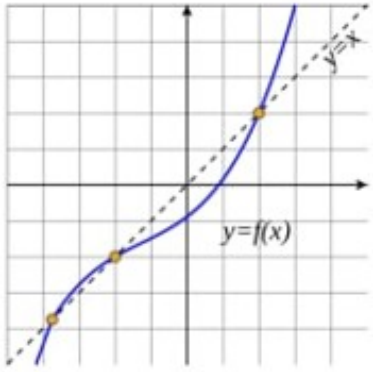

y=x
y=f(x)

Fixed Points

1. Fixed Point Exercises
2. Fixed Point Discussion
3. Topological Fixed Point Exercises
4. Diagonalization Fixed Point Exercises
5. Iteration Fixed Point Exercises
6. Hyperreal Brouwer
7. Formal Open Problem in Decision Theory
8. The Ubiquitous Converse Lawvere Problem
9. Reflective oracles as a solution to the converse Lawvere problem

Fixed Point Exercises

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Sometimes people ask me what math they should study in order to get into agent foundations. My first answer is that I have found the introductory class in every subfield to be helpful, but I have found the later classes to be much less helpful. My second answer is to learn enough math to understand all fixed point theorems. These two answers are actually very similar. Fixed point theorems span all across mathematics, and are central to (my way of) thinking about agent foundations.

This post is the start of a sequence on fixed point theorems. It will be followed by several posts of exercises that use and prove such theorems. While these exercises aren't directly connected to AI safety, I think they're quite useful for preparing to think about agent foundations research. Afterwards, I will discuss the core ideas in the theorems and where they've shown up in alignment research.

The math involved is not much deeper than a first course in the various subjects (logic, set theory, topology, computability theory, etc). If you don't know the terms, a bit of googling, wikipedia and math.stackexchange should easily get you most of the way. Note that the posts can be tackled in any order.

Here are some ways you can use these exercises:

- You can host a local MIRIx group, and go through the exercises together. This might be useful to give a local group an affordance to work on math rather than only reading papers.
- You can work on them by yourself for a while, and post questions when you get stuck. You can also post your solutions to help others, let others see an alternate way of doing a problem, or help you realize that there is a problem with your solution.
- You can skip to the discussion (which has some spoilers), learn a bunch of theorems from Wikipedia, and use this as a starting point for trying to understand some MIRI papers.
- You can use answering these questions as a goalpost for learning a bunch of introductory math from a large collection of different subfields.
- You can show off by pointing out that some of the questions are wrong, and then I will probably fix them and thank you.

The first set of exercises is here.

Thanks to Sam Eisenstat for helping develop these exercises, Ben Pace for helping edit the sequence, and many AISFP participants for testing them and noticing errors.

Meta

Read the following.

Please use the (new) spoilers feature - the symbol '>' followed by '!' followed by space - in your comments to hide all solutions, partial solutions, and other discussions of the math. The comments will be moderated strictly to cover up spoilers!

I recommend putting all the object level points in spoilers and leaving metadata outside of the spoilers, like so:

> Here's my solution / partial solution / confusion for question #5:

And put your idea in here! (reminder: LaTex is cmd-4 / ctrl-4)

Fixed Point Discussion

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Warning: This post contains some important spoilers for Topological Fixed Point Exercises, Diagonalization Fixed Point Exercises, and Iteration Fixed Point Exercises. If you plan to even try the exercises, reading this post will significantly reduce the value you can get from doing them.

Core Ideas

A fixed point of a function f is an input x such that f(x) = x. Fixed point theorems show that various types of functions must have fixed points, and sometimes give methods for finding those fixed points.

Fixed point theorems come in three flavors: Topological, Diagonal, and Iterative. (I sometimes refer to them by central examples as Brouwer, Lawvere, and Banach, respectively.)

Topological fixed points are non constructive. If f is continuous, f(0) > 0, and f(1) < 1, then we know f must have some fixed point between 0 and 1, since f(x) must somewhere transition from being greater than x to being less than x. This does not tell us where it happens. This can be especially troublesome when there are multiple fixed points, and there is no principled way to choose between them.

Diagonal fixed points are constructed with a weird trick where you feed a code for a function into that function itself. Given a function f, if you can construct a function g, which on input x, interprets x as a function, runs x on itself, and then runs f on the result (i.e. g(x) := f(x(x)).), then g(g) is a fixed point of f because g(g) = f(g(g)). This is not just an example; everything in the cluster looks like this. It is a weird trick, but it is actually very important.

Iterative fixed points can be found through iteration. For example, if $f(x) = -\frac{x}{2}$, then starting with any x value, iterating f forever will converge to the unique fixed point x = 0.

(There is a fourth cluster in number theory discussed here, but I am leaving it out, since it does not seem relevant to AI, and because I am not sure whether to put it by itself or to tack it onto the topological cluster.)

Topological Fixed Points

Examples of topological fixed point theorems include Sperner's lemma, the Brouwer fixed point theorem, the Kakutani fixed point theorem, the intermediate value theorem, and the Poincaré-Miranda theorem.

- Brouwer is the central example of the cluster. Brouwer states that any continuous function f from a compact convex set to itself has a fixed point.
- Sperner's Lemma is a discrete analogue which is used in one proof of Brouwer.
- Kakutani is a strengthening of Brouwer to degenerate set valued functions, that look almost like continuous functions.
- Poincaré-Miranda is an alternate formulation of Brouwer, which is about finding zeros rather than fixed points.
- The Intermediate Value Theorem is a special case of Poincaré-Miranda. To a first approximation, you can think of all of these theorems as one big theorem.

Topological fixed point theorems also have some very large applications. The Kakutani fixed point theorem is used in game theory to show that Nash equilibria exist, and to show that markets have equilibrium prices! Sperner's lemma is also used in some envy-free fair division results. Brouwer is also used is to show the existence of some differential equations.

In MIRI's agent foundations work, Kakutani is used to construct probabilistic truth predicates and reflective oracles, and Brouwer is used to construct logical inductors.

These applications all use topological fixed points very directly, and so carry with them most of the philosophical baggage of topological fixed points. For example, while Nash equilibria exist, they are not unique, are computationally hard to find, and feel non-constructive and arbitrary.

Diagonal Fixed Points

Diagonal fixed point theorems are all centered around the basic structure of g(g),

where g(x) := f(x(x)), as mentioned previously.

The pattern is used in many places.

- In CS theory, it is used to construct quines and the Y-combinator in lambda calculus, and to prove Rice's theorem and that the halting problem is undecidable.
- In formal logic, it is used to prove the diagonal lemma and important corollaries, like Gödel's incompleteness theorem, Löb's theorem, and Tarski's undefinability theorem. It is used to show the uncountability of the real numbers with Cantor's Diagonal Argument.
- Lawvere's fixed point theorem is the most general version of the argument, and can be used to show all of the above as a corollary.

In MIRI's agent foundations work, this shows up in the Löbian obstacle to self-trust, Löbian handshakes in Modal Combat and Bounded Open Source Prisoner's Dilemma,

as well as providing a basic foundation for why an agent reasoning about itself might make sense at all through quines.

Iterative Fixed Points

Iterative fixed point theorems are less of one cluster than the others; I will factor it into two sub-clusters, centered around the Banach fixed point theorem and Tarski fixed point theorem. (Each the same size as the original.)

The Tarski cluster is about fixed points of monotonic functions on (partially) ordered sets found by iteration. Tarski's fixed point theorem states that any order preserving function on a complete lattice has a fixed point (and further the set of fixed points forms a complete lattice). The least fixed point can be found by starting with the least element and iterating the function transfinitely. This, for example, implies that every monotonic function on from $[0, 1]$ to itself has a fixed point, even if it is not continuous. Kleene's fixed point theorem strengthens the assumptions of Tarski by adding a form of continuity (and also removes some irrelevant assumptions), which gives us that the least fixed point can be found by iterating the function only ω times. The fixed point lemma for normal functions is similar to Kleene, but with ordinals rather than partial orders. It states that any strictly increasing continuous function on ordinals has arbitrarily large fixed points.

The Banach cluster is about fixed points of contractive functions on metric spaces found by iteration. A contractive function is a function that sends points closer together. A function f is contractive if there exists an $\epsilon > 0$ such that for all $x \neq y$ $d(f(x), f(y)) \leq (1 - \epsilon)d(x, y)$. Banach's fixed point theorem state that any contractive function has a unique fixed point. This fixed point is $\lim_{n \to \infty} f^n(x)$ for any starting point x. An application of this to linear functions is that any ergodic stationary Markov chain has a stationary distribution (which is a fixed point of the transition map), which is converged to via iteration. This is also used in showing that correlated equilibria exist and can be found quickly. Banach can also be used to show that gradient descent converges exponentially quickly on a strongly convex function.

Interdisciplinary Nature

I think of Pure Mathematics as divided at the top into 5 subfields: Algebra, Analysis, Topology, Logic, and Combinatorics.

The mapping of the key fixed point theorems discussed in the exercises into these categories is surjective:

- Lawvere's fixed point theorem is Algebra
- Banach's fixed point theorem is Analysis
- Brouwer fixed point theorem is Topology
- Gödel's first incompleteness theorem is Logic
- Sperner's lemma is Combinatorics.

On top of that, major applications of fixed point theorems show up in Differential Equations, CS theory, Machine Learning, Game Theory, and Economics.

Tomorrow's AI Alignment Forum sequences post will be two short posts 'Approval-directed bootstrapping' and 'Humans consulting HCH' by Paul Christiano in the sequence Iterated Amplification.

The next posting in this sequence will be four posts of Agent Foundations research that use fixed point theorems, on Wednesday 28th November. These will be re-posts of content from the now-defunct Agent Foundations forum, all of whose content is now findable on the AI Alignment Forum (and all old links will soon be re-directed to the AI Alignment Forum).

Topological Fixed Point Exercises

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

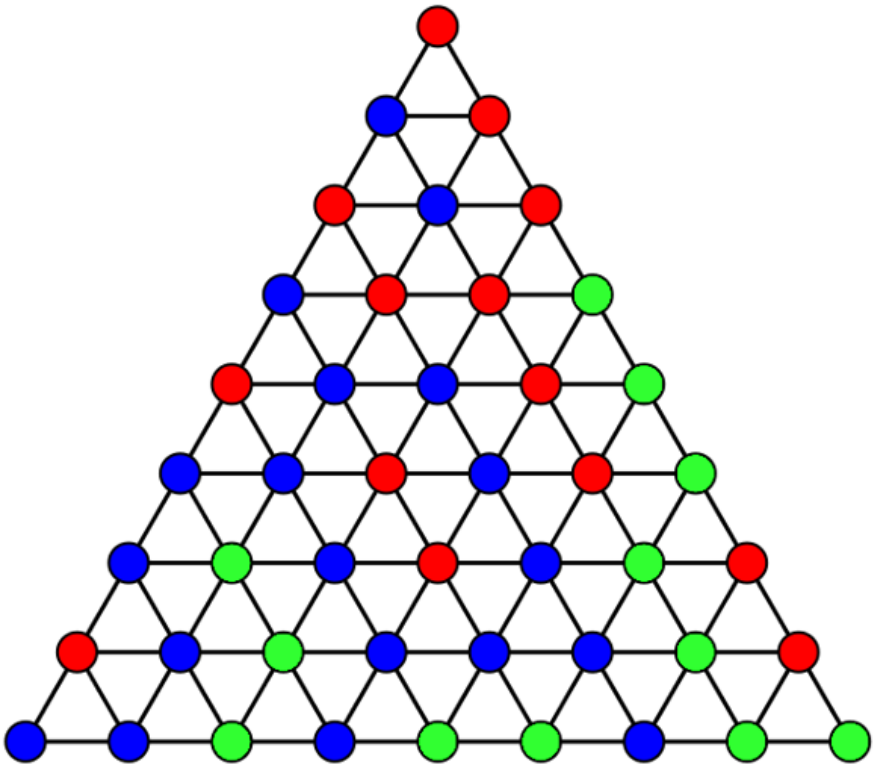
This is one of three sets of fixed point exercises. The first post in this sequence is here, giving context.

1. (1-D Sperner's lemma) Consider a path built out of n edges as shown. Color each vertex blue or green such that the leftmost vertex is blue and the rightmost vertex is green. Show that an odd number of the edges will be bichromatic.

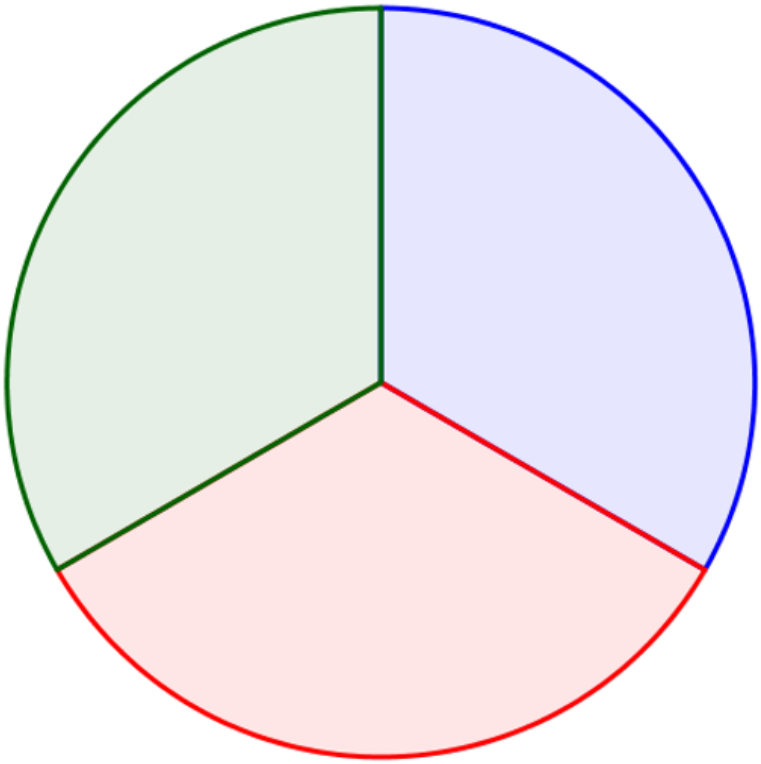
2. (Intermediate value theorem) The Bolzano-Weierstrass theorem states that any bounded sequence in R^n has a convergent subsequence. The intermediate value theorem states that if you have a continuous function $f : [0, 1] \to R$ such that $f(0) \leq 0$ and $f(1) \geq 0$, then there exists an $x \in [0, 1]$ such that $f(x) = 0$. Prove the intermediate value theorem. It may be helpful later on if your proof uses 1-D Sperner's lemma and the Bolzano-Weierstrass theorem

3. (1-D Brouwer fixed point theorem) Show that any continuous function $f : [0, 1] \to [0, 1]$ has a fixed point (i.e. a point $x \in [0, 1]$ with $f(x) = x$). Why is this not true for the open interval $(0, 1)$?

4. (2-D Sperner's lemma) Consider a triangle built out of n^2 smaller triangles as shown. Color each vertex red, blue, or green, such that none of the vertices on the large bottom edge are red, none of the vertices on the large left edge are green, and none of the vertices on the large right edge are blue. Show that an odd number of the small triangles will be trichromatic.



5. Color the all the points in the disk as shown. Let f be a continuous function from a closed triangle to the disk, such that the bottom edge is sent to non-red points, the left edge is sent to non-green points, and the right edge is sent to non-blue points. Show that f sends some point in the triangle to the center.



6. Show that any continuous function f from closed triangle to itself has a fixed point.

7. (2-D Brouwer fixed point theorem) Show that any continuous function from a compact convex subset of R^2 to itself has a fixed point. (You may use the fact that given any closed convex subset S of R^n, the function from R^n to S which projects each point to the nearest point in S is well defined and continuous.)

8. Reflect on how non-constructive all of the above fixed-point findings are. Find a parameterized class of functions where for each $t \in [0,1]$, $f_t : [0,1] \to [0,1]$, and the function $t \mapsto f_t$ is continuous, but there is no continuous way to pick out a single fixed point from each function (i.e. no continuous function g such that g(t) is a fixed point of f_t for all t).

9. (Sperner's lemma) Generalize exercises 1 and 4 to an arbitrary dimension simplex.

10. (Brouwer fixed point theorem) Show that any continuous function from a compact convex subset of R^n to itself has a fixed point.

11. Given two nonempty compact subsets $A, B \subseteq R^n$, the Hausdorff distance between them is the supremum

$$\max \{ \sup_{a \in A} d(a, B), \sup_{b \in B} d(b, A) \}$$

over all points in either subset of the distance from that point to the other subset. We call a set valued function $f : S \to 2^T$ a continuous Hausdorff limit if there is a sequence $\{f_n\}$ of continuous functions from S to T whose graphs, $\{(x, y) \mid y = f_n(x)\} \subseteq S \times T$, converge to the graph of f, $\{(x, y) \mid f(x) \ni y\} \subseteq S \times T$, in Hausdorff distance. Show that every continuous Hausdorff limit $f : T \to 2^T$ from a compact convex subset of R^n to itself has a fixed point (a point x such that $x \in f(x)$).

12. Let S and T be nonempty compact convex subsets of R^n. We say that a set valued function, $f : S \to 2^T$ is a Kakutani function if the graph of f, $\{(x, y) \mid f(x) \ni y\} \subseteq S \times T$, is closed, and f(x) is convex and nonempty for all $x \in S$. For example, we could take S and T to be the interval $[0, 1]$, and we could have $f : S \to 2^T$ send each $x < \frac{1}{2}$ to $\{0\}$, map $x = \frac{1}{2}$ to the whole interval $[0, 1]$, and map $x > \frac{1}{2}$ to $\{1\}$. Show that every Kakutani function is a continuous Hausdorff limit. (Hint: Start with the case where S is a unit cube. Construct f_n by breaking S into small cubes of side length 2^{-n}. Constuct a smaller cube of side length 2^{-n-1} within each 2^{-n} cube. Send each small 2^{-n-1} to the convex hull of the images of all points in the 2^{-n} cube with a continuous function, and glue these together with straight lines. Make sure you don't accidentally get extra limit points.)

13. (Kakutani fixed point theorem) Show that every Kakutani function from a compact convex subset of $S \subseteq R^n$ to itself has a fixed point.

Please use the spoilers feature - the symbol '>' followed by '!' followed by space -in your comments to hide all solutions, partial solutions, and other discussions of the math. The comments will be moderated strictly to hide spoilers!

I recommend putting all the object level points in spoilers and leaving metadata outside of the spoilers, like so: "I think I've solved problem #5, here's my solution <spoilers>" or "I'd like help with problem #3, here's what I understand <spoilers>" so that people can choose what to read.

Diagonalization Fixed Point Exercises

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This is the second of three sets of fixed point exercises. The first post in this sequence is here, giving context.

1. Recall Cantor's diagonal argument for the uncountability of the real numbers. Apply the same technique to convince yourself than for any set S, the cardinality of S is less than the cardinality of the power set P(S) (i.e. there is no surjection from S to P(S)).

2. Suppose that a nonempty set T has a function f from T to T which lacks fixed points (i.e. $f(x) \neq x$ for all $x \in T$). Convince yourself that there is no surjection from S to $S \to T$, for any nonempty S. (We will write the set of functions from A to B either as $A \to B$ or B^A; these are the same.)

3. For nonempty S and T, suppose you are given $g : S \to T^S$ a surjective function from the set S to the set of functions from S to T, and let f be a function from T to itself. The previous result implies that there exists an x in T such that $f(x) = x$. Can you use your proof to describe x in terms of f and g?

4. Given sets A and B, let $\text{Comp}(A, B)$ denote the space of total computable functions from A to B. We say that a function from C to $\text{Comp}(A, B)$ is computable if and only if the corresponding function $f' : C \times A \to B$ (given by $f'(c, a) = f(c)(a)$) is computable. Show that there is no surjective computable function from the set S of all strings to $\text{Comp}(S, \{T, F\})$.

5. Show that the previous result implies that there is no computable function $\text{halt}(x, y)$ from $S \times S \to \{T, F\}$ which outputs T if and only if the first input is a code for a Turing machine that halts when given the second input.

6. Given topological spaces A and B, let Cont(A, B) be the space with the set of continuous functions from A to B as its underlying set, and with topology such that $f : C \to \mathrm{Cont}(A, B)$ is continuous if and only if the corresponding function $f' : C \times A \to B$ (given by $f'(c, a) = f(c)(a)$) is continuous, assuming such a space exists. Convince yourself that there is no space X which continuously surjects onto Cont(X, S), where S is the circle.

7. In your preferred programming language, write a quine, that is, a program whose output is a string equal to its own source code.

8. Write a program that defines a function f taking a string as input, and produces its output by applying f to its source code. For example, if f reverses the given string, then the program should outputs its source code backwards.

9. Given two sets A and B of sentences, let Syn(A, B) be the set of all functions from A to B defined by substituting the Gödel number of a sentence in A into a fixed formula. Let S_0 be the set of all sentences in the language of arithmetic with one unbounded universal quantifier and arbitrarily many bounded quantifiers, and let S_1 be the set of all formulas with one free variables of that same quantifier complexity. By representing syntax using arithmetic, it is possible to give a function $f \in \mathrm{Syn}(S_1 \times S_1, S_0)$ that substitutes its second argument into its first argument. Pick some coding of formulas as natural numbers, where we denote the number coding for a formula φ as $\ulcorner\varphi\urcorner$. Using this, show that for any formula $\phi \in S_1$, there is a formula $\psi \in S_0$ such that $\phi(\ulcorner\psi\urcorner) \leftrightarrow \psi$.

10. (Gödel's second incompleteness theorem) In the set S_1, there is a formula $\neg\mathrm{Bew}$ such that $\neg\mathrm{Bew}(\ulcorner\psi\urcorner)$ holds iff the sentence ψ is not provable in Peano arithmetic. Using this, show that Peano arithmetic cannot prove its own consistency.

11. (Löb's theorem) More generally, the diagonal lemma states that for any formula ϕ with a single free variable, there is a formula ψ such that, provably, $\phi(\ulcorner\psi\urcorner) \leftrightarrow \psi$. Now, suppose that Peano arithmetic proves that $\mathrm{Bew}(\psi) \to \psi$ for some formula ψ. Show that Peano arithmetic also proves ψ itself. Some facts that you may need

are that (a) when a sentence ψ is provable, the sentence $\mathrm{Bew}(\psi)$ is itself provable, (b) Peano arithmetic proves this fact, that is, Peano arithmetic proves $\mathrm{Bew}(\psi) \to \mathrm{Bew}(\mathrm{Bew}(\psi))$, for any sentence ψ and (c) Peano arithmetic proves the fact that if χ and $\chi \to \zeta$ are provable, then ζ is provable.

12. (Tarski's theorem) Show that there does not exist a formula ϕ with one free variable such that for each sentence ψ, the statement $\phi(\ulcorner\psi\urcorner) \leftrightarrow \psi$ holds.

13. Looking back at all these exercises, think about the relationship between them.

Please use the spoilers feature - the symbol '>' followed by '!' followed by space -in your comments to hide all solutions, partial solutions, and other discussions of the math. The comments will be moderated strictly to hide spoilers!

I recommend putting all the object level points in spoilers and including metadata outside of the spoilers, like so: "I think I've solved problem #5, here's my solution <spoilers>" or "I'd like help with problem #3, here's what I understand <spoilers>" so that people can choose what to read.

Iteration Fixed Point Exercises

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This is the third of three sets of fixed point exercises. The first post in this sequence is here, giving context.

Note: Questions 1-5 form a coherent sequence and questions 6-10 form a separate coherent sequence. You can jump between the sequences.

1. Let (X, d) be a complete metric space. A function $f : X \to X$ is called a contraction if there exists a $q < 1$ such that for all $x, y \in X$, $d(f(x), f(y)) \leq q \cdot d(x, y)$. Show that if f is a contraction, then for any x, the sequence $\{x_n = f^n(x_0)\}$ converges. Show further that it converges exponentially quickly (i.e. the distance between the nth term and the limit point is bounded above by $c \cdot a^n$ for some $a < 1$)

2. (Banach contraction mapping theorem) Show that if (X, d) is a complete metric space and f is a contraction, then f has a unique fixed point.

3. If we only require that $d(f(x), f(y)) < d(x, y)$ for all $x \neq y$, then we say f is a weak contraction. Find a complete metric space (X, d) and a weak contraction $f : X \to X$ with no fixed points.

4. A function $f : R^n \to R$ is convex if $f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y)$, for all $t \in [0, 1]$ and $x, y \in R^n$. A function f is strongly convex if you can subtract a positive parabaloid from it and it is still convex. (i.e. f is strongly convex if $x \mapsto f(x) - \varepsilon||x||^2$ is convex for some $\varepsilon > 0$.) Let f be a strongly convex smooth function from R^n to R, and suppose that the magnitude of the second derivative $\|\nabla^2 f\|$ is bounded. Show that there exists an $\varepsilon > 0$ such that the function $g : R^n \to R^n$ given by $x \mapsto x - \varepsilon(\nabla f)(x)$ is a contraction. Conclude that gradient descent with a sufficiently small constant step size converges exponentially quickly on a strongly convex smooth function.

5. A finite stationary Markov chain is a finite set S of states, along with probabilistic rule $A : S \to \Delta S$ for transitioning between the states, where ΔS represents the space of probability distributions on S. Note that the transition rule has no memory, and depends only on the previous state. If for any pair of states $s, t \in \Delta S$, the probability of passing from s to t in one step is positive, then the Markov chain (S, A) is ergodic. Given an ergodic finite stationary Markov chain, use the Banach contraction mapping theorem to show that there is a unique distribution over states which is fixed under application of transition rule. Show that, starting from any state s, the limit distribution $\lim_{n\to\infty} A^n(s)$ exists and is equal to the stationary distribution.

6. A function f from a partially ordered set to another partially ordered set is called monotonic if $x \leq y$ implies that $f(x) \leq f(y)$. Given a partially ordered set $(P, \leq)$ with finitely many elements, and a monotonic function from P to itself, show that if $f(x) \geq x$ or $f(x) \leq x$, then $f^n(x)$ is a fixed point of f for all $n > |P|$.

7. A complete lattice $(L, \leq)$ is a partially ordered set in which each subset of elements has a least upper bound and greatest lower bound. Under the same hypotheses as the previous exercise, extend the notion of $f^n(x)$ for natural numbers n to $f^\alpha(x)$ for ordinals α, and show that $f^\alpha(x)$ is a fixed point of f for all $x \in X$ with $f(x) \leq x$ or $f(x) \geq x$ and all $|\alpha| > |L|$ ($|A| \leq |B|$ means there is an injection from A to B, and $|A| > |B|$ means there is no such injection).

8. (Knaster-Tarski fixed point theorem) Show that the set of fixed points of a monotonic function on a complete lattice themselves form a complete lattice. (Note that since the empty set is always a subset, a complete lattice must be nonempty.)

9. Show that for any set A, $(P(A), \subseteq)$ forms a complete lattice, and that any injective function from A to B defines a monotonic function from $(P(A), \subseteq)$ to $(P(B), \subseteq)$. Given injections $f : A \to B$ and $g : B \to A$, construct a subset A' of A and a subset of B' of B such that $B' = f(A')$ and $A - A' = g(B - B')$.

10. (Cantor–Schröder–Bernstein theorem) Given sets A and B, show that if $|A| \leq |B|$ and $|A| \geq |B|$, then $|A| = |B|$. ($|A| \leq |B|$ means there is an injection from A to B, and $|A| = |B|$ means there is a bijection)

Please use the spoilers feature - the symbol '>' followed by '!' followed by space -in your comments to hide all solutions, partial solutions, and other discussions of the math. The comments will be moderated strictly to hide spoilers!

I recommend putting all the object level points in spoilers and including metadata outside of the spoilers, like so: "I think I've solved problem #5, here's my solution <spoilers>" or "I'd like help with problem #3, here's what I understand <spoilers>" so that people can choose what to read.

Tomorrow's AI Alignment Forum Sequences post will be "Approval-directed agents: overview" by Paul Christiano in the sequence Iterated Amplification.

The next post in this sequence will be released on Saturday 24th November, and will be 'Fixed Point Discussion'.

Hyperreal Brouwer

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

(This post was originally published on Oct 6th 2017, and has been temporarily brought forwarded as part of the AI Alignment Forum launch sequence on fixed points.)

This post explains how to view Kakutani's fixed point theorem as a special case of Brouwer's fixed point theorem with hyperreal numbers. This post is just math intuitions, but I found them useful in thinking about Kakutani's fixed point theorem and many things in agent foundations. This came out of conversations with Sam Eisenstat.

Brouwer's fixed theorem says that a continuous function from a compact convex subset of R^n to itself has fixed point. Kakutani's fixed point is similar, but instead of continuous functions, it uses Kakutani functions, which are set valued functions with closed graph which are point wise nonempty and convex.

When I think about Kakutani functions, I usually think about them as limits of continuous functions. For example, consider the kakutani function f from $[-1, 1]$ to itself which sends negative inputs to 1, positive inputs to -1, and sends 0 to the entire interval. You can view f as a the limit of a sequence of functions f_n sends x to $\min(\max(-n \cdot x, -1), 1)$. This is not a point wise limit, since if it was 0 would be sent to 0, rather than the entire interval. Instead, it is a limit in the Hausdorff metric between the graphs of the functions.

Given two compact nonempty subsets X and Y of R^n, the Hausdorff distance between X and Y is the maximum over all points in X or Y of the Euclidean distance between that point and the closest point in the other set. Since X and Y are compact, this maximum is achieved.

Given compact convex subsets X and Y of R^n, we say that a sequence of continuous functions f_n from X to Y converges in graph Hausdorff distance to the closed graph set valued function f if the graphs of f_n viewed as subsets of $X \times Y$ converges to the graph of f in Hausdorff distance.

We say that a closed graph function f from X to Y is a continuous graph limit of if there exists a sequence of continuous functions which converges to f in graph Hausdorff distance.

Theorem 1: Every continuous graph limit f **from a compact convex subset of** R^n **to itself has a fixed point (a point contained in its image).**

Proof: f is a limit of continuous functions f_n each of which has a fixed point by Brouwer. Choose one fixed point from each f_n to get a sequence of points which has a convergent subsequence by Bolzano–Weierstrass. Let x be the limit of this convergent subsequence. If f(x) did not contain x, then (x, x) would not be in the graph of f. Since f is closed graph, a ball around (x, x) would not be in the graph of f, which contradicts the fact that $\{f_n\}$ must contain functions with graphs arbitrarily close to the graph of f with fixed points arbitrarily close to x and thus points in their graphs arbitrarily close to (x, x). □

This theorem is not equivalent to Kakutani's fixed point theorem. There exist continuous graph limits which are not point wise convex (but only in more than one dimension). For example the function from $[-1, 1]^2$ to $[-1, 1]^2$ which sends every point to the circle of points distance 1 from 0 is not Kakutani, but is a continuous graph limit. It is the limit of functions f_n given by $(x, y) \mapsto (\cos(n \cdot x), \sin(n \cdot x))$.

However, this theorem is strictly stronger that Kakutani's fixed point theorem (although sometimes harder to use, since showing a function is Kakutani might be easier than showing it is a continuous graph limit)

Theorem 2: Given compact convex subsets X **and** Y **of** R^n**, every Kakutani function** f **from** X **to** Y **is a continuous graph limit.**

Proof: We define a function f_n as follows. Take a finite set S of radius 1/n open balls in $X \times Y$ such that each ball intersects the graph of f, the X coordinates of the centers of all the balls are distinct, and the balls cover the entire graph of f. This induces a covering S_X of X by radius 1/n open balls by taking balls centered at the X coordinates of the centers of S. We continuously map each point in X to a weighted average of balls in S_X as follows. If a point is the center of some ball, it is sent to 100% that ball. Otherwise, it is sent to a combination of all the balls in which it is contained with weight proportional to (the reciprocal of the distance to the center of that ball) minus

n. This gives a function f_n from X to Y by mapping each point to the weighted average of the Y coordinates of the centers of the balls in S with weights equal to the weights of the corresponding balls in S_X above. One can verify that f_n is continuous.

Observe that the graph of f_n contains the centers of all balls in S. Thus, f_n contains points within 1/n of every point is the graph of f. Thus, if f_n did not converge to f, it must be because infinitely many f_n contain points a distance from the graph of f bounded away from 0. Consider a convergent subsequence of these points. This gives a point (x, y) not in the graph of f and a subsequence of the f_n with points in their graphs converging to (x, y).

Let d be half the distance between y and f(x), and consider the set T of all points in Y at most d from the nearest point in f(x). Note that T is convex. Note that all (x', y') in the graph of f with x' sufficiently close to x must have y' in T, since otherwise there would be (x', y') with x' converging to x which must have a convergent subsequence with y' converging to a point not in f(x), contradicting the fact that f has closed graph.

However f_n must send all points within distance ε of x to a point in the convex hull of the images under f of points within $\varepsilon + 1/n$ of x. But, we showed that for ε sufficiently small and 1/n sufficiently large all of these points must be in T. Therefore, for all sufficiently large n, f_n must send all points within ε of x to points in T, which are bounded away from y, contradicting the assumption that points in the f_n converge to (x, y). □

Corollary: Kakutani's fixed point theorem

We have proven (a strengthening of) Kakutani's Fixed Point Theorem from Brouwer's fixed point theorem, and given a way to think about Kakutani functions as just limits of graphs of continuous functions, and thus have better intuitions about what (a superset of) Kakutani functions look like. We will now take this further and think about Kakutani as a consequence of an analogue of Brouwer using Hyperreal infinitesimal numbers. (I am going to be informal now. I am not going to use standard notation here. I am not going to make sense to people who don't already know something about non-standard analysis. Sorry.)

Given a compact convex subset X of R^n, we can define $*X$ to be the set of all equivalence classes of infinite sequences of elements of X, where two sequences are equivalent if they agree on a set that matters according to some ultrafilter U on N. A function $*f$ from $*X$ to itself is defined by a sequence of functions from X to itself $\{f_n\}$, where you apply the functions pointwise. I will call a function hyper-continuous if each of the component functions is continuous. (I am not sure what this is actually called.) Each point in $*X$ has a standard part, which is a point in X, which is the unique point such that a subset of components that matter converges to X.

Claim: Every hyper continuous function $*f$ from $*X$ to itself has a fixed point.

Proof: Just take the sequence of the fixed points of the individual component functions. □

Claim: Continuous graph limits f from X to Y are exactly those closed graph set-valued functions such that there exists a hyper-continuous function $*f : *X \to *Y$ such that $y \in f(x)$ if and only if there exist points $*x \in *X$ and $*y \in *Y$ with standard parts x and y respectively such that $f(*x) = *y$

Proof: Use the same sequence of functions with graphs converging to that of f as the sequence of functions defining $*f$.□

Thus, we can view continuous graph limits (and thus Kakutani functions) as something you get when looking at just the standard part of a hyper-continuous function from the hyper version of X to itself. The fixed point will fix everything, including the infinitesimal parts, and we do not have to deal with any set-valued functions.

For example, consider our original function f from $[-1, 1]$ to itself which sends negative inputs to 1, positive inputs to -1, and sends 0 to the entire interval. We can view this as a function $*f$ involving infinitesimals where everything with positive real part is sent to something with real part -1, everything with negative real part is sent to something with real part 1, and the infinitesimal numbers very close to 0 are sent to something in-between. If we use the sequence of functions from above and let the infinitesimal ε be $\{1/n\}$, then zooming in on the inputs between $-\varepsilon$ and ε, $*f$ will just be a steep linear function with slope $-1/\varepsilon$.

Now to be even more vague and connect things back up with agent foundations, perhaps this can give some good intuitions about what is happening with reflective oracles and probabilistic truth predicates. The oracle/truth predicate is effectively "zooming in" on the area around a specific probability, and when you stack oracle calls or truth predicates within each other, you can zoom in further. The fact that the probabilistic truth predicate does not know that it is reflectively consistent, can be viewed as it not believing a sentence akin to "If I assign probability less that ε to ϕ, then I also assign probability less that ε^2 to ϕ," which seems very reasonable. It also makes reflective oracles and the probabilistic truth predicates look more similar to other approaches to the same problem that are more hierarchy forming solutions to the same problem like normal halting oracles. Here the hierarchy comes from zooming in further and further on the infinitesimal in the Kakutani function.

This post was originally published on Oct 6th 2017, and has been brought forwarded as part of the AI Alignment Forum launch sequences.

Tomorrow's AIAF sequences post will be 'Iterated Amplification and Distillation' by Ajeya Cotra, in the sequence on iterated amplification.

Formal Open Problem in Decision Theory

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

(This post was originally published on March 31st 2017, and has been brought forwarded as part of the AI Alignment Forum launch sequence on fixed points.)

In this post, I present a new formal open problem. A positive answer would be valuable for decision theory research. A negative answer would be helpful, mostly for figuring out what is the closest we can get to a positive answer. I also give some motivation for the problem, and some partial progress.

Open Problem: Does there exist a topological space X (in some convenient category of topological spaces) such that there exists a continuous surjection from X to the space $[0,1]^X$ (of continuous functions from X to $[0,1]$)?

Motivation:

Topological Naturalized Agents: Consider an agent who makes some observations and then takes an action. For simplicity, we assume there are only two possible actions, A and B. We also assume that the agent can randomize, so we can think of this agent as outputting a real number in $[0,1]$, representing its probability of taking action A.

Thus, we can think of an agent as having a policy which is a function from the space Y of possible observations to $[0,1]$. We will require that our agent behaves continuously as a function of its observations, so we will think of the space of all possible policies as the space of continuous functions from Y to $[0,1]$, denoted $[0,1]^Y$.

We will let X denote the space of all possible agents, and we will have a function $f : X \to [0,1]^Y$ which takes in an agent, and outputs that agent's policy.

Now, consider what happens when there are other agents in the environment. For simplicity, we will assume that our agent observes one other agent, and makes no other observations. Thus, we want to consider the case where $Y = X$, so $f : X \to [0,1]^X$.

We want f to be continuous, because we want a small change in an agent to correspond to a small change in the agent's policy. This is particularly important since other agents will be implementing continuous functions on agents, and we would like any continuous function on policies to be able to be considered valid continuous function on agents.

We also want f to be surjective. This means that our space of agents is sufficiently rich that for any possible continuous policy, there is an agent in our space that implements that policy.

In order to meet all these criteria simultaneously, we need a space X of agents, and a continuous surjection $f : X \to [0,1]^X$.

Unifying Fixed Point Theorems: While we are primarily interested in the above motivation, there is another secondary motivation, which may be more compelling for those less interested in agent foundations.

There are (at least) two main clusters of fixed point theorems that have come up many times in decision theory, and mathematics in general.

First, there is the Lawvere cluster of theorems. This includes the Lawvere fixed point theorem, the diagonal lemma, and the existence of Quines and fixed point combinators. These are used to prove Gödel's incompleteness Theorem, Cantor's Theorem, Löb's Theorem, and achieve robust cooperation in the Prisoner's Dilemma in modal framework and bounded variants. All of these can be seen as corollaries of Lawvere's fixed point theorem, which states that in a cartesian closed category, if there is a point-surjective map $f : X \to Y^X$, then every morphism $g : Y \to Y$ has a fixed point.

Second, there is the Brouwer cluster of theorems. This includes Brouwer's fixed point theorem, The Kakutani fixed point theorem, Poincaré–Miranda, and the intermediate value theorem. These are used to prove the existence of Nash Equilibria, Logical Inductors, and Reflective Oracles.

If we had a topological space and a continuous surjection $X \to [0,1]^X$, this would allow us to prove the one-dimensional Brouwer fixed point theorem directly using the Lawvere fixed point theorem, and thus unify these two important clusters.

Thanks to Qiaochu Yuan for pointing out the connection to Lawvere's fixed point theorem (and actually asking this question three years ago).

Partial Progress:

Most Diagonalization Intuitions Do Not Apply: A common initial reaction to this question is to conjecture that such an X does not exist, due to cardinality or diagonalization intuitions. However, note that all of the diagonalization theorems pass

through (some modification of) the same lemma: Lawvere's fixed point theorem. However, this lemma does not apply here!

For example, in the category of sets, the reason that there is no surjection from any set X to the power set, $\{T, F\}^X$, is because if there were such a surjection, Lawvere's fixed point theorem would imply that every function from $\{T, F\}$ to itself has a fixed point (which is clearly not the case, since there is a function that swaps T and F).

However, we already know by Brouwer's fixed point theorem that every continuous function from the interval $[0, 1]$ to itself has a fixed point, so the standard diagonalization intuitions do not work here.

Impossible if You Replace $[0, 1]$ with e.g. S^1: This also provides a quick sanity check on attempts to construct an X. Any construction that would not be meaningfully different if the interval $[0, 1]$ is replaced with the circle S^1 is doomed from the start. This is because a continuous surjection $X \to (S^1)^X$ would violate Lawvere's fixed point theorem, since there is a continuous map from S^1 to itself without fixed points.

Impossible if you Require a Homeomorphism: When I first asked this question I asked for a homeomorphism between X and $[0, 1]^X$. Sam Eisenstat has given a very clever argument why this is impossible. You can read it here. In short, using a homeomorphism, you would be able to use Lawvere to construct a continuous map that send a function from $[0, 1]$ to itself to a fixed point of that function. However, no such continuous map exists.

Notes:

If you prefer not to think about the topology of $[0, 1]^X$, you can instead find a space X, and a continuous map $h : X \times X \to [0, 1]$, such that for every continuous function $f : X \to [0, 1]$, there exists an $x_f \in X$, such that for all $x \in X$, $h(x_f, x) = f(x)$.

Many of the details in the motivation could be different. I would like to see progress on similar questions. For example, you could add some computability condition to the space of functions. However, I am also very curious which way this specific question will go.

This post came out of many conversations, with many people, including: Sam, Qiaochu, Tsvi, Jessica, Patrick, Nate, Ryan, Marcello, Alex Mennen, Jack Gallagher, and James Cook.

This post was originally published on March 31st 2017, and has been brought forwarded as part of the AI Alignment Forum launch sequences.

Tomorrow's AIAF sequences post will be 'Iterated Amplification and Distillation' by Ajeya Cotra, in the sequence on iterated amplification.

The Ubiquitous Converse Lawvere Problem

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

(This post was originally published on Oct 20th 2017, and is 1 of 4 posts brought forwarded today as part of the AI Alignment Forum launch sequence on fixed points.)

In this post, I give a stronger version of the open question presented here, and give a motivation for this stronger property. This came out of conversations with Marcello, Sam, and Tsvi.

Definition: A continuous function $f : X \to Y$ is called ubiquitous if for every continuous function $g : X \to Y$, there exists a point $x \in X$ such that $f(x) = g(x)$.

Open Problem: Does there exist a topological space X with a ubiquitous function $f : X \to [0,1]^X$?

I will refer to the original problem as the Converse Lawvere Problem, and the new version as the the Ubiquitous Converse Lawvere Problem. I will refer to a space satisfying the conditions of (Ubiquitous) Converse Lawvere Problem, a (Ubiquitous) Converse Lawvere Space, abbreviated (U)CLS. Note that a UCLS is also a CLS, since a ubiquitous is always surjective, since g can be any constant function.

Motivation: True FairBot

Let X be a Converse Lawvere Space. Note that since such an X might not exist, the following claims might be vacuous. Let $f : X \to [0,1]^X$ be a continuous surjection.

We will view X as a space of possible agents in an open source prisoner's dilemma game. Given two agents $A, B \in X$, we will interpret $f_A(B)$ as the probability with which A cooperates when playing against B. We will define $U_A(B) := 2f_B(A) - f_A(B)$, and interpret this as the utility of agent A when playing in the prisoner's dilemma with B.

Since f is surjective, every continuous policy is implemented by some agent. In particular, this means gives:

Claim: For any agent $A \in X$, there exists another agent $A' \in X$ such that $f_{A'}(B) = f_B(A)$. i.e. A' responds to B the way that B responds to A.

Proof: The function $B \mapsto f_B(A)$ is a continuous function, since $B \mapsto f_B$ is continuous, and evaluation is continuous. Thus, there is a policy $B \mapsto f_B(A)$ in $[0,1]^X$. Since f is surjective, this policy must be the image under f of some agent A', so $f_{A'}(B) = f_B(A)$.

Thus, for any fixed agent A, we have some other agent A' that responds to any B the way B responds to A. However, it would be nice if $A' = A$, to create a FairBot that responds to any opponent the way that that opponent responds to it. Unfortunately, to construct such a FairBot, we need the extra assumption that f is ubiquitous.

Claim: If f is ubiquitous, then exists a true fair bot in X: an agent $FB \in X$, such that $f_{FB}(A) = f_A(FB)$ for all agents $A \in X$.

Proof: Given an agent $B \in X$, there exists an policy $g_B \in [0,1]^X$ such that $g_B(A) = f_A(B)$ for all A, since $A \mapsto f_A(B)$ is continuous. Further, the function $B \mapsto g_B$ is continuous, since the function $A, B \mapsto f_A(B)$ and the definition of the exponential topology. Since f is ubiquitous, there must be some $FB \in X$ such that $f_{FB} = g_{FB}$. But then, for all A, we have $f_{FB}(A) = g_{FB}(A) = f_A(FB)$.

Note that we may not need the full power of ubiquitous here, but it is the simplest property I see that gets the result.

Note that this FairBot is fair in a stronger sense than the FairBot of modal combat, in that it always has the same output as its opponent. This may make you suspicious, since the you can also construct an UnfairBot, UB such that $f_{UB}(A) = 1 - f_A(UB)$ for all A. This would have caused a problem in the modal combat framework, since you can put a FairBot and an UnfairBot together to form a paradox. However, we do not have this problem, since we deal with probabilities, and simply have $f_{UB}(FB) = f_{FB}(UB) = 1/2$. Note that the exact phenomenon that allows this to possibly work is the fixed point property of the interval $[0,1]$ which is the only reason that we cannot use diagonalization to show that no CLS exists.

Finally, note that we already have a combat framework that has a true FairBot: the reflective oracle framework. In fact, the reflective oracle framework may have all the benefits we would hope to get out of a UCLS. (other than the benefit of simplicity of not having to deal with computability and hemicontinuity).

This post was originally published on Oct 20th 2017, and has been brought forwarded as part of the AI Alignment Forum launch sequences.

Tomorrow's AIAF sequences post will be 'Iterated Amplification and Distillation' by Ajeya Cotra, in the sequence on iterated amplification.

Reflective oracles as a solution to the converse Lawvere problem

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

(This post was originally published on Nov 30th 2017, and is 1 of 4 posts brought forwarded today as part of the AI Alignment Forum launch sequence on fixed points.)

1 Introduction

Before the work of Turing, one could justifiably be skeptical of the idea of a universal computable function. After all, there is no computable function $f: \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ such that for all computable $g: \mathbb{N} \to \mathbb{N}$ there is some index i_g such that $f(i_g, n) = g(n)$ for all n. If there were, we could pick $g(n) = f(n, n) + 1$, and then

$$g(i_g) = f(i_g, i_g) + 1 = g(i_g) + 1,$$

a contradiction. Of course, universal Turing machines don't run into this obstacle; as Gödel put it, "By a kind of miracle it is not necessary to distinguish orders, and the diagonal procedure does not lead outside the defined notion." [1]

The miracle of Turing machines is that there is a partial computable function $f: \mathbb{N} \times \mathbb{N} \to \mathbb{N} \cup \{\bot\}$ such that for all partial computable $g: \mathbb{N} \to \mathbb{N} \cup \{\bot\}$ there is an index i such that $f(i, n) = g(n)$ for all n. Here, we look at a different "miracle", that of reflective oracles [2,3]. As we will see in Theorem 1, given a reflective oracle O, there is a (stochastic) O-computable function $f: \mathbb{N} \times \mathbb{N} \to \{0, 1\}$ such that for any (stochastic) O-computable function $g: \mathbb{N} \to \{0, 1\}$, there is some index i such that $f(i, n)$ and $g(n)$ have the same distribution for all n. This existence theorem seems to skirt even closer to the contradiction mentioned above.

We use this idea to answer "in spirit" the converse Lawvere problem posed in [4]. These methods also generalize to prove a similar analogue of the ubiquitous converse Lawvere problem from [5]. The original questions, stated in terms of topology, remain open, but I find that the model proposed here, using computability, is equally satisfying from the point of view of studying reflective agents. Those references can be consulted for more motivation on these problems from the perspective of reflective agency.

Section 3 proves the main lemma, and proves the converse Lawvere theorem for reflective oracles. In section 4, we use that to give a (circular) proof of Brouwer's fixed

point theorem, as mentioned in [4]. In section 5, we prove the ubiquitous converse Lawvere theorem for reflective oracles.

2 Definitions

For any measurable space X, the set of probability measures on X is denoted ΔX.

A *probabilistic oracle* is a map $N \to [0,1]$. A *probabilistic oracle machine* is a probabilistic Turing machine that can query a probabilistic oracle O. On a given query $n \in N$, it gets the answer 1 with probability $O(n)$ and the answer 0 otherwise. Different queries are independent events, so this completely determines the behavior of such a machine.

We denote by ϕ_i^O the stochastic partial O-computable function $N \to \Delta(N \cup \{\bot\})$, where $\bot$ represents nonhalting, computed by the probabilistic Turing machine with index i. The notation $\phi_i^O(n)\downarrow$ indicates the event that ϕ_i^O halts on input n, and $\phi_i^O(n)\downarrow = m$ is the event that $\phi_i^O(n)$ halts and outputs m. Finally, $\phi_i^O(n)\uparrow$ is the event that ϕ_i^O does not halt on input n.

We use $\langle\cdot\rangle$ to represent a computable coding function, in order to biject arbitrary countable sets to the naturals for the purpose of computability.

A *reflective oracle* is a probabilistic oracle O such that for all $i, n \in N$ and $p \in [0,1]_Q$,

$$P(\phi_i^O(n)\downarrow = 1) > p \implies O(\langle i,n,p\rangle) = 1$$

$$P(\neg\phi_i^O(n)\downarrow = 0) < p \implies O(\langle i,n,p\rangle) = 0.$$

For more on reflective oracles, see Fallenstein et al., 2015 [2].

A function $f: N \to [0,1]$ is O-*computable* if there is an index i such that for all $n \in N$, we have

$$P(\phi_i^O(n)\downarrow \in \{0,1\}) = 1$$

and

$$P(\phi_i^O(n)\downarrow = 1) = f(n).$$

That is, ϕ_i^O represents f by probabilistically outputting either 0 or 1.

For any $m \in \mathbb{N}$, a function $f\colon \mathbb{N} \to [0,1]^m$ is O-*computable* if each coordinate $f_j\colon \mathbb{N} \to [0,1]$ for $1 \leq j \leq m$ is O-computable.

A function $f\colon \mathbb{N} \to [0,1]^{\mathbb{N}}$ is O-*computable* if the corresponding function $\mathbb{N} \to [0,1]$ given by $\langle n, m\rangle \mapsto f(n)(m)$ is O-computable.

For any point $p \in [0,1]$, a probabilistic oracle O is *compatible* with p if for all $r \in [0,1]_{\mathbb{Q}}$, we have $O(\langle r\rangle) = 0$ if $p < r$ and $O(\langle r\rangle) = 1$ if $p > r$. More generally, for any $m \in \mathbb{N}$ and any $p \in [0,1]^m$, a probabilistic oracle O is *compatible* with p if for all j such that $1 \leq j \leq m$ and all $r \in [0,1]_{\mathbb{Q}}$, we have $O(\langle j, r\rangle) = 0$ if $p_j < r$ and $O(\langle j, r\rangle) = 1$ if $p_j > r$.

A function $f\colon [0,1]^m \to [0,1]^m$ is O-*computable* if for each coordinate $f_j\colon [0,1]^m \to [0,1]$, there is an index i such that whenever P is compatible with p, we have

$$P(\phi_i^{O,P}(0)\downarrow \in \{0,1\}) = 1$$

and

$$P(\phi_i^{O,P}(0)\downarrow = 1) = f_j(p).$$

A map $f\colon \mathbb{N} \to [0,1]^{\mathbb{N}}$ is O-*computably ubiquitous* if for every O-computable map $e\colon \mathbb{N} \to [0,1]^{\mathbb{N}}$, there is some $n \in \mathbb{N}$ such that $f(n) = e(n)$.

3 Converse Lawvere property

We first need a lemma that tells us that we can replace certain partial O-computable functions with total ones when working relative to a reflective oracle. As discussed in the introduction, this contrasts strongly with the situation for computable functions. All of our theorems will make essential use of this lemma.

Lemma 1 (totalizing): There is a computable map $\tau : \mathbb{N} \to \mathbb{N}$ such that for all $i, n \in \mathbb{N}$ and any reflective oracle O, we have

$$P(\phi^{O}_{\tau(i)}(n){\downarrow} \in \{0,1\}) = 1$$

and for $b \in \{0,1\}$,

$$P(\phi^{O}_{\tau(i)}(n){\downarrow} = b) \geq P(\phi^{O}_{i}(n){\downarrow} = b).$$

Proof: We construct τ using a recursive procedure that ensures that $P(\phi^{O}_{\tau(i)}(n){\downarrow} = b)$ upper-bounds $P(\phi^{O}_{i}(n){\downarrow} = b)$ using what may be thought of as repeated subdivision or binary search. This is essentially the same as the function flip in [3], but we handle computability issues differently. Let $S \subseteq [0,1] \times [0,1]$ be the set of pairs of dyadic rationals (p, q) such that $p < q$. We recursively define a stochastic O-computable function $t : \mathbb{N} \times \mathbb{N} \times S \to \Delta\{0,1\}$; the intent is to have $P(t(i,n,p,q) = 1)$ equal

$$\frac{P(\phi^{O}_{\tau(i)}(n){\downarrow} = 1) - p}{q - p}$$

if that quantity is in the interval $[0,1]$, and take the closest possible value, either 0 or 1, otherwise. Then, we will be able to define $\phi^{O}_{\tau(i)}(n)$ to be $t(i,n,0,1)$.

Construct t so that a call $t(i,n,p,q)$ first queries $O(\langle i,n,r\rangle)$, where $r = \frac{p+q}{2}$, and also flips a fair coin C. Then, it either outputs 0, 1, or the result of a recursive call, as follows:

$$t(i,n,p,q) = \begin{cases} 0 & \text{if } O(\langle i,n,r\rangle) = 0 \text{ and } C = 0 \\ t(i,n,p,r) & \text{if } O(\langle i,n,r\rangle) = 0 \text{ and } C = 1 \\ 1 & \text{if } O(\langle i,n,r\rangle) = 1 \text{ and } C = 0 \\ t(i,n,r,q) & \text{if } O(\langle i,n,r\rangle) = 1 \text{ and } C = 1. \end{cases}$$

We can now choose τ so that $\phi^O_{\tau(i)}(n) = t(i,n,0,1)$.

This algorithm upper bounds the probabilities $P(\phi^O_i(n)\downarrow = 0)$ and $P(\phi^O_i(n)\downarrow = 1)$ by binary search. Once the initial call $t(i,n,0,1)$ has recursed to $t(i,n,p,q)$, it has already halted outputting 1 with probability p, confirming that this is an upper bound since it received answer 1 from a call to $O(\langle i,n,p\rangle)$. Similarly, it has output 0 with probability $1-q$, confirming this bound since it received the answer 0 from a call to $O(\langle i,n,q\rangle)$. Further, since each call to t halts without recursing with probability $\frac{1}{2}$, t halts almost surely. Thus, we get the totality property

$$P(\phi^O_{\tau(i)}(n)\downarrow \in \{0,1\}) = 1$$

and the bounds

$$P(\phi^O_{\tau(i)}(n)\downarrow = b) \geq P(\phi^O_i(n)\downarrow = b)$$

for $b \in \{0,1\}$. □

Now we can prove our main theorem.

Theorem 1 (converse Lawvere for reflective oracles): Let O be a reflective oracle. Then, there is an O-computable map $f: \mathbb{N} \to [0,1]^{\mathbb{N}}$ such that for all O-computable $g: \mathbb{N} \to [0,1]$, there is some index i such that $g = f(i)$.

Proof: Using τ from the totalizing lemma, let

$$f(i)(n) = P(\phi^{O}_{\tau(i)}(n)\downarrow = 1).$$

Given any O-computable $g\colon \mathbb{N} \to [0,1]$, there is some i such that

$$P(\phi^{O}_{i}(n)\downarrow = 1) = g(n)$$

$$P(\phi^{O}_{i}(n)\downarrow = 0) = 1 - g(n).$$

Then,

$$f(i)(n) = P(\phi^{O}_{\tau(i)}(n)\downarrow = 1) \geq P(\phi^{O}_{i}(n)\downarrow = 1) = g(n)$$

and similarly $1 - f(i)(n) \geq 1 - g(n)$, so $f(i) = g$. □

This theorem gives a computable analogue to the problem posed in [4]. The analogy would be strengthened if we worked in a cartesian closed category where the present notion of O-computability gave the morphisms, and where $[0,1]^{\mathbb{N}}$ is an exponential object. In addition, the totalizing lemma would have a nice analogue in this setting. I expect that all this can be done using something like an effective topos [6], but I leave this for future work.

4 Recovering Brouwer's fixed point theorem

As mentioned in [4], the intermediate value theorem would follow from the existence of a space with the converse Lawvere property, that is, a space X that surjects onto the mapping space $[0,1]^X$. Further, Brouwer's fixed point theorem on the n-ball, B^n, would follow from the existence of a topological space X with a surjection $X \to (B^n)^X$.

We can do something similar to conclude Brouwer's fixed point theorem from the converse Lawvere theorem for reflective oracles. Of course, this is circular; Kakutani's fixed point theorem, a generalization of Brouwer's fixed point theorem, is used to prove the existence of reflective oracles. Still, it is interesting to see how this can be done.

We start with two technical lemmas telling us some basic facts about reflective-oracle-computable functions $[0,1]^m \to [0,1]^m$. Using these, it will be easy to derive Brouwer's fixed point theorem.

Lemma 2 (continuous implies relatively computable): Take $m \in \mathbb{N}$ and let $h\colon [0,1]^m \to [0,1]^m$ be continuous. Then, there is a (deterministic) oracle O such that h is O-computable.

Proof: For each coordinate h_j of h, each rectangle

$$R = [\ell_1, u_1] \times \cdots \times [\ell_m, u_m] \subseteq [0,1]^m$$

with rational endpoints, and each pair of rationals $\ell_0, u_0 \in [0,1]_{\mathbb{Q}}$ with $\ell_0 < u_0$, let $O(\langle j, \ell_0, u_0, \ldots, \ell_m, u_m\rangle)$ be 1 if $h_j(R) \subseteq [\ell_0, u_0]$ and 0 otherwise. To see that h is O-computable, we compute any $h_j(p)$ for any j with $1 \leq j \leq m$, and any $p \in [0,1]^m$, making use of O and any oracle P compatible with p.

We proceed by a search process similar to the argument in the totalizing lemma. Start with $R^0 = [0,1]^m$, $\ell_0^0 = 0$ and $u_0^0 = 1$. At each step s, perform an exhaustive search for a rectangle

$$R^{s+1} = [\ell_1^{s+1}, u_1^{s+1}] \times \cdots \times [\ell_m^{s+1}, u_m^{s+1}] \subseteq R^s$$

and points ℓ_0^{s+1}, u_0^{s+1} such that a query to $P(\langle k, \ell_k^{s+1}\rangle)$ returns 1 for all k, a query to any $P(\langle k, u_k^{s+1}\rangle)$ returns 0, a query to $O(\langle j, \ell_0^{s+1}, u_0^{s+1}, R^{s+1}\rangle)$ returns 1, and where either $\ell_0^{s+1} = \ell_0^s$ and $u_0^{s+1} = \frac{1}{3}\ell_0^s + \frac{2}{3}u_0^s$, or $\ell_0^{s+1} = \frac{2}{3}\ell_0^s + \frac{1}{3}u_0^s$ and $u_0^{s+1} = u_0^s$. In the first case, output 0 with probability $\frac{1}{3}$ and continue with probability $\frac{2}{3}$, and in the second case, output 1 with probability $\frac{1}{3}$ and continue with probability $\frac{2}{3}$.

By construction, $p \in R^s$ and $h_j(R^s) \subseteq [\ell_0^s, u_0^s]$ at each stage s. Since h_j is continuous, there is some neighbourhood of p on which its image is contained in either

$[\ell_0^s, \frac{1}{3}\ell_0^s + \frac{2}{3}u_0^s]$ or $[\frac{2}{3}\ell_0^s + \frac{1}{3}u_0^s, u_0^s]$. There are two possibilities to consider. If $p \in \operatorname{int} R^s$, then there is some rectangle R contained in such a neighbourhood of p, and with $p \in \operatorname{int} R$. This rectangle would be chosen if considered, so the algorithm will move beyond step s.

The remaining possibility is that p is on the boundary of R^s; say, $p_k = \ell_k^s$. Since R^s was chosen previously though, we know that querying $P(k, \ell_k^{s+1})$ has returned 1 at least once, so $P(k, \ell_k^{s+1}) \neq 0$. Thus, the algorithm will almost surely eventually accept R^s or another rectangle.

Putting this together, this algorithm almost surely halts, outputting either 0 or 1. By stage s, it has halted outputting 1 with probability ℓ_0^s and outputting 0 with probability $1 - u_0^s$, so overall it outputs 1 with probability $h_j(p)$. Thus, h is O-computable. □

Lemma 3 (composition): Let O be a reflective oracle and let $g: \mathbb{N} \to [0,1]^m$ and $h: [0,1]^m \to [0,1]^m$ be O-computable. Then, $h \circ g: \mathbb{N} \to [0,1]^m$ is O-computable.

Proof: For each j with $1 \leq j \leq m$, take $i_j \in \mathbb{N}$ such that $\phi_{i_j}^O$ witnesses the computability of g_j. Then, $O(\langle i_j, n, r\rangle) = 0$ if $g_j(n) < r$ and $O(\langle i_j, n, r\rangle) = 1$ if $g_j(n) > r$, so O lets us simulate a probabilistic oracle compatible with $g(n)$. Hence, by the O-computability of h, for each k with $1 \leq k \leq m$, we have a probabilistic O-machine that always halts, outputting either 0 or 1, and that on input n outputs 1 with probability $h_k \circ g(n)$. □

Theorem 2 (Brouwer's fixed point theorem): Take $m \in \mathbb{N}$ and $h: [0,1]^m \to [0,1]^m$. Then, h has a fixed point.

Proof: By Lemma 2, we have an oracle O such that h is O-computable. By relativizing the construction of a reflective oracle [2,3] to O, we get a reflective oracle $\tilde{O}$ above O.

Notice that h is $\tilde{O}$-computable. Letting f be the $\tilde{O}$-computable function

$$f(i)(n) = P(\phi^{\tilde{O}}_{\tau(i)}(n)\downarrow = 1)$$

constructed in the converse Lawvere theorem for reflective oracles, define $f_m : \mathbb{N} \to ([0,1]^m)^{\mathbb{N}}$ by

$$f_m(\langle i_1, \ldots, i_m\rangle)(n) = (f(i_1)(n), \ldots, f(i_m)(n)).$$

The rest will now follow the proof of Lawvere's fixed point theorem [7].

Define $g : \mathbb{N} \to [0,1]^m$ by $g(n) = h(f_m(n)(n))$; this is $\tilde{O}$-computable by Lemma 3. Now, by converse Lawvere theorem, for each coordinate $1 \leq j \leq m$ of g, there is some $i_j \in \mathbb{N}$ such that $g_j = f(i_j)$. Letting $i = \langle i_1, \ldots, i_m\rangle$, we have

$$g_j(i) = h_j(f_m(i)(i)) = h_j(g(i)),$$

so $g(i) = h(g(i))$, and so $g(i)$ is a fixed point of h. □

5 Ubiquitous converse Lawvere property

Theorem 3 (ubiquitous converse Lawvere): Let O be a reflective oracle. Then, there is an O-computable, O-computably ubiquitous map $f : \mathbb{N} \to [0,1]^{\mathbb{N}}$.

Proof: This follows by a combination of the recursion theorem and the totalizing lemma. We use the same map f from Theorem 1. Let $e : \mathbb{N} \to [0,1]^{\mathbb{N}}$ be any O-computable map. There is a computable map $s : \mathbb{N} \to \mathbb{N}$ such that for all $i, n \in \mathbb{N}$, we have

$$P(\phi^{O}_{s(i)}(n)\downarrow \in \{0,1\}) = 1,$$

and

$$e(i)(n) = P(\phi^{o}_{s(i)}(n) \downarrow = 1).$$

By the recursion theorem, there is some i such that $\phi^{o}_{s(i)} = \phi^{o}_{i}$. Then,

$$e(i)(n) = P(\phi^{o}_{s(i)}(n) \downarrow = 1) = P(\phi^{o}_{i}(n) \downarrow = 1)$$

$$= P(\phi^{o}_{\tau(i)}(n) \downarrow = 1) = f(i)(n),$$

so $e(i) = f(i)$. □

References

[1] Kurt Gödel. 1946. "Remarks before the Princeton bicentennial conference of problems in mathematics." Reprinted in: Martin Davis. 1965. "The Undecidable. Basic Papers on Undecidable Propositions, Unsolvable Problems, and Computable Functions." Raven Press.

[2] Benja Fallenstein, Jessica Taylor, and Paul F. Christiano. 2015. "Reflective Oracles: A Foundation for Classical Game Theory." arXiv: 1508.04145.

[3] Benya Fallenstein, Nate Soares, and Jessica Taylor. 2015. "Reflective variants of Solomonoff induction and AIXI." In *Artificial General Intelligence*. Springer.

[4] Scott Garrabrant. 2017. "Formal Open Problem in Decision Theory." https://agentfoundations.org/item?id=1356.

[5] Scott Garrabrant. 2017. "The Ubiquitous Converse Lawvere Problem." https://agentfoundations.org/item?id=1372

[6] Jaap van Oosten. 2008. "Realizability: an introduction to its categorical side." *Studies in Logic and the Foundations of Mathematics*, vol. 152. Elsevier.

[7] F. William Lawvere. 1969. "Diagonal arguments and cartesian closed categories." In *Category Theory, Homology Theory and their Applications, II (Battelle Institute Conference, Seattle, Wash., 1968, Vol. Two)*, pages 134–145. Springer.

This post was originally published on Nov 30th 2017, and has been brought forwarded as part of the AI Alignment Forum launch sequences.

Tomorrow's AIAF sequences post will be 'Iterated Amplification and Distillation' by Ajeya Cotra, in the sequence on iterated amplification.