for g is greater or equal than the expected from f. That makes a lot of sense to me.
  The reason it holds for h comes from the fact that expectation only depends on the minimal points of an infradistribution (Proposition 3°, just after the definition of minimal points). And recall that B satisfies Condition 5 as an infradistribution: its minimal points are all a-measures. This matters because that ensures that in $m(f) + b = \sum_{x \in X_n} m(x)f(x) + b$, the $m(x)$ terms are all strictly positive (for $(m, b) \in B^{min}$). Since utility functions are also positive, this means that $f \leq g$ entails $h(f) \leq h(g)$ -- because $\forall (m, b) \in B^{min} : m(f) + b \leq m(g) + b$.
  So this actually depends on B being an infradistribution, specifically the condition that $B^{min} \subseteq \mathcal{M}^a(X_n)$.
- **(Uniform continuity)** h is uniformly continuous. Well that's literally Condition 5b for infradistributions.
- **(Normalization)** $h(0) = 0$ and $h(1) = 1$. This is also literally a condition on infradistributions, Condition 7.
- **(Range)** $range(f) \not\subseteq [0, 1] \implies h(f) = -\infty$. So the only functions it makes sense to consider are ones constrained to $[0, 1]$. This property follows from Condition 4 on B, upper completion. Indeed, say $x \in X_n$ is such that $f(x) > 1$. Then given any sa-measure in B, we can add to it as many times as we want the sa-measure with -1 measure on x and $b = 1$ (to compensate), thanks to upper completion. Hence we can create sa-measures in B with smaller and smaller $m(f) + b$ forever, which means that taking the infinimum, $h(f) = -\infty$.

To summarize, Conditions 4 to 7 for infradistributions as sets do most of the work, while Conditions 1 to 3 are not considered since they just increase the size of the set without changing the expectation (technically Condition 1 is necessary everywhere, but it's trivial).

Like a healthy relationship, a good duality goes both ways. Hence Theorem 5°, which shows how to get an infradistribution as a set from a infradistribution as a functional (satisfying the conditions studied above). The proof of this one is way more involved, which is why I won't go into it.

That being said, there is a nice way to visualize the set of sa-measures coming from an expectation in the finite dimensional case. Let's say $|X_n| = 1$. So there is only one outcome x. Let's say we have an h satisfying all the properties above. Notably, it's concave. Then the sa-measures of the corresponding infradistribution as a set are all the pairs $(m(x), b)$ such that $m(x)f(x) + b \geq h(f)$ for all f. Visually, any line above h (which is basically a function of $[0, 1]$, since f is completely determined by its output for x).



In this plot, h is the blue function, and all other functions correspond to sa-measures in the dual "infradistribution as a set". This provides a really cool geometrical intuition for some conditions on infradistributions. For example, upper completeness only means that we can add any line to one of our lines/sa-measures, and we'll still be above h. Or minimal points being a-measures means that they are the tangents of h (like the pink and yellow one on the plot). And it generalizes in higher dimensions, by replacing lines with hyperplanes.

(To be clear, I didn't come up with this geometric perspective; Diffractor explained it to me during a discussion about the duality).

So infradistributions are both sets of sa-measures and functionals, both satisfying specific conditions. The functional perspective is cleaner for proofs, but I'll keep with the set perspective in the rest of this post.

Back to Belief Function: Causality and Nirvana

Recall that "nice" belief functions return infradistributions on the outcome set of a policy. This is never stated explicitly in Belief Function and Decision Theory°, but follows from the first conditions on belief functions from this section°.

Other conditions matter for manipulating belief functions, like consistency and Hausdorff-continuity. But the point of this section isn't to make you master Belief Function and Decision Theory°; it's to give you a path through it. And the last big idea on the path is Causality, and it's relation to the Nirvana Trick.

Indeed, if you've read the sequence before, you might be surprised by me not mentioning the Nirvana trick already. My reason is that I only understood it correctly after getting causality, and causality requires all the background I layed out already.

The Nirvana Trick: Making Murphy Useful

Recall that we have Knightian Uncertainty over the environments we consider. So instead of maximizing the expected utility over a distribution of environments, we use worst-case reasoning, by assuming the environment is chosen by an adversary Murphy. This is a pretty neat setting, until we consider environments that depend on the policy. This happens notably in Newcomb-like problems° (of which Parfit's Hitchhiker is an example), which are an important fighting ground for decision-theories.

Now, it's not so much that representing such environments is impossible; instead, it's that what we think of as environments is usually simpler. Notably, what happens depends only on the action taken by the policy, not on the one it could have taken in other situations. This is also a setting where our intuitions about decisions are notably simpler, because we don't have to think about predictions and causality in their full extent.

The Nirvana trick can be seen as a way to keep this intuition of environments, while still having a dependence of the environment on the policy. It starts with the policy-dependent environment, and then creates one policy-independent environment for each policy, by hard-coding this policy in the parameter slot of the policy-dependent environment. But that doesn't guarantee that the hardcoded policy will match the actual policy. This is where Nirvana appears: if the policy acts differently than the hardcoded policy, it "goes to Nirvana", meaning it gets maximum return (either through an infinite reward at that step or with a reward of 1 for each future step). Murphy, which wants to minimize your utility, will thus never choose an environment where Nirvana can be reached, that is never choose the ones with a different policy in the parameter slot.

To understand better the use of the Nirvana trick, we need to define different kinds of belief functions (called hypotheses) such that adding or removing Nirvana goes from one to the other.

Causality, Pseudocausality and Acausality

The three types of belief functions (called hypotheses) considered in Belief Function and Decision Theory° are causal, pseudocausal and acausal. Intuitively, a causal hypothesis corresponds to a set of environments which doesn't depend on the policy; a pseudocausal hypothesis corresponds to a set of environments which depends on the policy in some imperfect way; and an acausal hypothesis corresponds to a set of environments completely and exactly determined by the policy.

Causality can be made formal through the introduction of outcome functions (Definition 15°), functions from a policy to a *single* sa-measure on the outcome set of this policy. On the other hand, recall that belief functions return an infradistribution, which is a set of sa-measures on the same set. Compared with the usual Bayesian setting, a belief function returns something analogous to a probability distribution over probability distribution over histories, while an outcome function returns something analogous to a single probability distribution over histories. An outcome function thus plays the role of an environment, which takes in a policy and gives a distribution over the outcomes/histories generated.

There is one additional subtlety about outcome functions that plays a big role in the rest of the formalism. If you look at Definition 15° in Belief Function and Decision Theory°, it requires something about the projection of partial policies. The projection mapping (Definition 9°) sends a sa-measure over a policy π_1 to a sa-measure over a policy π_2, if π_1 is defined on strictly more histories than π_2 and they agree when they're both defined. Basically, if π_1 extends π_2, we can project back a measure over the outcomes of π_1 to a measure over the outcomes of π_2, by summing the measure of all outcomes of π_1 that share as prefix a given outcome of π_2.

Outcome functions must agree with that, in the sense that the outcome function applied to π_2 must return the projection of what the outcome function returns when applied to π_1. In that sense it's a real environment, because if you extend a policy, it only splits the probability given to each prefix, not moves probability between prefixes.

Causal, pseudocausal and acausal hypotheses are defined through constraints related to the outcome functions corresponding to a belief function. They all share the first 9 Conditions on belief functions given here°.

- **Causality** requires Condition C°: that for every policy π_{pa} and every sa-measure M from $\theta(\pi_{pa})$ (the application of the belief function to π_{pa}), there is an outcome function f such that the output of f on π_{pa} is M, and the output of f on all other policies is included in the corresponding output of the belief function.

So for every distribution over history (for a policy), there is an environment sending this policy to this distribution, and any other policy to an accepted distribution over histories (by the belief function). Take all these outcome functions, and you get a set of environments that completely capture the behavior of the belief function.
Remember that with outcome functions comes a constraint on projections. This constraint does the heavy lifting here: it forces the environments to be policy-independent. This is because it ensures that revealing more actions of the policy (extending it) cannot change what happened before these new actions. Changing an empty policy to the policy that always 1-box in a causal version of transparent Newcomb (so no Omega) doesn't change what is possible to do in a given environment; it merely splits the probability into the extended outcomes.

- **Pseudocausality** requires Condition P°. This condition is slightly more complex. First, it requires the belief function to be Nirvana-free: to not consider outcomes leading to Nirvana. So when thinking about pseudocausality, we don't use the Nirvana trick.
Starting with a setting without Nirvana, pseudocausality asks that for every sa-measure M from the infradistribution of a given partial policy π, M is also included in the infradistribution for any other policy that generates only outcomes for which M has non-zero measure. So if M only cares about outcomes where both policies agree, it should be either in no infradistribution or in both. This property captures the fact that if two distinct policies don't reveal their difference by taking different actions in a given environment, then this environment should be possible for both or none, but not just one.
It's a weaker form of policy-independency than Condition C°, because it removes the constraint on projection completely -- this definition doesn't even use outcome functions. Note though that there is still some constraint on projection, common to all hypotheses, in the form of Condition 7° on belief functions, consistency. But it's not about policy-dependency, so I won't talk in detail about it.
Perhaps my biggest initial confusion with Condition P° came from the fact that transparent Newcomb with imperfect prediction satisfies it. Intuitively, the environment there should definitely depend on the policy, including on what is done in the other branch. But the trick lies in realizing that imperfect prediction means that no possibility for the transparent box (empty or full) can have probability 0. Thus Condition P° doesn't really constrain this problem, because if two policies are different, it will always be revealed by an outcome with non-null measure.
- **Acausality** doesn't require Condition C° or Condition P°. It's for cases where the belief function is so dependent on the policy that pseudocausality fails to hold. Typically, the transparent Newcomb problem with perfect prediction is acausal, because it doesn't

satisfy either Condition C° or Condition P°.
To see why, we can focus on Condition P° as it's the weaker condition. Perfect prediction invalidates Condition P° because it means for example that the sa-measure giving all measure to the box being empty is in the infradistribution for the policy that 1-box when full and 2-box when empty, yet it isn't in the infradistribution for the policy that always 2-box. In the latter case, Omega will know that the policy will 2-box on seeing the box full, and thus will make the box full every time.

Perhaps one of the most important results philosophically of Infra-Bayesianism is that one can go from pseudocausality to causality by the Nirvana trick, and from causality to pseudocausality by removing Nirvana (Theorem 3.1°). So the first direction basically means that if thinking about policy-dependency fries your brain, you can just add Nirvana, and voilà, everything is policy-independent and causal again. And equivalently, if you have a causal setting with the Nirvana trick, you can remove the trick at the price of only ensuring pseudocausality.

This looks really useful, because in my own experience, non causal situations are really confusing. Having a formal means to convert to a more causal case (at the price of using the Nirvana trick) could thus help in clarifying some issues with decision theory and Newcomb-like problems.

(The same sort of result holds between acausal hypotheses and so-called surcausal hypotheses, but this one requires digging into so many subtle details that I will not present it here.)

Conclusion

Infra-Bayesianism provides a framework for studying learning theory for RL in the context of non-realizability. It is based around infradistribution, sets of distributions with additional data, which satisfy additional conditions for both philosophical and mathematical reasons. Among the applications of Infra-Bayesianism, it can be used to study different decision theory problems in a common framework, and ensure updates which fit with what UDT would do at the beginning of time.

I hope that this post gave you a better idea of Infra-Bayesianism, and whether or not you want to take the time to dig deeper. If you do, I also hope that what I wrote will make navigation a bit easier.

Distillation & Pedagogy 4 Logical Uncertainty 2 Decision Theory 2 Infra-Bayesianism 2 AI 4 Frontpage

Previous:

Universality Unwrapped

1 comments 29 karma

Next:

Epistemology of HCH

2 comments 17 karma

Log in to save where you left off

Mentioned in

43 My take on Vanessa Kosoy's take on AGI safety

15 Elementary Infra-Bayesianism

Moderation Log