Universal Artificial Intelligence

Department of Philosophy
Central South University
xieshenlixi@163.com
github

November 3, 2020

Digression

I'm a philosopher.
What do you expect from me?

Digression

什么是哲学? 哲学是神学与科学的中间地带 — 罗素

Good philosophy in my eyes

- Beyes — *How to turn one's 'prior beliefs' into 'posterior beliefs'?*
- Cantor — *What is 'infinity'? What is 'set'?*
- Leibniz — *What are the extent and limits of reason?* — Universal Characteristic & Rational Calculus.
- Hilbert — *How to justify non-constructive reasoning?*
- Gödel — *What is the difference between 'proof' and 'truth'?*
- Tarski — *What is 'truth'? What are 'logical notions'?*
- Turing — *What is 'effective procedure'?*
- Kolmogorov — *What is 'simplicity'/'randomness'?*
- Solomonoff — *What is learnable? How to make induction?*
- Hutter/Schmidhuber — *What is 'intelligence'/'consciousness'?*

好的哲学工作是把哲学变成不是哲学的工作。
好的科学工作是把哲学变成不是哲学的工作。

Philosophy

Could a tadpole hope that it will survive to be a frog?

- Metaphysics: What kinds of things exist? (e.g. tadpole, material things, mental states, relationships)
- Epistemology: What can we know and how do we know it? Can we ever know about the contents of "other" minds?
- Philosophy of mind: What are mental states and processes? Are certain material states sufficient to produce mental states?
- Ethics: How should we think about the rights of a tadpole? Do we have the right to cause them pain?
- Philosophy of science: What are scientific theories? Explanations? Evidence? Can theories ever be proved, or refuted, and if so how? What's the relationship between the development of new concepts and the development of new theories?
- Conceptual analysis: What do we mean by X? What does it mean to say that a tadpole hope hopes for something?
- ...

A brief history of ontology

Descartes	Central role of mind. Dualism of mind and matter.
Kant	Reality is unknowable. Metaphysics is impossible. We can only know the quasi-fictional domains which we ourselves create. The mind imposes its categories on the data of experience.
Russell	Logic as mirror of reality.
Wittgenstein	Centrality of language and of language games.
Quine	Ontological commitment (study not: what there is, but: what sciences believe there is when logically formalized)
Hayes	Ontology to build Robots. Ontological commitment = study what people believe there is when logically formalized.

Contents

History of AI

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Reinforcement Learning

General Reinforcement Learning

Questions

- What is (artificial) intelligence?
- What does an intelligent system look like?
- Can there be a behavioural criterion for intelligence?
- Can computers think?
- Can connectionist networks think?
- Can physical symbol systems think?
- Do computing systems have "emergent" properties?
- Isn't changing the weight on a neural link a sort of symbol manipulation?
- Do computers have to be conscious to think?
- Are thinking computers mathematically possible?
- How will we know if we've done it?
- If we can do it, should we?

What is (Artificial) Intelligence?

What is AI?	**humanly**	**rationally**
Think	Cognitive Science	Laws of Thought
Act	Turing test, Behaviorism	Doing the Right Thing

Think like a human

- What cognitive capabilities are necessary to produce intelligent performance?
- **Cognitive science:** Models of the human thinking processes.
- **Advantages:** Models of the human thinking processes./Intelligible.
- **Difficulties:** The best artificial design for an intelligent system need not mirror the human mind.

Physical symbol system hypothesis(Newell & Simon): a physical symbol system has the necessary and sufficient means for general intelligent action.
Any system (human or machine) exhibiting intelligence must operate by manipulating data structures composed of symbols.

Act like a human: Turing Test

- Alan M. Turing, "Computing Machinery and Intelligence"
- John R. Searle, "Minds, Brains, and Programs"
- Interrogator in one room, human in another, system in a third.
- Interrogator in one room, human in another, system in a third.
- Interrogator tries to guess which is which.
- Interrogator tries to guess which is which.
- Chinese Room argument. — Syntax is insufficient for dealing with semantics.

Needs: natural language, knowledge representation, automated reasoning, machine learning.
Difficulties: Ambiguous./Not constructive./Cannot be formalized mathematically.

If we don't use the Turing test, what measure should we use?

Think rationally: Logicist AI

- What are the laws of thought?
- How should we think?
- **Logic:** Automatic reasoning procedure.
- **Advantages:** Precise./Search algorithm.
- **Difficulties:** Formalization of informal knowledge./Computational Cost.

Act rationally: Agents

- **Rational agent:** Autonomous system, capable of perceiving and interacting with its environment, of exploration (information gathering), learning and adaptation, of formulating goals and designing plans to reach those goals.
- The agent is rational, in the sense that it acts to achieve the best (expected) outcome, according to a performance measure, conditioned to its knowledge of the world and given computational resources.
- What to do, for example, when we must make a decision faced with insufficient information?

The Environment of Rational Agents

- Accessible vs. inaccessible (fully observable vs. partially observable)
  Are the relevant aspects of the environment accessible to the sensors?
- Deterministic vs. stochastic
  Is the next state of the environment completely determined by the current state and the selected action?
- Episodic vs. sequential
  Can the quality of an action be evaluated within an episode (perception + action), or are future developments decisive for the evaluation of quality?
- Static vs. dynamic
  Can the environment change while the agent is deliberating?
- Discrete vs. continuous
  Is the environment discrete or continuous?
- Single agent vs. multi-agent
  Which entities have to be regarded as agents? There are competitive and cooperative scenarios.

Solving Problems by Searching

- Before an agent can start searching for solutions, it must formulate a goal and then use that goal to formulate a problem.
- A problem consists of five parts: The state space, initial situation, actions, goal test and path costs. A path from an initial state to a goal state is a solution.
- A general search algorithm can be used to solve any problem. Specific variants of the algorithm can use different search strategies.
- Search algorithms are judged on the basis of completeness, time complexity, space complexity and optimality.

Criteria for Search Strategies

Completeness Is the strategy guaranteed to find a solution when there is one?

Time Complexity How long does it take to find a solution?

Space Complexity How much memory does the search require?

Optimality Does the strategy find the best solution (with the lowest path cost)?

Example — Missionaries and Cannibals

- Three missionaries and three cannibals are on one side of a river.
- A boat is available that can hold at most two people.
- You must never leave a group of missionaries outnumbered by cannibals on the same bank.
- Find a strategy that brings everyone safely to the opposite bank.

States	(x, y, z) with $0 \leq x, y \leq 3$ and $z \in \{0, 1\}$, where x, y and z represent the number of missionaries, cannibals and boat currently on the original bank.
Initial State	$(3, 3, 1)$
Goal State	$(0, 0, 0)$
Path Costs	1 unit per crossing.

$$(3,3,1) \rightarrow (3,1,0) \vee (2,2,0) \rightarrow (3,2,1) \rightarrow (3,0,0,) \rightarrow (3,1,1) \rightarrow$$
$$(1,1,0) \rightarrow (2,2,1) \rightarrow (0,2,0) \rightarrow (0,3,1) \rightarrow (0,1,0) \rightarrow$$
$$(0,2,1) \vee (1,1,1) \rightarrow (0,0,0)$$

How do computers discover new knowledge?

- Fill in gaps in existing knowledge
- Emulate the brain
- Simulate evolution
- Systematically reduce uncertainty
- Notice similarities between old and new

Machine Learning

- Supervised Learning
  - Learn the relationship between "input" x and "output" y: search for a function f, such that $y \approx f(x)$
  - There is training data with labels available
    Regression: y is metric variable (with values in $\mathbb{R}$)
    Classification: y is categorical variable (unordered, discrete).
  - Semi-supervised learning: also uses available unlabeled data, e.g. assumes that similar inputs have similar outputs.
- Unsupervised Learning
  - There exist no outputs, search for patterns within the inputs x
    Clustering: find groups of similar items
    Dimensionality reduction: describe data in fewer features
    Outlier detection: what is out of the ordinary?
    Association rules: which things often happen together?
- Reinforcement learning

The Five Tribes of Machine Learning

Tribe	**Origins**	**Master Algorithm**
Symbolists	Logic, philosophy	Inverse deduction
Connectionists	Neuroscience	Backpropagation
Evolutionaries	Evolutionary biology	Genetic programming
Bayesians	Statistics	Probabilistic inference
Analogizers	Psychology	Kernel machines

Symbolists

- All intelligence can be reduced to manipulating symbols
- Logic, Decision trees
- Inverse deduction
- Easy to add knowledge
- Can combine knowledge, data, to fill in gaps
- Impossible to code everything in rules

Representation	Rules, trees, first order logic rules
Evaluation	Accuracy, information gain
Optimization	Top-down induction, inverse deduction
Algorithms	Decision trees, Logic programs

Connectionists

- Learning is what the brain does: reverse-engineer it
- Hebbian learning: Neurons that fire together, wire together
- Neural networks
- Backpropagation
- Can handle raw, high-dimensional data, constructs it own features
- Hard to add reasoning/explanations

Representation	Neural network
Evaluation	Squared error
Optimization	Gradient descent
Algorithms	Backpropagation

Evolutionaries

- Natural selection is the mother of all learning
- Evolutionary algorithms
- Crossover, mutation
- Can learn structure, wide hypothesis space
- Needs a way to 'fill' the structure

Representation	Genetic programs (often trees)
Evaluation	Fitness function
Optimization	Genetic search
Algorithms	Genetic programming (crossover, mutation)

Bayesians

- Learning is a form of uncertain inference
- Graphical models, Gaussian processes, HMMs, Kalman filter
- Uses Bayes theorem to incorporate new evidence into our beliefs
- Can deal with noisy, incomplete, contradictory data
- Depends on the prior
- Hard to do unite logic and probability

Representation	Graphical models, Markov networks
Evaluation	Posterior probability
Optimization	Probabilistic inference
Algorithms	Bayes theorem and derivates

Analogizers

- You are what you resemble
- Recognizes similarities between situations and infers other similarities
- k-Nearest Neighbor, Support Vector Machines
- Transfer solution from previous situations to new situations
- Hard to do rules and structure

Representation	Memory, support vectors
Evaluation	Margin
Optimization	Kernel machines
Algorithms	k-Nearest Neighbor, Support Vector Machines

The Master Algorithm?

Tribe	**Problem**	**Solution**
Symbolists	Knowledge composition	Inverse deduction
Connectionists	Credit assignment	Backpropagation
Evolutionaries	Structure discovery	Genetic programming
Bayesians	Uncertainty	Probabilistic inference
Analogizers	Similarity	Kernel machines

- Representation: The hypothesis space.
  - Probabilistic logic
  - Weighted formulas $\longrightarrow$ Distribution over states
- Evaluation: How to choose one hypothesis over the other?
  - Posterior probability
  - User-defined objective function
- Optimization: How do we search the hypothesis space?
  - Formula discovery: Genetic programming
  - Weight learning: Backpropagation

Elegant/Extensible/Expressive/Efficient/Educable/Evolvable?

Learn to Learn

1. Good Old-Fashioned AI
   - Handcraft predictions
   - Learn nothing
2. Shallow Learning
   - Handcraft features
   - Learn predictions
3. Deep Learning
   - Handcraft algorithm (optimiser, target, architecture, ...)
   - Learn features and predictions end-to-end
4. Meta Learning
   - Handcraft nothing
   - Learn algorithm and features and predictions end-to-end

Can Machines Think?

- Theological objections.
- Argument from informality of behavior. — Human behavior is far too complex to be captured by any simple set of rules./Learning from experience.
- Argument from incompleteness theorems. — No formal system incl. AIs, but only humans can "see" that Gödel's unprovable sentence is true./Lucas can't consistently assert that this sentence is true.
- Machines can't be conscious or feel emotions. — Reductionism doesn't really answer the question: why can't machines be conscious or feel emotions?
- Machines don't have Human Quality X.
- Machines just do what we tell them to do. — Maybe people just do what their neurons tell them to do.
- Machines are digital. Mental states can emerge from neural substrate only. — Only the functionality/behavior matters.
- Non-computable Physics & Brains.

Brain Dissection

- The "brain in a vat" experiment: (no) real experience.
- The brain prosthesis experiment:
  - Replacing some neurons in the brain by functionally identical electronic prostheses would neither effect external behavior nor internal experience of the subject.
  - Successively replace one neuron after the other until the whole brain is electronic.

Fermi Paradox — Where are the aliens?

- There are none, i.e. we're all alone.
- We can't detect them because...
  - we're too primitive or too far apart
  - there are predators or all fear them
  - we're lied to, live in a simulation
  - ...

Why do we need to align machine agents?

- **Goal orthogonality.**
  Intelligence and final goals are orthogonal: Any utility function can be combined with a powerful epistemology and decision theory.
- **Instrumental convergence.**
  Different long-term goals imply similar short-term strategies.
  - Self-preservation
  - Retention of goals through time
  - Cognitive enhancement
  - Technological perfection
  - Resource acquisition
  - ...
- **Capability gain.**
  There are potential ways for artificial agents to greatly gain in cognitive power and strategic options.
- **Alignment difficulty.**
  It's hard to transmit our values to AI systems or avert adversarial incentives.

Asimov's Laws of Robotics

- **Law Zero**
  A robot may not injure humanity, or, through inaction, allow humanity to come to harm.
- **Law One**
  A robot may not injure a human being or, through inaction, allow a human being to come to harm, unless this would violate a higher order law.
- **Law Two**
  A robot must obey orders given it by human beings except where such orders would conflict with a higher order law.
- **Law Three**
  A robot must protect its own existence as long as such protection does not conflict with a higher order law.

An Extended Set of the Laws of Robotics

- **The Meta-Law**
  A robot may not act unless its actions are subject to the Laws of Robotics.
- **Law Zero**
  A robot may not injure humanity, or, through inaction, allow humanity to come to harm.
- **Law One**
  A robot may not injure a human being, or, through inaction, allow a human being to come to harm, unless this would violate a higher order Law.
- **Law Two**
  1. A robot must obey orders given it by human beings, except where such orders would conflict with a higher order Law.
  2. A robot must obey orders given it by superordinate robots, except where such orders would conflict with a higher order Law.
- **Law Three**
  1. A robot must protect the existence of a superordinate robot as long as such protection does not conflict with a higher order Law.
  2. A robot must protect its own existence as long as such protection does not conflict with a higher order Law.
- **Law Four**
  A robot must perform the duties for which it has been programmed, except where that would conflict with a higher order law.
- **The Procreation Law**
  A robot may not take any part in the design or manufacture of a robot unless the new robot's actions are subject to the Laws of Robotics.

Ethical Concerns

- Is it morally justified to create intelligent systems with these constraints?
- Would it be possible to do so?
- Should intelligent systems have free will? Can we prevent them from having free will? What could it mean for a machine to have its own goals? Do we have a kind of freedom machines could never have?
- Will intelligent systems have consciousness?
- If they do, will it drive them insane to be constrained by artificial ethics placed on them by humans?
- If intelligent systems develop their own ethics and morality, will we like what they come up with?

Machine Ethics

- People might lose their jobs to automation.
- People might have too much (or too little) leisure time.
- People might lose their sense of being unique.
- People might lose some of their privacy rights.
- The use of AI systems might result in a loss of accountability.
  — Who is responsible if a physician follows the advice of a medical expert system, whose diagnosis turns out to be wrong?
- The success of AI might mean the end of the human race.

What If We Do Succeed?

- Natural selection is replaced by artificial evolution.
  — AI systems will be our mind children.
- Once a machine surpasses the intelligence of a human it can design even smarter machines.
- This will lead to an intelligence explosion and a technological singularity at which the human era ends.
- Prediction beyond this event horizon will be impossible.
- Alternative 1: We keep the machines under control.
  - Capability control (limiting what the system can or does do).
  - Motivation selection (controlling what the system wants to do).
- Alternative 2: Humans merge with or extend their brain by AI.

Single-Shot Situation

Our first superhuman AI must be a safe one for we may not get a second chance!

Singularity

Time	Speed Explosion. — Computing speed doubles every two subjective years of work.
Quantitative	Population Explosion. — Computing costs halve for a certain amount of work.
Qualitative	Intelligence Explosion. — Proportionality Thesis: An increase in intelligence leads to similar increases in the capacity to design intelligent systems.

Comparison of Main Ethical Theories

	Consequentialism	**Deontology**	**Virtue Ethics**
Description	An action is right if it maximizes happiness	An action is right if it is in accordance with a moral rule	An action is right if it is what a virtuous person would do in the circumstances
Central Concern	The results matter, not the actions themselves	Persons must be seen as ends and may never be used as means	Emphasize the character of the agent making the actions
Guiding Value	Good (often seen as maximum happiness)	Right (rationality is doing one's moral duty)	Virtue (leading to the attainment of eudaimonia)
Practical Reasoning	The best for most (means-ends reasoning)	Follow the rule (rational reasoning)	Practice human qualities (social practice)
Deliberation Focus	Consequences (What is outcome of action?)	Action (Is action compatible with some imperative?)	Motives (Is action motivated by virtue?)

Capability Control Methods

- Boxing: the system can only act through restricted channels.
  — The AI could persuade someone to free it from its box.
- Incentives: access to other AIs, cryptographic reward tokens.
- Stunting: imposing constraints on the system's cognitive abilities
- Tripwires: diagnostic tests run periodically to check for dangerous activity, with shutdown a consequence of detection.

Motivation Selection Methods

- Direct specification of motivations
  - Rule-based methods: give the machine a set of rules that define its final goals.
  - Direct consequentialist methods: specify some measure that is to be maximised. (e.g. human happiness.)
- Augmentation
  - Start with an AI with human-level intelligence, that has an acceptable motivation system: then enhance its cognitive faculties to make it superintelligent.
- Indirect normativity
  - Rather than specifying a normative standard directly, we specify a process for deriving a standard. We then build the system so it is motivated to carry out this process.
  - Value learning. — inverse reinforcement learning. — wireheading.

Prisoner's Dilemma

- Difficult to prevent arms races.
- The winner takes all (of what remains).
- Arms races are dangerous because parties sacrifice safety for speed! — When headed the wrong way, the last thing we need is progress.

International Cooperation

In face of uncertainty, cooperation is robust!

Why should I care about the world when I am dead and gone? I want it to go fast, damn it! This increases the chance I have of experiencing a more technologically advanced future.

AI Alignment

- Important to ensure it's aligned with our interests
  - But how do we specify beneficial goals?
  - How do we make sure system actually pursues them?
  - How do we correct the system if we get it wrong?
- Want solid theoretical understanding of problem & solution
  - What is correct reasoning and decision making?
  - Probability theory, decision theory, game theory, statistical learning theory, Bayesian networks, formal verification, ...

Technical Research Questions

1. Reliable self-modification
2. Logical uncertainty (reasoning without logical omniscience)
3. Reflective stability of decision theory
4. Decision theory for Newcomb-like problems
5. Corrigibility (accepting modifications)
6. The shutdown problem
7. Value loading
8. Indirect specification of decision theory
9. Domesticity (goal specification for limited impact)
10. The competence gap
11. Weighting options or outcomes for variance-normalizing solution to moral uncertainty
12. Program analysis for self-improvement
13. Reading values and beliefs of AIs
14. Pascal's mugging
15. Infinite ethics
16. Mathematical modelling of intelligence explosion

Before the prospect of an intelligence explosion, we humans are like children playing with a bomb. Such is the mismatch between the power of our play-thing and the immaturity of our conduct. Superintelligence is a challenge for which we are not ready now and will not be ready for a long time. We have little idea when the detonation will occur, though if we hold the device to our ear we can hear a faint ticking sound.

— Nick Bostrom

Contents

History of AI

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Reinforcement Learning

General Reinforcement Learning

Contents

History of AI

Philosophy of Induction
- History
- How to Choose the Prior?

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Reinforcement Learning

General Reinforcement Learning

Hume & Russell

Proposition (Hume)

Induction is just a mental habit, and necessity is something in the mind and not in the events.

Leibniz-Wittgenstein-Goodman

Proposition (Leibniz)

Since for any finite number of points there are always infinitely many curves going through them, any finite set of data is compatible with infinitely many inductive generalizations.

Proposition (Wittgenstein)

Since any finite course of action is in accord with infinitely many rules, no universal rule can be learned by examples.

Proposition (Goodman)

All emeralds discovered till 2050 are green, and blue thereafter.

Mill — Homogeneous Universe

Proposition (Mill)

Induction can be turned into a deduction, by adding principles about the world (such as 'the future resembles the past', or 'space-time is homogeneous').

Homogeneous?

Problem

$1, 3, 5, 7, 9, 11, 13, 15, ?$

Solution

$$2n - 1 \tag{17?}$$

$$2n - 1 + \prod_{i=1}^{8}(n - i) \tag{17+8!?}$$

$$\vdots \tag{??}$$

Epicurus vs Occam

Proposition (Epicurus)

If more than one hypothesis is consistent with the observations, keep them all.

Proposition (Occam's Razor)

Among the hypotheses that are consistent with the observed phenomena, select the simplest one.

- Entities should not be multiplied beyond necessity.
- Wherever possible, logical constructions are to be substituted for inferred entities.
- It is vain to do with more what can be done with fewer.

Why Simplicity? — Gestalt Psychology

Figure: Gestalt Psychology

Why Simplicity?

God does not play dice.
God always takes the simplest way.
Subtle is the Lord, but malicious He is not.
The most incomprehensible thing about the world is that it is comprehensible.
What really interests me is whether God could have created the world any differently; in other words, whether the requirement of logical simplicity admits a margin of freedom.
When I am judging a theory, I ask myself whether, if I were God, I would have arranged the world in such a way.

— *Einstein*

- Principle of least/stationary action
- Noether's theorem
- ...

Why Simplicity?

program	$\xrightarrow{\textit{Computer}}$	*output*
axioms	$\xrightarrow{\textit{Deduction}}$	*theorems*
scientific theory	$\xrightarrow{\textit{Calculations}}$	*experimental data*
encoded message	$\xrightarrow{\textit{Decoder}}$	*original message*
software	$\xrightarrow{\textit{Universal Constructor}}$	*physical system*
DNA	$\xrightarrow{\textit{Pregnancy}}$	*organism*
Ideas	$\xrightarrow{\textit{Mind of God}}$	*Universe*

Popper — The Logic of Scientific Discovery

Proposition (Popper)

- *A single observational event may prove hypotheses wrong, but no finite sequence of events can verify them correct.*
- *Induction is theoretically unjustifiable and becomes in practice the choice of the simplest generalization that resists falsification.*
- *The simpler a hypothesis, the easier it is to be falsified.*
- *Falsifiability is as subjective as simplicity, there is no objective criterion.*

Kaynes $\Longrightarrow$ Carnap

- Assign to inductive generalizations probabilities that should converge to 1 as the generalizations are supported by more and more independent events.

  — Kaynes

- Observational events provide, if not proofs, at least positive confirmations of scientific hypotheses. Chose the generalization that confirm more evidence.

  — Carnap

Philosophy of Induction

What is learnable? How to learn?
How can we know that what we learned is true?

History

Possible Worlds/Hypothesis (Epicurus/Leibniz)

$+$

Homogeneous Universe(s) (Mill/Turing)

$+$

Simplicity Criterion (Occam/Kolmogorov)

$+$

Prior Belief (Carnap/Solomonoff)

$+$

Update Belief (Bayes)

$\Downarrow$

Convergence to Truth

$$P(h|e) = \frac{P(e|h)P(h)}{\sum_{h \in \mathcal{H}} P(e|h)P(h)} \xrightarrow{\ell(e)\to\infty} 1$$

MDL vs Bayesian Mixture

When solving a problem of interest, do not solve a more general problem as an intermediate step.

Bayesianism

Theorem (Convergence Theorem)

$$\sum_{t=1}^{n} \mathbb{E}_\mu \left[\left(\sqrt{\frac{\rho(a|x_{<t})}{\mu(a|x_{<t})}} - 1 \right)^2 \right] \leq \sum_{t=1}^{n} \mathbb{E}_\mu \left[\sum_{a \in \mathcal{X}} \left(\sqrt{\rho(a|x_{<t})} - \sqrt{\mu(a|x_{<t})} \right)^2 \right] \leq D_n(\mu \| \rho)$$

$$\sum_{t=1}^{n} \mathbb{E}_\mu \left[\sum_{a \in \mathcal{X}} (\rho(a|x_{<t}) - \mu(a|x_{<t}))^2 \right] \leq D_n(\mu \| \rho)$$

$$\frac{1}{2n} \left(\sum_{t=1}^{n} \mathbb{E}_\mu \left[\sum_{a \in \mathcal{X}} \left| \rho(a|x_{<t}) - \mu(a|x_{<t}) \right| \right] \right)^2 \leq D_n(\mu \| \rho)$$

where

$$D_n(\mu \| \rho) := \mathbb{E}_\mu \left[\ln \frac{\mu(x_{1:n})}{\rho(x_{1:n})} \right]$$

Theorem

$$\xi = \operatorname*{argmin}_{\rho} \mathbb{E}_w \left[D(\mu \| \rho) \right] \quad \textit{where} \quad \xi(x) := \sum_{\nu \in \mathcal{M}} w_\nu \nu(x)$$

Bayesian Decisions

Suppose $\mathrm{Loss}(x_t, y_t) \in [0,1]$

$$y_t^{\Lambda_\rho}(x_{<t}) \;:=\; \arg\min_{y_t} \sum_{x_t} \rho(x_t|x_{<t})\,\mathrm{Loss}(x_t, y_t)$$

$$L^{\Lambda_\rho}(x_{<t}) \;:=\; \mathbb{E}_\mu\left[\mathrm{Loss}\left(x_t, y_t^{\Lambda_\rho}\right)\middle| x_{<t}\right]$$

$$L_n^{\Lambda_\rho} \;:=\; \sum_{t=1}^{n} \mathbb{E}_\mu\left[L^{\Lambda_\rho}(x_{<t})\right]$$

Theorem

$$\left(\sqrt{L_n^{\Lambda_\xi}} - \sqrt{L_n^{\Lambda_\mu}}\right)^2 \le 2\sum_{t=1}^{n} \mathbb{E}_\mu\left[\sum_{a\in\mathcal{X}}\left(\sqrt{\xi(a|x_{<t})} - \sqrt{\mu(a|x_{<t})}\,\right)^2\right] \le 2D_n(\mu\|\xi)$$

$$L_n^{\Lambda_\xi} - L_n^{\Lambda_\mu} \le 2D_n(\mu\|\xi) + 2\sqrt{L_n^{\Lambda_\mu} D_n(\mu\|\xi)}$$

Problem

How to choose the model class and prior?

- choose the smallest model class that will contain the true environment.
- choose the priors that best reflect a rational a-priori belief in each of these environments.
  1. Convergence of Bayesian mixture to true environment.
  2. Confirmation of "the sun will always rise".
  3. Invariance Criterion.
     reparametrization & regrouping invariant.

Contents

History of AI

Philosophy of Induction
- History
- How to Choose the Prior?

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Reinforcement Learning

General Reinforcement Learning

Invariance Criterion

By applying some principle to a parameter θ we get prior $w(\theta)$. If we consider some new parametrization θ' via $f : \theta \mapsto \theta'$, then we get a prior $\tilde{w}(\theta')$ by transforming the original prior via f.
for discrete class $\mathcal{M}$,

$$\tilde{w}(\theta') := \sum_{\theta : f(\theta) = \theta'} w(\theta)$$

for continuous parametric class $\mathcal{M}$,

$$\tilde{w}(\theta') := \int \delta(f(\theta) - \theta') w(\theta) \mathrm{d}\theta \qquad \text{(Dirac-delta)}$$

Regrouping-invariant:

$$\tilde{w}(\theta') = w'(\theta')$$

where $w'(\theta')$ is obtained by applying the same principle to the new parametrization.
We say the principle is reparametrization-invariant when f is bijective.

How to Assign Prior? Indifference Principle/MaxEnt

How to Assign Prior?

- The principle of indifference. Assume $w(\theta) = 1$ and $\theta' = \sqrt{\theta}$.

$$\tilde{w}(\theta') = \int_0^1 \delta\big(\sqrt{\theta} - \theta'\big) w(\theta) \mathrm{d}\theta = 2\sqrt{\theta} \neq w'(\theta')$$

- Maximize the entropy subject to some constraints provided by empirical data or considerations of symmetry, probabilistic laws, and so on.
- Occam's razor.

How to confirm "All Ravens are Black"?

Problem 1 — All Ravens are Black $\theta = 1$

Suppose θ is the percentage of ravens that are black.
"All ravens are black" $\equiv \theta = 1$. or, 1^∞ or, $\forall x : R(x) \to B(x)$

$$P(\theta = 1) = \int_1^1 w(\theta)\mathrm{d}\theta = 0 \qquad (0 \text{ prior})$$

$$\Downarrow$$

$$P(\theta = 1|1^n) = \frac{P(1^n|\theta = 1)P(\theta = 1)}{P(1^n)} = 0 \qquad (\times)$$

Problem 2 — The Sun will always Rise 1^∞

Indifference Principle

$$\left.\begin{array}{c}\displaystyle\int_0^1 w(\theta)\mathrm{d}\theta = 1 \\ \forall\theta,\theta' : w(\theta) = w(\theta')\end{array}\right\} \implies \forall\theta : w(\theta) = 1$$

The sun will rise tomorrow. ✓

$$P(x) = \int_0^1 P(x|\theta)w(\theta)\mathrm{d}\theta \implies P(1|1^n) = \frac{n+1}{n+2}$$

The sun will always rise. ×

$$P(1^\infty|1^n) = \lim_{k\to\infty} P(1^k|1^n) = \lim_{k\to\infty} \frac{n+1}{n+k+1} = 0$$

Solution 1 — Soft Hypothesis — No absolute truth!

$$H_\varepsilon = \{\theta : \theta \in (1-\varepsilon, 1]\}$$

$$P(H_\varepsilon) = \int_{1-\varepsilon}^{1} w(\theta)\mathrm{d}\theta = \varepsilon > 0$$

$$\begin{aligned}
P(H_\varepsilon|1^n) &= \int_{1-\varepsilon}^{1} w(\theta|1^n)\mathrm{d}\theta \\
&= \int_{1-\varepsilon}^{1} \frac{P(1^n|\theta)w(\theta)}{P(1^n)}\mathrm{d}\theta \\
&= \int_{1-\varepsilon}^{1} (n+1)\theta^n \mathrm{d}\theta \\
&= \theta^{n+1}\Big|_{1-\varepsilon}^{1} \\
&= 1-(1-\varepsilon)^{n+1} \xrightarrow{n\to\infty} 1
\end{aligned}$$

Solution 2 — ad hoc

$$w(\theta) := \frac{1}{2}(1 + \delta(1 - \theta))$$

Dirac-delta sifting property: $\displaystyle\int f(\theta)\delta(\theta - a)\,\mathrm{d}\theta = f(a)$

$$\begin{aligned}
P(x) &= \int_0^1 P(x|\theta)w(\theta)\mathrm{d}\theta \\
&= \int_0^1 \theta^s(1-\theta)^f \cdot \frac{1}{2}(1 + \delta(1-\theta))\mathrm{d}\theta \\
&= \frac{1}{2}\int_0^1 \theta^s(1-\theta)^f\,(1 + \delta(\theta - 1))\mathrm{d}\theta \\
&= \frac{1}{2}\left(\frac{s!f!}{(s+f+1)!} + 1^s \cdot (1-1)^f\right) && \text{[sifting property]} \\
&= \frac{1}{2}\left(\frac{s!f!}{(s+f+1)!} + \delta_{f,0}\right)
\end{aligned}$$

Solution 2 — ad hoc

$$P(1^\infty|1^n) = \lim_{k\to\infty} \frac{P(1^{n+k})}{P(1^n)} = \lim_{k\to\infty} \frac{\frac{1}{2}\left(\frac{(n+k)!0!}{(n+k+1)!} + 1\right)}{\frac{1}{2}\left(\frac{n!0!}{(n+1)!} + 1\right)} = \frac{n+1}{n+2} \xrightarrow{n\to\infty} 1$$

$$P(\theta \geq a) = \int_a^1 \frac{1}{2}(1 + \delta(\theta - 1))\mathrm{d}\theta = 1 - \frac{1}{2}a$$

$$P(\theta = 1) = \frac{1}{2}$$

$$P(\theta = 1|1^n) = \frac{P(1^n|\theta = 1)P(\theta = 1)}{P(1^n)} = \frac{1 \cdot \frac{1}{2}}{\frac{1}{2}\left(\frac{n!0!}{(n+1)!} + 1\right)} = \frac{n+1}{n+2} \xrightarrow{n\to\infty} 1$$

Why $\theta = 1$ special?

Contents

History of AI

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Reinforcement Learning

General Reinforcement Learning

Natural Wish List

- (computability) $P_n(A)$ is computable.
- (convergence) $P(A) = \lim_{n \to \infty} P_n(A)$
- (coherent limit) $P(A \wedge B) + P(A \vee B) = P(A) + P(B)$
- (non-dogmatism) If $\nvdash A$ then $P(A) < 1$, and if $\nvdash \neg A$ then $P(A) > 0$.

Unary Pure Inductive Logic

- $\mathscr{L}$ contains countable constants $\mathcal{C}$ and m unary predicates.
- $\mathcal{R} = \{R_1, R_2, \ldots, R_m\}$ with no function symbols nor equality.
- $Q_i := \bigwedge_{j=1}^{m} \pm R_j$ for $1 \leq i \leq 2^m =: r$.
- $\mathcal{Q} = \{Q_1, \cdots, Q_r\}$ is a r-fold classification system of some Universe with domain $\mathcal{C}$.

Definition (Probability on Sentences)

A probability on sentences is a non-negative function $w : \mathcal{S} \to [0, 1]$ s.t.

P_1. $\vDash A \implies w(A) = 1$

P_2. $A \vDash \neg B \implies w(A \vee B) = w(A) + w(B)$

P_3. $w\big(\exists x A(x)\big) = \lim_{n\to\infty} w\left(\bigvee_{i=1}^{n} A(a_i)\right)$

Properties

Theorem

(i) $w(\neg A) = 1 - w(A)$

(ii) $\vDash \neg A \implies w(A) = 0$

(iii) *The following are equivalent:*

1. $w(A) = 1 \implies \vDash A$
2. $w(A) = 0 \implies \vDash \neg A$

(iv) $A \vDash B \implies w(A) \leq w(B)$

(v) $\vDash A \leftrightarrow B \implies w(A) = w(B)$

(vi) $w(A) + w(B) = w(A \wedge B) + w(A \vee B)$

Theorem (Extension Theorem)

For any probability function over quantifier-free sentences $w : \mathcal{S} \to [0, 1]$ *satisfying* P_1, P_2, w *has an unique extension to* $\overline{w} : \mathcal{S} \to [0, 1]$ *satisfying* P_1, P_2, P_3.

Possible Worlds

- **state description** $\bigwedge_{i=1}^{n} Q_{h_i}(a_i)$
  where $h : \mathcal{C} \to \mathcal{Q}$
- **structure description** $\{n_i\}_{i=1}^{r}$
  where $n_i := \sum_{j=1}^{n} [\![h_j = i]\!]$
- **rank description** $\{m_i\}_{i=0}^{n}$
  where $m_i := \sum_{j=1}^{r} [\![n_j = i]\!]$

  Obviously, $\sum_{i=1}^{n} i \cdot m_i = n$ and $\sum_{i=0}^{n} m_i = r$.

Indifference Principle

A All state descriptions have equal weight.

B All structure descriptions have equal weight.

C Each nonempty subset of the alphabet is equally likely.

D Each nonzero cardinality is equally likely.

E All rank descriptions have equal weight.

Given n individuals, there are r^n possible state descriptions,

$$\left|\left\{(n_1, \ldots, n_r) : \sum_{i=1}^{r} n_i = n\right\}\right| = \binom{n+r-1}{r-1}$$

possible structure descriptions, and

$$p(n,r) := \left|\left\{(m_0, \ldots, m_n) : \sum_{i=1}^{n} i \cdot m_i = n \ \ \& \ \ \sum_{i=0}^{n} m_i = r \ \ \& \ \ \forall i : m_i \geq 0\right\}\right|$$

possible rank descriptions.

(A) State Description ×

According to (A),

$$m^\dagger\left(\bigwedge_{i=1}^{n} Q_{h_i}(a_i)\right) = \frac{1}{r^n}$$

$$c^\dagger\left(Q_j(a_{n+1}) \,\middle|\, \bigwedge_{i=1}^{n} Q_{h_i}(a_i)\right) = \frac{m^\dagger\left(\bigwedge_{i=1}^{n} Q_{h_i}(a_i) \wedge Q_j(a_{n+1})\right)}{m^\dagger\left(\bigwedge_{i=1}^{n} Q_{h_i}(a_i)\right)} = \frac{1}{r}$$

(B) Structure Description ✓

According to (B),

$$m^*(n_1, \ldots, n_r) = \frac{1}{\binom{n+r-1}{r-1}}$$

Structure Description $(n_1, \ldots, n_r)$ corresponds to $\binom{n}{n_1, \ldots, n_r}$ State Descriptions.

$$m^*\left(\bigwedge_{i=1}^{n} Q_{h_i}(a_i)\right) = \frac{m^*(n_1, \ldots, n_r)}{\binom{n}{n_1, \ldots, n_r}} = \frac{1}{\binom{n+r-1}{r-1}\binom{n}{n_1, \ldots, n_r}}$$

Sometimes we write $m^*(h_{1:n}) := m^*\left(\bigwedge_{i=1}^{n} Q_{h_i}(a_i)\right)$ for short.

Carnap's Degree of Confirmation

Carnap's Degree of Confirmation

$$c^*\left(Q_j(a_{n+1}) \,\middle|\, \bigwedge_{i=1}^{n} Q_{h_i}(a_i)\right) = \frac{m^*\left(\bigwedge_{i=1}^{n} Q_{h_i}(a_i) \wedge Q_j(a_{n+1})\right)}{m^*\left(\bigwedge_{i=1}^{n} Q_{h_i}(a_i)\right)} = \frac{n_j + 1}{n + r}$$

frequency — independent identical distribution(i.i.d)
extension

$$c^*(A)$$

(C,D,E)

According to (C),

$$m^{\$}(h_{1:n}) = \frac{1}{\left(\sum_{i=1}^{\min\{r,n\}} \binom{r}{i}\right) \binom{n-1}{r-m_0-1} \binom{n}{n_1,\ldots,n_r}}$$

According to (D),

$$m^{\#}(h_{1:n}) = \frac{1}{\min\{r,n\} \binom{r}{r-m_0} \binom{n-1}{r-m_0-1} \binom{n}{n_1,\ldots,n_r}}$$

According to (E),

$$m^{\tau}(h_{1:n}) = \frac{1}{\binom{n}{n_1,\ldots,n_r} \binom{r}{m_0,\ldots,m_n} p(n,r)}$$

Rank Description $(m_0, \ldots, m_n)$ corresponds to $\binom{r}{m_0,\ldots,m_n}$ Structure Descriptions.

What is the right w?

- Constant Exchangeability Principle.
  For any permutation σ of $\mathbb{N}^+$,

$$w(A(a_1, \ldots, a_n)) = w(A(a_{\sigma(1)}, \ldots, a_{\sigma(n)})) \tag{Ex}$$

- Atom Exchangeability Principle.
  For any permutation τ of $\{1, 2, \ldots, r\}$,

$$w\left(\bigwedge_{i=1}^{n} Q_{h_i}(a_i)\right) = w\left(\bigwedge_{i=1}^{n} Q_{\tau(h_i)}(a_i)\right) \tag{Ax}$$

- Sufficientness Postulate.

$$w\left(Q_j(a_{n+1}) \,\middle|\, \bigwedge_{i=1}^{n} Q_{h_i}(a_i)\right) = f_j(n_j, n) \tag{SP}$$

What is the right w?

Principle Ex asserts that $w\left(\bigwedge_{i=1}^{n} Q_{h_i}(a_i)\right)$ depends only on the vector $\langle n_{h_i} : 1 \leq i \leq n\rangle$, so that it is independent on the order of observing the individuals, while in the presence of Ex, principle Ax asserts that $w\left(\bigwedge_{i=1}^{n} Q_{h_i}(a_i)\right)$ depends only on $\{n_i : 1 \leq i \leq r\}$, and $w(Q_i(a_1)) = 1/r$ for all $1 \leq i \leq r$.

Carnap's λ-continuum

Theorem

Suppose language $\mathscr{L}$ has at least two predicates i.e. $m \geq 2$, then the probability function w on $\mathscr{L}$ satisfies Ex, SP iff $w = c_\lambda$ for some $0 \leq \lambda \leq \infty$. Namely,

$$f_i(n_i, n) = \frac{n_i + \lambda\gamma_i}{n + \lambda}$$

where $\gamma_i = f_i(0,0)$ and $\lambda = \dfrac{f_i(0,1)}{f_i(0,0) - f_i(0,1)}$.

By adding Ax, $\forall i : \gamma_i = \dfrac{1}{r}$.

Shortcoming 1 — All Ravens are Black? ×

$$c^*\left(\forall x\left(R(x) \to B(x)\right)\right) \leq \lim_{n\to\infty} \prod_{i=0}^{n-1} \frac{i+r-1}{i+r} = 0$$

Convergence Speed? Yes and No!

$$\prod_{n\geq 1} a_n = 0 \iff \sum_{n\geq 1}(1-a_n) = \infty \quad \text{for} \quad \forall n : 0 < a_n \leq 1$$

Shortcoming 2 — No-Free-Lunch for Carnap!

Strengthened "Hume" — Wolpert & Macready 1997, Igel & Toussaint 2004:

No-Free-Lunch Theorem!

All state descriptions with the same structure description have equal weight!

Block Uniform

No Free Lunch Theorem — Strengthened "Hume"

Theorem (No Free Lunch Theorem)

If and only if the probability distribution P is block uniform, i.e.

$$\forall f, g \in \mathcal{Y}^{\mathcal{X}} : \forall y \in \mathcal{Y} \left(\left| f^{-1}(y) \right| = \left| g^{-1}(y) \right| \right) \implies P(f) = P(g)$$

then for any two algorithms A, A', any value $k \in \mathbb{R}$, any $m \in \{1, \ldots, |\mathcal{X}|\}$, and any performance measure L,

$$\sum_{f \in \mathcal{Y}^{\mathcal{X}}} P(f) \left[\!\left[k = L\left(T_m^y(A, f)\right) \right]\!\right] = \sum_{f \in \mathcal{Y}^{\mathcal{X}}} P(f) \left[\!\left[k = L\left(T_m^y(A', f)\right) \right]\!\right]$$

where $T_n := \langle (x_1, f(x_1)), \ldots, (x_n, f(x_n)) \rangle$, $T_n^x := \langle x_1, \ldots, x_n \rangle$, $T_n^y := \langle f(x_1), \ldots, f(x_n) \rangle$ and $A : T_n \mapsto x_{n+1} \in \mathcal{X} \setminus T_n^x$.

equally well and equally poorly

No learning is possible without some prior knowledge!

No-Free-Lunch for Pure Inductive Logic

- No learning is possible for $c^\dagger$. It seems possible to learn with c^* and c_λ. Unfortunately, No-Free-Lunch Theorem!
- Take $\mathcal{X} := \mathcal{C}, \mathcal{Y} := \mathcal{Q}$ in the No-Free-Lunch Theorem. The state description $h : \mathcal{C} \to \mathcal{Q}$ can be taken as a *classification* function.
- Then "all state descriptions $h : \mathcal{C} \to \mathcal{Q}$ with the same structure description $\{n_i : 1 \leq i \leq r\}$ have equal weight" is block uniform!
- Define induction algorithm $A(h_{1:n}) := \operatorname*{argmax}_j c^* \left(Q_j(a_{n+1}) | h_{1:n}\right)$, and loss function $L\left(A, n, h\right) := [\![A(h_{1:n}) \neq h_{n+1}]\!]$. Then for any A',

$$\sum_{h \in \mathcal{Y}^{\mathcal{X}}} m^*(h_{1:n}) L\left(A, n, h\right) = \sum_{h \in \mathcal{Y}^{\mathcal{X}}} m^*(h_{1:n}) L\left(A', n, h\right)$$

- Similarly for rank description (E), which is related to Good-Turing estimate. And similarly for Ristad's methods $(C)(D)$.

Time Series and Solomonoff Induction

- However, if we take time into consideration, define

$$M(h_{1:n}) := \sum_{p:U(p)=h_{1:n}*} 2^{-\ell(p)}$$

then we can get free lunch with M', since M' biases non-random state descriptions and is not block uniform. where

$$M'(\epsilon) := 1$$

$$M'(h_{1:n}) := M'(h_{<n}) \frac{M(h_{1:n})}{\sum_{Q \in \mathcal{Q}} M(h_{<n}Q)}$$

and we can extend M' from state descriptions to sentences.

- Besides, by c^* the probability of "the sun will rise tomorrow" is $\frac{n+1}{n+2}$, but c^* fails to confirm "the sun will always rise", while M' can confirm it.

$$\lim_{n\to\infty} M'(\text{the sun will always rise}|\text{the sun rises in the first } n \text{ days}) = 1$$

PAC (Probably Approximately Correct) Learning

Definition (PAC-Learnability)

A hypothesis space $\mathcal{H} \subset 2^{\mathcal{X}}$ is PAC-learnable iff there exists a sample complexity function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm A with the following property:

- for every $\varepsilon, \delta \in (0,1)$
- for every distribution $\mathcal{D}$ over $\mathcal{X}$, and for every labeling function $f : \mathcal{X} \to \{0,1\}$

when running A on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ i.i.d. training samples S generated by $\mathcal{D}$ and labeled by f, the algorithm A returns a hypothesis $A(S) \in \mathcal{H}$ s.t.

$$P_{S \sim \mathcal{D}^m}\Big(P_{x \sim \mathcal{D}}\big(A(S)(x) \neq f(x)\big) > \varepsilon\Big) < \delta$$

$$\text{No-Free-Lunch} \implies m_{2^{\mathcal{X}}}(\varepsilon, \delta) = \infty$$

Agnostic PAC Learning

Definition (Agnostic PAC-Learnability)

A hypothesis space $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ is agnostic PAC-learnable under a class Δ of distributions with respect to $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ and a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}^+$, iff there exists a sample complexity function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm A with the following property:

- for every $\varepsilon, \delta \in (0,1)$
- for every distribution $\mathcal{D} \in \Delta$ over $\mathcal{Z}$

when running A on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ i.i.d. training samples S generated by $\mathcal{D}$, the algorithm A returns a hypothesis $A(S) \in \mathcal{H}$ s.t.

$$P_{S \sim \mathcal{D}^m}\left(L_{\mathcal{D}}(A(S)) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) > \varepsilon\right) < \delta$$

where $L_{\mathcal{D}}(h) := \mathbb{E}_{z \sim \mathcal{D}}\left[\ell(h, z)\right]$.

PAC-learnable under $\Delta := \{\mathcal{D} : \mathcal{D} \overset{\times}{\lesssim} \xi\} \iff$ PAC-learnable under $\xi(x) := \sum_{\nu \in \mathcal{M}} 2^{-K(\nu)} \nu(x)$.

VC-Dimension

Definition (Shattering)

A hypothesis space $\mathcal{H} \subset 2^{\mathcal{X}}$ shatters a set $C \subset \mathcal{X}$ iff $\mathcal{H}\restriction_C = 2^C$.

Definition (VC-Dimension)

$\mathrm{VC}(\mathcal{H}) := \sup\left\{|C| : \mathcal{H} \text{ shatters } C\right\}$.

If someone can explain every phenomena, her explanations are worthless.

Theorem

If $\mathrm{VC}(\mathcal{H}) = \infty$, then $\mathcal{H}$ is not PAC-learnable.

Theorem (Fundamental Theorem of PAC Learning)

$\mathcal{H}$ is PAC-learnable iff $\mathrm{VC}(\mathcal{H}) < \infty$. Indeed, there exists C_1, C_2 s.t.

$$C_1 \frac{\mathrm{VC}(\mathcal{H}) + \log(1/\delta)}{\varepsilon} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \frac{\mathrm{VC}(\mathcal{H}) \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}$$

Contents

History of AI

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Reinforcement Learning

General Reinforcement Learning

Formal Learning Theory

- Carnap's inductive logic is a design for a 'learning machine' that can extrapolate certain kinds of empirical regularities from the data with which it is supplied, and the task of inductive logic is to construct a 'universal learning machine'.
- If there is such a thing as a correct definition of 'degree of confirmation' which can be fixed once and for all, then a machine that predicts in accordance with it would be a cleverest possible learning machine.
- Either there are better and better 'degree of confirmation' functions, but no 'best possible', or there is a 'best possible' but it is not computable by a machine.

Formal Learning Theory

- Putnam 1963
- Gold 1967 $\hat{f}(n+1) = g(\langle f(0), \ldots, f(n)\rangle)$ $\varphi_{\lim_{n\to\infty} g(\langle f(0),\ldots,f(n))} = f$
- Solomonoff 1960

Ray Solomonoff

Figure: Solomonoff

- 0101010101010101010101010101010101
- 0101001101111011011101010100000 1
- 0010010000111111011010101000 1000 $\{\pi\}$

1. What is regularity/pattern/law/principle/model/hypothesis/theory?
2. What is phenomenon/data/experience?
3. What is randomness/noise?
4. What is typicalness/unpredictability/incompressibility?
5. What is simplicity/complexity?
6. What is learning?
7. What is beauty/interesting/curiosity/novelty/surprise/creativity?
8. What is intelligence?

Text $\Longrightarrow$ Meaning?

numeral $\xrightarrow{\text{short algorithm}}$ value

汝有田舍翁，家资殷盛，而累世不识“之”“乎”。一岁，聘楚士训其子。楚士始训之搦管临朱，书一画，训曰：“一”字。书二画，训曰：“二”字。书三画，训曰：“三”字。其子辄欣欣然掷笔，归告其父曰：“儿得矣！儿得矣！可无烦先生，重费馆谷也，请谢去。”其父喜从之，具币谢遣楚士。逾时，其父拟征召姻友万氏者饮，令子晨起治状，久之不成。父趣之。其子恚曰：“天下姓字多矣，奈何姓万？自晨起至今，才完五百画也。”

Kolmogorov Complexity

Definition (Kolmogorov Complexity)

$$K(x) := \min_p \{\ell(p) : U(p) = x\}$$

where U is a universal prefix Turing machine.

"Independence" of universal Turing machine!

How to quantify "simplicity"? "randomness"?

Figure: Kolmogorov

Kraft-Chaitin Theorem

Theorem (Kraft Inequality)

1. *Let* $f : \mathcal{X} \to 2^{<\omega}$ *be uniquely decodable. Let* $\ell_x := \ell(f(x))$*. Then* f *satisfies the Kraft inequality*

$$\sum_{x \in \mathcal{X}} 2^{-\ell_x} \leq 1$$

2. *Conversely, for any set of code length* $\{\ell_x : x \in \mathcal{X}\}$ *satisfying the above Kraft inequality, there exists a prefix code* f *such that* $\ell_x = \ell(f(x))$*.*

Theorem (Kraft-Chaitin Theorem)

For any c.e. set $(\ell_i, x_i)_{i \in \omega} \subset \mathbb{N} \times 2^{<\omega}$ *with* $\sum_{i \in \omega} 2^{-\ell_i} \leq 1$*, one can effectively obtain a prefix-free machine* M *and strings* p_i *of length* ℓ_i *such that* $M(p_i) = x_i$ *for all* i *and* $\mathrm{dom}(M) = \{p_i : i \in \omega\}$*.*

Properties

1. $K(n) \stackrel{+}{\leq} \log^* n \leq \log n + 2\log\log n$
2. $K(x) \stackrel{+}{\leq} K(x|\ell(x)) + K(\ell(x)) \stackrel{+}{\leq} \ell(x) + \log^* \ell(x) \leq \ell(x) + 2\log \ell(x)$
3. $\sum_x 2^{-K(x)} \leq 1$
4. $K(x|y) \stackrel{+}{\leq} K(x) \stackrel{+}{\leq} K(x,y)$
5. $K(xy) \stackrel{+}{\leq} K(x,y) \stackrel{+}{\leq} K(x) + K(x|y) \stackrel{+}{\leq} K(x) + K(y)$
6. $K(x) \stackrel{+}{=} K(x, K(x))$
7. $K(x|y, K(y)) + K(y) \stackrel{+}{=} K(x,y) \stackrel{+}{=} K(y,x) \stackrel{+}{=} K(y|x, K(x)) + K(x)$
8. $K(f(x)) \stackrel{+}{\leq} K(x) + K(f)$ for computable f
9. $K(x) \stackrel{+}{\leq} -\log \mu(x) + K(\mu)$ if μ is lower semicomputable and $\sum_x \mu(x) \leq 1$
10. $\sum_{x:f(x)=y} 2^{-K(x)} \stackrel{\times}{=} 2^{-K(y)}$ if f is computable and $K(f) = O(1)$
11. $0 \leq \mathbb{E}_\mu[K] - H(\mu) \stackrel{+}{\leq} K(\mu)$ for computable probability distribution μ

Universal Similarity Metric

$$d(x,y) := \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}}$$
$$\approx \frac{K_T(xy) - \min\{K_T(x), K_T(y)\}}{\max\{K_T(x), K_T(y)\}}$$

- T : Lempel-Ziv/gzip/bzip2/PPMZ, or $K_T(x) := -\log P_{google}(x)$
- compute similarity matrix $\big(d(x_i, x_j)\big)_{ij}$
- cluster similar objects.

Algorithmic Probability

Definition (Algorithmic Probability)

$$M(x) := \sum_{p:U(p)=x*} 2^{-\ell(p)}$$

where U is a universal monotone Turing machine.

$$\sum_{p:U(p)=x*} 2^{-\ell(p)} \gg 2^{-\ell(x)}$$

Toss Coin onto Universal Turing Machine

Bolztmann Brain vs Brain in a Vat

$K(\nu)$ is too large.

Aspect 1 — Popper

$\mathcal{H}$: truth $\leftarrow$ simplcity/generality/aesthetic/utilitarian/...

Make a weighted prediction based on all consistent programs, with short programs weighted higher.

Aspect 2 — Deterministic vs Stochastic

$\mathcal{M} := \{\nu_1, \nu_2, \dots\}$ lower semicomputable semi-measure.

$$\xi(x) := \sum_{\nu \in \mathcal{M}} 2^{-K(\nu)} \nu(x)$$

$$M(x) \stackrel{\times}{=} \xi(x)$$

$\boxed{w_\nu := 2^{-K(\nu)} \text{ is reparametrization \& regrouping invariant.}}$

$$\tilde{w}_{\theta'} = w_{f^{-1}(\theta')} = 2^{-K(f^{-1}(\theta'))} \stackrel{\times}{=} 2^{-K(\theta')} \stackrel{\times}{=} w'_{\theta'}$$

$$\tilde{w}_{\theta'} = \sum_{\theta : f(\theta) = \theta'} 2^{-K(\theta)} \stackrel{\times}{=} 2^{-K(\theta')} \stackrel{\times}{=} w'_{\theta'}$$

Aspect 3 — Frequency Interpretation

$$
\begin{aligned}
M(x) &= \sum_p 2^{-\ell(p)} [\![U(p) = x*]\!] \\
&= \lim_{n\to\infty} \frac{\sum_{p:\ell(p)\le n} 2^{n-\ell(p)} [\![U(p) = x*]\!]}{2^n} \\
&\approx \lim_{n\to\infty} \frac{\left|\left\{p : \ell(p) = n \;\&\; U(p) = x*\right\}\right|}{2^n}
\end{aligned}
$$

$$
\textbf{algorithmic probability} = \frac{\left|\text{consistent worlds}\right|}{\left|\text{all possible worlds}\right|}
$$

$$
\begin{cases}
\text{Carnap — frequency of phenomena — i.i.d} \\
\text{Solomonoff — frequency of causes — arbitrary order Markov chain}
\end{cases}
$$

Aspect 4 — Free-Lunch for Solomonoff!

$$\xi \to \mu$$

break "block uniform" — bias non-random functions

Algorithmic Coding Theorem

$$m(x) := \sum_{p:U(p)\downarrow=x} 2^{-\ell(p)}$$

Theorem (Algorithmic Coding Theorem)

$$K(x) \stackrel{+}{=} -\log m(x)$$

The probability of a string being produced by a random algorithm is inversely proportional to its algorithmic complexity.
If a string has many long descriptions then it also has a short description.

$$KM(x) := -\log M(x)$$

$$KM(x) \leq Km(x) \stackrel{+}{\leq} KM(x) + K(KM(x))$$

Completeness Theorem

$$M'(\epsilon) := 1$$

$$M'(x_{1:t}) := M'(x_{<t}) \frac{M(x_{1:t})}{\sum_{x \in \mathcal{X}} M(x_{<t}x)} = \frac{M(x_{1:t})}{M(\epsilon)} \prod_{i=1}^{t} \frac{M(x_{<i})}{\sum_{x \in \mathcal{X}} M(x_{<i}x)}$$

Theorem (Completeness Theorem)

For any computable measure μ,

$$\sum_{t=1}^{\infty} \sum_{x_{1:t} \in \mathcal{X}^t} \mu(x_{<t}) \Big(M'(x_t | x_{<t}) - \mu(x_t | x_{<t}) \Big)^2 \leq D(\mu \| M) \stackrel{+}{\leq} K(\mu) \ln 2$$

Emergence of Simple Laws of Physics

Completeness Theorem

For any computable measure μ, there is a set $A \subset \mathcal{X}^*$ with $\mu(A) = 1$ s.t. for all $\mathbf{x} \in A$

$$\sum_{y \in \mathcal{X}} \left(\sqrt{M'(y|\mathbf{x})} - \sqrt{\mu(y|\mathbf{x})} \right)^2 \xrightarrow{n \to \infty} 0$$

Emergence of Simple Laws of Physics

For any computable measure μ,

$$M' \left\{ \sum_{y \in \mathcal{X}} \left(\sqrt{M'(y|\mathbf{x})} - \sqrt{\mu(y|\mathbf{x})} \right)^2 \xrightarrow{n \to \infty} 0 \right\} \geq 2^{-K(\mu)}$$

Universal Prediction of Selected Bits

Theorem (Universal Prediction of Selected Bits)

Let $f : \{0,1\}^* \to \{0,1,\epsilon\}$ *be a total recursive function and* $x \in 2^\omega$ *satisfying* $f(x_{<n}) = x_n$ *whenever* $f(x_{<n}) \neq \epsilon$. *If* $f(x_{<n_i}) \neq \epsilon$ *for an infinite sequence* $n_1, n_2, \ldots$ *then*

$$\lim_{i\to\infty} M'(x_{n_i}|x_{<n_i}) = 1$$

Pure Universal Inductive Logic?

- $M(\Theta(a_{1:n})) := \sum_{p:U(p)=h_{1:n}*} 2^{-\ell(p)}$

  where $\Theta(a_{1:n}) := \bigwedge_{i=1}^{n} Q_{h_i}(a_i)$.

- $M'(A(\vec{a})) := \sum_{\Theta(\vec{b})\vDash A(\vec{a})} M'(\Theta(\vec{b}))$

  where $\vDash A(\vec{a}) \leftrightarrow \bigvee_{\Theta(\vec{b})\vDash A(\vec{a})} \Theta(\vec{b})$. FDNF

$$\boxed{\sum_{t=1}^{\infty} \sum_{A(a_{1:t})} \mu(A(a_{<t}))\Big(M'(A(a_t)\,|\,A(a_{<t})) - \mu(A(a_t)\,|\,A(a_{<t}))\Big)^2 \stackrel{+}{\leq} K(\mu)\ln 2}$$

where $A(a_{1:t}) := \bigwedge_{i=1}^{t} A(a_i/x)$.

All Ravens are Black! ✓

Theorem (All Ravens are Black)

$$\lim_{n\to\infty} M'\left(\forall x(R(x) \to B(x)) \,\middle|\, \bigwedge_{i=1}^{n} (\neg R(a_i) \vee B(a_i))\right) = 1$$

Theorem (Confirmation by Random Sampling)

If the sampling function $t : \mathbb{N} \to \mathbb{N}$ satisfies $\forall i : t_i \leq t_{i+1}$ and $\chi_{1:\infty}$ is Martin-Löf random, where $\chi_i := [\![\exists k(t_k = i)]\!]$, then

$$M'\left(\forall x A(x) \,\middle|\, \bigwedge_{i=1}^{n} A(a_{t_i})\right) \xrightarrow{n\to\infty} 1$$

$$M(1|1^n) \xrightarrow{n\to\infty} 1 \qquad M(0|1^n) \stackrel{\times}{=} 2^{-K(n)} \qquad \sum_{n=0}^{\infty} M(0|1^n) < \infty$$

Why Solomonoff Prior?

$$H(w) \leq \mathbb{E}_w[K] \leq H(w) + K(w)$$

Maximum Entropy + Occam's Razor

$$\underset{w \vDash \sum_{\nu \in \mathcal{M}} w_\nu = 1}{minimize} \frac{\mathbb{E}_w[K]}{H(w)}$$

$$\Downarrow$$

$$w^*_\nu = \frac{2^{-K(\nu)}}{\sum_{\nu \in \mathcal{M}} 2^{-K(\nu)}}$$

Advantages & Disadvantages

- free-lunch
- universality — finite error
- data sparse problem — arbitrary order Markov chain — universal smoothing method
- confirmation of $\forall x : R(x) \to B(x)$
- incomputability
- weakly depends on universal Turing machine

A statistical mechanical interpretation of AIT

an energy eigenstate n	$\Longrightarrow$	a program p s.t. $U(p)\downarrow$
the energy E_n of n	$\Longrightarrow$	the length $\ell(p)$ of p
Boltzmann constant k	$\Longrightarrow$	$1/\ln 2$

$Z = \sum_n e^{-\frac{E_n}{kT}}$	$\Longrightarrow$	$Z = \sum_{p:U(p)\downarrow} 2^{-\frac{\ell(p)}{T}}$	Partition function
$F = -kT \ln Z$	$\Longrightarrow$	$F = -T \log Z$	Free energy
$P(n) = \frac{1}{Z} e^{-\frac{E_n}{kT}}$	$\Longrightarrow$	$P(p) = \frac{1}{Z} 2^{-\frac{\ell(p)}{T}}$	Boltzmann distribution
$E = \sum_n P(n) E_n$	$\Longrightarrow$	$E = \sum_{p:U(p)\downarrow} P(p)\ell(p)$	Energy
$S = \frac{E-F}{T}$	$\Longrightarrow$	$S = \frac{E-F}{T} = H(P)$	Entropy

$$T \geq 1 \Longrightarrow H(P) = \infty \qquad 0 < T < 1 \Longrightarrow H(P) < \infty$$

$$\boxed{\text{Temperature=Compression Rate}}$$

$$T = \lim_{n\to\infty} \frac{K(Z_{1:n})}{n} = \lim_{n\to\infty} \frac{K(F_{1:n})}{n} = \lim_{n\to\infty} \frac{K(E_{1:n})}{n} = \lim_{n\to\infty} \frac{K(S_{1:n})}{n}$$

Fixpoint theorem: for $T \in (0,1)$, if Z is computable, then $\lim_{n\to\infty} \frac{K(T_{1:n})}{n} = T$.

Computable Universal Predicter

$$Z_h = \sum_{p:U(p)=h*} 2^{-\frac{\ell(p)}{T}}$$

$$Z(h) = \sum_{p:U(p)=h*} 2^{-\frac{\ell(p)+\log t(p,h)}{T}}$$

$$\boxed{T \text{ is computable} \implies Z(h) \text{ is computable}}$$

$$\sum_{t=1}^{\infty} |1 - Z(h_t|h_{<t})| \leq \frac{Km(h)\ln 2 + \ln t(p,h)}{T}$$

Stochastic Case

$$E_{\{h,\nu\}} = -\log w_\epsilon^\nu - \log \nu(h)$$

$$Z_h = \sum_{\nu \in \mathcal{M}} 2^{-E_{\{h,\nu\}}} = \sum_{\nu \in \mathcal{M}} w_\epsilon^\nu \nu(h) = \xi(h)$$

$$P_h(\nu) = \frac{2^{-E_{\{h,\nu\}}}}{Z_h} = \frac{w_\epsilon^\nu \nu(h)}{\xi(h)} = w_h^\nu$$

$$F_h = -\log Z_h = \underbrace{-\log \xi(h)}_{\approx K(h)} = \underbrace{\mathbb{E}_{w_h}[-\log \nu(h)]}_{noise} + \underbrace{D(w_h \| w_\epsilon)}_{surprise}$$

Deduction vs Induction

	Induction		**Deduction**
Type of inference	generalization/prediction	$\Leftrightarrow$	specialization/derivation
Framework	probability axioms	$\widehat{=}$	logical axioms
Assumptions	prior	$\widehat{=}$	non-logical axioms
Inference rule	Bayes rule	$\widehat{=}$	modus ponens
Results	posterior	$\widehat{=}$	theorems
Universal scheme	Solomonoff probability	$\widehat{=}$	ZFC
Universal inference	universal induction	$\widehat{=}$	universal theorem prover
Limitation	uncomputable (Turing)	$\widehat{=}$	imcomplete (Gödel)
In practice	approximations	$\widehat{=}$	semi-formal proofs
Operation	computation	$\widehat{=}$	proof

Logic vs Statistics

Field	**Logical Approach**	**Statistical Approach**
Knowledge representation	First order logic	Graphical models
Automated reasoning	Satisfiability testing	Markov chain Monte Carlo
Machine learning	Inductive logic programming	Neural networks
Planning	Classical planning	Markov decision processes
Natural language processing	Definite clause grammars	Probabilistic context-free grammars

Logic	Statistics
rule-based	data-driven
rigour	possibility
knowable	black-box
simple & perfect world	complex & uncertain world
✗	✓

Prediction with Expert Advice

Assume that there is some large, possibly infinite, class of 'experts' which make predictions. The aim is to observe how each of these experts perform and develop independent predictions based on this performance.

- Follow the (perturbed) leader.
- Predicts according to a majority vote by the "good" experts.
- Multiplicative Weights. — take expert which performed best in past with high probability and others with smaller probability.
- Regularization. Choose the class of all computable experts, and penalize "complex" experts.
- Universal Portfolios.

Universal Portfolios

- the agent chooses a distribution $\mathbf{b}_t \in \Delta_n := \{\mathbf{x} \in [0,1]^n : \|\mathbf{x}\|_1 = 1\}$ of wealth over n goods.
- nature chooses returns $\mathbf{x}_t \in (\mathbb{R}^+)^n$, where

$$(\mathbf{x}_t)_i = \frac{\text{price of good } i \text{ at end of } t}{\text{price of good } i \text{ at beginning of } t}$$

- the total wealth. $W_t(\mathbf{b}, \mathbf{x}) = W_1 \prod_{k=1}^{t} \mathbf{b}_k^\top(\mathbf{x}_{<k})\mathbf{x}_k$
- regret. $R_t := \max_{\mathbf{b} \in \Delta_n} \sum_{k=1}^{t} \log \mathbf{b}_k^\top(\mathbf{x}_{<k})\mathbf{x}_k - \sum_{k=1}^{t} \log \mathbf{b}_k^\top(\mathbf{x}_{<k})\mathbf{x}_k$
- universal portfolios.

$$\hat{\mathbf{b}}_1 := \left(\tfrac{1}{n}, \ldots, \tfrac{1}{n}\right)$$

$$\hat{\mathbf{b}}_{t+1}(\mathbf{x}_{1:t}) := \frac{\int_{\Delta_n} \mathbf{b} W_t(\mathbf{b}, \mathbf{x}) \mathrm{d}\mathbf{b}}{\int_{\Delta_n} W_t(\mathbf{b}, \mathbf{x}) \mathrm{d}\mathbf{b}}$$

- Asymptotic Optimality. $\frac{1}{t} \log W_t(\hat{\mathbf{b}}, \mathbf{x}) \xrightarrow{t \to \infty} \frac{1}{t} \log \max_{\mathbf{b} \in \Delta_n} W_t(\mathbf{b}, \mathbf{x})$

Randomness

Typicalness The statistician's approach: A random sequence is the typical outcome of a random variable. Random sequences should not have effectively rare distinguishing properties.

Incompressibility The coder's approach: Rare patterns can be used to compress information. Random sequences should not be effectively described by a significantly shorter description than their literal representation.

Unpredictability The gambler's approach: A betting strategy can exploit rare patterns. Random sequences should be unpyellowictable. No effective martingale can make an infinite amount betting on the bits.

The Statistician's Approach

- A random sequence should be absolutely normal.
- If you select a subsequence, then it should satisfy the law of large numbers, the law of the iterated logarithm...
- But what selection functions should be allowed? Computable?
- Martin-Löf: we can effectively test whether a particular infinite sequence does not satisfy a particular law of randomness by effectively testing whether the law is violated on increasingly long initial segments. We should consider the intersection of all sets of measure one with recursively enumerable complements. (Such a complement set is expressed as the union of a recursively enumerable set of cylinders).

Cantor Space 2^ω

- For $x \in 2^{<\omega}$, the cylinder set $\Gamma_x := \{y \in 2^\omega : x \prec y\}$ is the basic open set. It corresponds to the interval $[0.x, 0.x + 2^{-\ell(x)})$.
- For $A \subset 2^{<\omega}$, the open set generated by A is $\Gamma_A := \bigcup_{x \in A} \Gamma_x$.
- The Lebesgue measure $\mu(\Gamma_x) := 2^{-\ell(x)}$, $\mu(x) := \mu(\Gamma_x)$.
- The outer measure of $C \subset 2^\omega$ is $\mu^*(C) := \inf \left\{ \sum_{x \in A} 2^{-\ell(x)} : C \subset \Gamma_A \right\}$.
- The inner measure of C is $\mu_*(C) := 1 - \mu^*(2^\omega \setminus C)$.
- If C is measurable, then $\mu^*(C) = \mu_*(C)$.
- $A \subset 2^\omega$ has measure 0 iff there is a sequence $\{V_n\}_{n \in \omega}$ of open sets s.t. $A \subset \bigcap_{n \in \omega} V_n$ and $\lim_{n \to \infty} \mu(V_n) = 0$.

Martin-Löf Randomness

Definition (Martin-Löf Randomness)

- A total lower semicomputable function $\delta : 2^{<\omega} \to \omega$ is a Martin-Löf test iff $\forall n : \mu(V_n) \leq 2^{-n}$, where $V_n := \{x : \delta(x) \geq n\}$.
- $x \in 2^\omega$ is ML-random iff for every ML-test δ, $\sup_n \delta(x_{1:n}) < \infty$.

$$\delta(x) < \infty \iff x \notin \bigcap_{n=1}^{\infty} V_n$$

Definition (Martin-Löf Randomness)

- A Martin-Löf test is a uniformly c.e. (i.e., $\left\{\langle n, x\rangle : x \in V_n\right\}$ is c.e.) sequence of open sets $\{V_n\}_{n\in\omega}$ s.t. $\forall n : \mu(V_n) \leq 2^{-n}$.
- $x \in 2^\omega$ is ML-random iff for every ML-test $\{V_n\}_{n\in\omega}$, $x \notin \bigcap_{n=1}^{\infty} V_n$.

Martin-Löf Randomness

Definition (Universal Martin-Löf Test)

A ML-test δ_0 is *universal* iff for every ML-test δ, $\exists c \forall x : \delta_0(x) \geq \delta(x) - c$.

A ML-test $\{U_n\}_{n\in\omega}$ is *universal* iff for every ML-test $\{V_n\}_{n\in\omega}$, $\bigcap_{n\in\omega} U_n \supset \bigcap_{n\in\omega} V_n$.

$$\boxed{\delta(x) := \ell(x) - K(x|\ell(x)) \text{ is a universal ML-test.}}$$

$$R_b := \left\{x \in 2^\omega : \exists n\big(K(x_{1:n}) < n - b\big)\right\}$$

$$\boxed{\{R_b\}_{b\in\omega} \text{ is a universal ML-test.}}$$

Theorem (Schnorr 1973)

A sequence $x \in 2^\omega$ is ML-random iff it is 1-random.

The Gambler's Approach

- A martingale is a function $d : 2^{<\omega} \to [0, \infty)$ s.t. for every $\sigma \in 2^{<\omega}$

$$d(\sigma) = \frac{d(\sigma 0) + d(\sigma 1)}{2}$$

- A supermartingale is a function $d : 2^{<\omega} \to [0, \infty)$ s.t. for every $\sigma \in 2^{<\omega}$

$$d(\sigma) \geq \frac{d(\sigma 0) + d(\sigma 1)}{2}$$

- A (super)martingale d succeeds on $x \in 2^{\omega}$ iff $\limsup_{n \to \infty} d(x_{1:n}) = \infty$.

Theorem

A sequence $x \in 2^{\omega}$ is ML-random iff no c.e. (super)martingale succeeds on it.

The Coder's Approach

Definition (1-Randomness)

$x \in 2^\omega$ is 1-random iff

$$\exists c \forall n : K(x_{1:n}) \geq n - c$$

Theorem

The following are equivalent.

- $x \in 2^\omega$ *is ML-random.*
- *No c.e. (super)martingale succeeds on it.*
- $\exists c \forall n : K(x_{1:n}) \geq n - c$
- $\forall n : Km(x_{1:n}) \stackrel{+}{=} n$
- $\lim_{n \to \infty} K(x_{1:n}) - n = \infty$
- $\sum_{n=1}^{\infty} 2^{n - K(x_{1:n})} < \infty$
- $\sup_n 2^{n - K(x_{1:n})} < \infty$
- $C(x_{1:n}) \stackrel{+}{\geq} n - K(n)$
- $C(x_{1:n}) \stackrel{+}{\geq} n - f(n)$ *for every computable* f *s.t.* $\sum_{n=1}^{\infty} 2^{-f(n)} < \infty$.

Definition (Solovay Reducibility)

Let $a_n \to \alpha$ and $b_n \to \beta$ be two computable strictly increasing sequences of rationals converging to lower semicomputable reals α and β. We say that $\alpha \leq_S \beta$ iff there is a constant c and a total computable function f s.t. $\forall n : \alpha - a_{f(n)} \leq c(\beta - b_n)$.

Theorem

For lower semicomputable reals α, the following are equivalent.

- α *is* 1*-random*
- $\alpha \geq_S \beta$ *for all lower semicomputable reals* β.
- $\alpha \geq_S \Omega$
- $K(\alpha_{1:n}) \stackrel{+}{\geq} K(\beta_{1:n})$ *for all lower simicomputable reals* β.
- $K(\alpha_{1:n}) \stackrel{+}{\geq} K(\Omega_{1:n})$
- $K(\alpha_{1:n}) \stackrel{+}{=} K(\Omega_{1:n})$
- $\alpha = \Omega_U := \sum_{p:U(p)\downarrow} 2^{-\ell(p)}$ *for some universal prefix Turing machine* U.

μ/ξ-randomness

Definition (μ/ξ-randomness)

- A sequence $x \in 2^\omega$ is μ/ξ-random iff $\exists c \forall n : \xi(x_{1:n}) \leq c \cdot \mu(x_{1:n})$.
- A sequence $x \in 2^\omega$ is μ-ML-random iff $\exists c \forall n : M(x_{1:n}) \leq c \cdot \mu(x_{1:n})$.

- $x_{1:\infty}$ is μ-ML-random iff $\sup_n \delta(x_{1:n}|\mu) < \infty$, where

$$\delta(x|\mu) := \log \frac{M(x)}{\mu(x)}$$

- For a computable μ, $x_{1:\infty}$ is μ-ML-random iff

$$\forall n : Km(x_{1:n}) \stackrel{+}{=} -\log \mu(x_{1:n})$$

Randomness, Triviality, Logical Depth

- $A \subset \mathbb{N}$ is *low* iff $A' \leq_T \emptyset'$, and A is *high* iff $\emptyset'' \leq_T A'$.
- A is *low for ML-randomness* iff each ML-random set is already ML-random relative to A.
- A is *low for K* iff $\exists c \forall x : K(x) \leq K^A(x) + c$.
- $x \in 2^\omega$ is *K-trivial* iff $\exists c \forall n : K(x_{1:n}) \leq K(n) + c$.

Theorem

A is K-trivial $\iff$ A is low for ML-randomness $\iff$ A is low for K.

Some sequences are K-trivial but not computable.
Neither randoms, nor K-trivials, are deep.

Definition (Logical Depth)

The logical depth of x at a significance level b is

$$\mathrm{depth_b}(x) := \min\left\{t : U^t(p) = x \ \ \& \ \ \ell(p) - K(x) \leq b\right\}$$

x is called *shallow* iff $\mathrm{depth_b}(x) \overset{+}{\leq} \ell(x)$.
Let $\chi_i := [\![\varphi_i(i)\downarrow]\!]$. $\chi_{1:\infty}$ is deep. Ω is shallow.

Effective Complexity

$$\delta(x|A) := \log|A| - K(x|A)$$ [randomness deficiency]

$$\delta(x|\mu) := \log\frac{M(x)}{\mu(x)}$$ [μ-randomness deficiency]

$$\beta_x(k) := \min_A\{\delta(x|A) : x \in A \ \ \& \ \ K(A) \le k\}$$ [Best-Fit]

$$h_x(k) := \min_A\{\log|A| : x \in A \ \ \& \ \ K(A) \le k\}$$ [Kolmogorov structure function / ML]

$$h_x(k) := \min_\mu\{-\log\mu(x) : K(\mu) \le k\}$$ [[Kolmogorov structure function / ML]

$$A^*(x) := \iota A\Big[x \in A \ \ \& \ \ K(A) = \mu k\big[k + h_x(k) \stackrel{+}{=} K(x)\big]\Big]$$ [Kolmogorov minimal sufficient statistic]

$$\lambda_x(k) := \min_A\{K(A) + \log|A| : x \in A \ \ \& \ \ K(A) \le k\}$$ [MDL]

$$\lambda_x(k) := \min_\mu\{K(\mu) - \log\mu(x) : K(\mu) \le k\}$$ [MDL]

$$\Delta(x|A) := K(A) + \log|A| - K(x)$$ [discrepancy]

$$\mathrm{soph_c}(x) := \min_A\{K(A) : \Delta(x|A) < c\}$$ [sophistication]

$$\mathrm{csoph}(x) := \min_A\{K(A) + \Delta(x|A)\}$$ [coarse sophistication]

$$\Sigma(\mu) := K(\mu) + H(\mu)$$ [total information]

$$\mathcal{E}_{\delta,\Delta}(x|\mathcal{M}) := \min_{\mu\in\mathcal{M}}\Big\{K(\mu) : \Sigma(\mu) - K(x) \le \Delta \ \ \& \ \ \mu(x) \ge 2^{-H(\mu)(1+\delta)}\Big\}$$ [effective complexity]

$$\mathcal{E}_\delta(x|\mathcal{M}) := \min_{\mu\in\mathcal{M}}\Big\{K(\mu) + \Sigma(\mu) - K(x) : \mu(x) \ge 2^{-H(\mu)(1+\delta)}\Big\}$$ [coarse effective complexity]

K(x)
hx(k)
log |A|
K(x) - k
K(K(x))
K(x)
minimal sufficient statistic
k

ℓ (x)
λx(k)
K(x)
log |A|
hx(k)
βx(k)
minimal sufficient statistic
K(x)
k

Theorem

For x and k,

$$\lambda_x(k) \leq h_x(k) + k \overset{+}{\leq} \lambda_x(k) + K(k)$$

For k with $0 \leq k \leq K(x) - O(\log \ell(x))$,

$$\beta_x(k) + K(x) \overset{+}{\leq} \lambda_x(k)$$

$$\lambda_x(k + O(\log \ell(x))) \leq \beta_x(k) + K(x)$$

In other words, the equality

$$\beta_x(k) + K(x) = \lambda_x(k) = h_x(k) + k$$

holds within logarithmic additive terms in argument and value.

Sophistication and Computational Depth

Theorem

$$\mathrm{csoph}(x) = \min_c \{\mathrm{soph_c}(x) + c\}$$

$$K^t(x) := \min_p \{\ell(p) : U^t(p) = x\} \quad \text{(Time-bounded Kolmogorov Complexity)}$$

$$\mathrm{depth^t}(x) := K^t(x) - K(x) \quad \text{(Basic Computational Depth)}$$

$$\mathrm{depth_{BB}}(x) := \min_t \{\mathrm{depth^t}(x) + K(t)\} \quad \text{(Busy Beaver Computational Depth)}$$

Theorem

$$|\mathrm{csoph}(x) - \mathrm{depth_{BB}}(x)| \leq O(\log \ell(x))$$

Zurek's Physical Entropy

Definition (Physical Entropy)

Physical entropy $S(d)$ of a microstate d is the sum of the conditional Shannon entropy $H_d := -\sum_k P(k|d) \log P(k|d)$ and of the Kolmogorov Complexity $K(d)$.

$$S(d) := H_d + K(d)$$

Figure: random vs regular microstate

Maxwell's Demon & Landauer's Principle

Figure: The demon turns entropy into information, the information-erasure operation turns information into entropy. In the course of ideal measurement on an equilibrium ensemble, the decrease of the entropy must be compensated by the increase of the size of the minimal record, and vice versa. $\Delta H \approx -\langle \Delta K \rangle$.

Figure: Destroying information generates heat

Landauer's Principle: Logically irreversible computation costs energy. Erasing 1 bit of information dissipates at least $kT \ln 2$ of heat into the environment.

Gács' Algorithmic Entropy

Definition (Algorithmic Entropy)

- *coarse-grained algorithmic entropy* of a cell Γ with respect to μ
$$H_\mu(\Gamma) := \log \mu(\Gamma) + K(\Gamma|\mu)$$
- *fine-grained algorithmic entropy* of $x \in 2^\omega$ with respect to μ
$$H_\mu(x) := \inf_n \left\{ \log \mu(x_{1:n}) + K(x_{1:n}|\mu) \right\}$$

- $-H_\mu(x)$ is a universal ML-test: x is μ-ML-random iff $H_\mu(x) > -\infty$.
- The fine-grained algorithmic entropy of a microstate can be approximated by the coarse-grained algorithmic entropies of successively smaller cells containing it.
$$H_\mu(x) = \inf_n \left\{ H_\mu(\Gamma_{x_{1:n}}) \right\}$$

Contents

History of AI

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Reinforcement Learning

General Reinforcement Learning

Simpson Paradox — Should we treat scurvy with lemons?

	Recovery	No Recovery	Total	Recovery Rate
No Lemons	20	20	40	50%
Lemons	16	24	40	40%
Total	36	44	80	

Table: $P(\text{recovery}|\text{lemmon}) < P(\text{recovery}|\text{no lemmon})$

	Recovery	No Recovery	Total	Recovery Rate
No Lemons	2	8	10	20%
Lemons	9	21	30	30%
Total	11	29	40	

Table: $P(\text{recovery}|\text{lemmon}, \text{old}) > P(\text{recovery}|\text{no lemmon}, \text{old})$

	Recovery	No Recovery	Total	Recovery Rate
No Lemons	18	12	30	60%
Lemons	7	3	10	70%
Total	25	15	40	

Table: $P(\text{recovery}|\text{lemmon}, \text{young}) > P(\text{recovery}|\text{no lemmon}, \text{young})$

Resolution of Simpson's paradox — The do-operator

- What is the sailors' probability of recovery when **we see** a treatment with lemons?

$$P(\text{recovery}|\text{lemmon})$$

- What is the sailors' probability of recovery if **we do** treat them with lemons?

$$P(\text{recovery}|\,\mathrm{do}(\text{lemmon}))$$

- We should treat scurvy with lemons if

$$P(\text{recovery}|\,\mathrm{do}(\text{lemmon})) > P(\text{recovery}|\,\mathrm{do}(\text{no lemmon}))$$

$$P(\text{recovery}|\,\mathrm{do}(\text{lemmon})) = \sum_{\text{age}} P(\text{age})P(\text{recovery}|\text{lemmon}, \text{age}) = 0.5$$

$$P(\text{recovery}|\,\mathrm{do}(\text{no lemmon})) = \sum_{\text{age}} P(\text{age})P(\text{recovery}|\text{no lemmon}, \text{age}) = 0.4$$

The Causal Hierarchy

1. Association: "What if I see x?"

$$P(y|x)$$

2. Intervention: "What if I do x?"

$$P(y|\,\mathrm{do}(x))$$

3. Counterfactuals: "What if I had done things differently?"

$$P(y_{x'}|x)$$

4. Options: "With what probability?" — Graphical Representation

Explanation y because of x

Intervention y will be true if I do x

Counterfactuals y would be different if x' were true

Structural Causal Model

Definition

A structural causal model is a 4-tuple (V, U, F, P), where

- $V = \{V_1, \ldots, V_n\}$ are endogenous variables.
- $U = \{U_1, \ldots, U_m\}$ are background variables.
- $F = \{f_1, \ldots, f_n\}$ are functions determining V, $v_i = f_i(\mathrm{pa}_i, u_i)$.
- P is a distribution over U.

$P(u)$ and F induce a distribution $P(v)$ over observable variables.

Example

$Model(M)$

$$C = f_C(U_C)$$
$$S = f_S(C, U_S)$$
$$R = f_R(C, U_R)$$
$$W = f_W(S, R, U_W)$$

- Every missing arrow advertises an independency, conditional on a separating set.

$$C \perp W | (S, R) \qquad S \perp R | C$$

- $P_{S=1}(C, R, W) = P(C)P(R|C)P(W|R, S = 1) \neq P(C, R, W|S = 1)$

Model$(M_{S=1})$

$$C = f_C(U_C)$$
$$S = 1$$
$$R = f_R(C, U_R)$$
$$W = f_W(S, R, U_W)$$

Would the pavement be wet **had** the sprinkler been on?

d-separation

Definition (Blocking of Paths)

A path q is said to be blocked by a set S iff

- q contains a chain $V_i \to V_j \to V_k$ or a fork $V_i \leftarrow V_j \to V_k$ such that the middle node is in S, or
- q contains a collider $V_i \to V_j \leftarrow V_k$ such that the middle node is not in S and no descendant of V_j is in S.

Definition (d-separation)

S is said to d-separate X and Y in the DAG G, i.e., $(X \perp Y|S)_G$ iff S blocks every path from a vertex in X to a vertex in Y.

Example

$$\begin{tikzcd} X \arrow[dr] & & Y \arrow[dl] \\ & Z & \end{tikzcd} \qquad \begin{array}{c} X \perp Y \\ X \not\perp Y|Z \end{array} \qquad \begin{tikzcd} X & & Y \\ & Z \arrow[ul] \arrow[ur] & \end{tikzcd} \qquad \begin{array}{c} X \not\perp Y \\ X \perp Y|Z \end{array}$$

The sentence: "Y would be y (in situation u), had X been x", denoted $Y_x(u) = y$, means:
The solution for Y in a mutilated model M_x, (i.e., the equations for X replaced by $X = x$) with input $U = u$, is equal to y.

- The Law of Counterfactuals

$$Y_x(u) = Y_{M_x}(u)$$

(M generates and evaluates all counterfactuals.)

- The Law of Conditional Independence (d-separation)

$$(X \perp Y|Z)_G \implies (X \perp Y|Z)_P$$

(Separation in the model $\implies$ independence in the distribution.)

The do-operator

- The joint probability distribution

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | \mathrm{pa}_i)$$

- The do-operator

$$P(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n | \mathrm{do}(x_j = b)) = \prod_{\substack{i=1 \\ i \neq j}}^{n} P(x_i | \mathrm{pa}_i)$$

- The post-intervention distribution resulting from the action $\mathrm{do}(X = x)$ is given by

$$P_M(y | \mathrm{do}(x)) = P_{M_x}(y)$$

Perturbed Graphs

- $G_{\overline{X}}$ perturbed graph in which all edges pointing to X have been deleted
- $G_{\underline{X}}$ perturbed graph in which all edges pointing from X have been deleted
- $Z(W)$ set of nodes in Z which are not ancestors of W

Back-Door Criterion

- Any common ancestor of X and Y is a confounder.
- Confounders originate "back-door" paths that need to be blocked by conditioning.

$$\begin{tikzcd} G & U \arrow[dl] \arrow[dr] & \\ X \arrow[rr] & & Y \end{tikzcd} \qquad \begin{tikzcd} G_{\overline{X}} & U \arrow[dr] & \\ X \arrow[rr] & & Y \end{tikzcd}$$

$$P(x, y, u) = P(x|u)P(u)P(y|x, u) \qquad P(y, u|\,\mathrm{do}(x)) = P(u)P(y|x, u)$$

Back-door adjustment:

$$P(y|\,\mathrm{do}(x)) = \sum_u P(u)P(y|x, u)$$

A set of variables Z satisfies *the back-door criterion* relative to an ordered pair of variables (V_i, V_j) in a DAG G if:

1. no node in Z is a descendant of V_i; and
2. Z blocks every path between V_i and V_j that contains an arrow to V_i.

Front-Door Criterion

$P(x, y, z, u) = P(x|u)P(z|x)P(u)P(y|z, u)$ $\quad$ $P(y, z, u|\,\mathrm{do}(z)) = P(z|x)P(u)P(y|z, u)$

Front-door adjustment:

$$P(y|\,\mathrm{do}(x)) = \sum_z P(z|x) \sum_u P(u)P(y|z, u)$$

A set of variables Z satisfies the *front-door criterion* relative to an ordered pair of variables (C_i, V_j) in a DAG G if:

1. Z intercepts all directed paths from V_i to V_j; and
2. there is no unblocked back-door path from V_i to Z; and
3. all back-door paths from Z to V_j are blocked by V_i

do-Calculus

Rule 1 Insertion/deletion of observations

$$P(y|\,\mathrm{do}(x), z, w) = P(y|\,\mathrm{do}(x), w) \text{ if } (Y \perp Z|X, W)_{G_{\overline{X}}}$$

Rule 2 Action/observation exchange

$$P(y|\,\mathrm{do}(x), \mathrm{do}(z), w) = P(y|\,\mathrm{do}(x), z, w) \text{ if } (Y \perp Z|X, W)_{G_{\overline{X},\underline{Z}}}$$

Rule 3 Insertion/deletion of actions

$$P(y|\,\mathrm{do}(x), \mathrm{do}(z), w) = P(y|\,\mathrm{do}(x), w) \text{ if } (Y \perp Z|X, W)_{G_{\overline{X},\overline{Z(W)}}}$$

Example

G

$$
\begin{aligned}
P(c|\mathrm{do}(s)) &= \sum_t P(t|\mathrm{do}(s))P(c|\mathrm{do}(s),t) && \text{(Probability)}\\
&= \sum_t P(t|\mathrm{do}(s))P(c|\mathrm{do}(s),\mathrm{do}(t)) && \text{(Rule2 } G_{\overline{s},\underline{t}}\text{)}\\
&= \sum_t P(t|s)P(c|\mathrm{do}(s),\mathrm{do}(t)) && \text{(Rule2 } G_{\underline{s}}\text{)}\\
&= \sum_t P(t|s)P(c|\mathrm{do}(t)) && \text{(Rule3 } G_{\overline{t},\overline{s}}\text{)}\\
&= \sum_t P(t|s)\sum_u P(u|\mathrm{do}(t))P(c|\mathrm{do}(t),u) && \text{(Probability)}\\
&= \sum_t P(t|s)\sum_u P(u|\mathrm{do}(t))P(c|t,u) && \text{(Rule2 } G_{\underline{t}}\text{)}\\
&= \sum_t P(t|s)\sum_u P(u)P(c|t,u) && \text{(Rule3 } G_{\overline{t}}\text{)}
\end{aligned}
$$

Contents

History of AI

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Reinforcement Learning

General Reinforcement Learning

Pirate Game

Problem (Pirate Game)

5 *rational pirates have a treasure of 100 gold coins.*
The pirate world's rules of distribution are thus:

1. *The fiercest pirate proposes how to split the coins, and all pirates (including the proposer) vote for or against it;*
2. *If* 50% *or more of the pirates vote for it, then the coins will be shared that way. Otherwise, the proposer will be thrown overboard, and the procedure is repeated with the next fiercest pirate;*
3. *As the pirates are bloodthirsty, each pirate would prefer to throw another overboard, if all other results would otherwise be equal.*

Who Will Survive?

Problem (Who Will Survive?)

5 *rational prisoners are going to take beans from a bag with 100 beans. They will do it one by one. No communication is allowed between them. But they can count the beans left in the bag.*

1. *each must take at least* 1 *bean, and the last prisoner need not to take all the rest;*
2. *all prisoners who take the maximum/minimum number will die;*
3. *they will try to survive first and then try to kill more people.*

Monty Hall Problem

Problem (Monty Hall Problem)

You're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No.1, and the host, who knows what's behind the doors, opens another door, say No.3, which has a goat. He then says to you, "Do you want to pick door No.2?"

Bertrand's Paradox

Problem (Bertrand's Paradox)

Consider an equilateral triangle inscribed in a circle. Suppose a chord of the circle is chosen at random. What is the probability that the chord is longer than a side of the triangle?

$\frac{1}{2}$? $\frac{1}{3}$? $\frac{1}{4}$?

What is "randomness"? a process or a product?

Confirmation Problem

$$\frac{\begin{array}{c}\neg Rx \wedge \neg Bx \ \text{ confirms } \ \forall x(\neg Bx \rightarrow \neg Rx) \\ \forall x(\neg Bx \rightarrow \neg Rx) \leftrightarrow \forall x(Rx \rightarrow Bx)\end{array}}{\neg Rx \wedge \neg Bx \ \text{ confirms } \ \forall x(Rx \rightarrow Bx)}$$

Figure: Raven Paradox

Zero-Knowledge Proof

- Alice gave Bob a Sudoku puzzle.
- Bob tried and failed: "Is it an unsolvable one?"
- Alice: "Let me show you that I have a solution. But I will not show you the actual solution."

Prover vs Verifier

- **Completeness:** P can convince V if X is true.
- **Soundness:** no (efficient) P can convince V if X is not true.
- **Zero Knowledge:** no efficient V learns anything more than the validity of X.

Remark: The adversary can simulate the proof without knowing the prover's witness.

Theorem

Every NP statement can be proven in zero-knowledge.

Logic vs Probability — Cox Theorem

Probability theory extends propositional logic?

Assumption (Cox's Assumptions for Beliefs)

1. $A \leftrightarrow B \implies b(\cdot|A) = b(\cdot|B)$ & $b(A|\cdot) = b(B|\cdot)$.
2. *there is a continuous binary operation* $\otimes$ *that is strictly increasing in each coordinate s.t.* $b(A \wedge B|C) = b(A|C) \otimes b(B|A \wedge C)$.
3. *for any rational numbers* $r_1, r_2, r_3 \in (0,1)$ *there are* $A, B, C, D \in \Omega$ *s.t.* $r_1 = b(A|D)$, $r_2 = b(B|A \wedge D)$ *and* $r_3 = b(C|A \wedge B \wedge D)$.
4. *thers is a continuous nonnegative nonincreasing function* $N : [0,1] \to [0,1]$ *s.t.* $b(\neg A|C) = N(b(A|C))$.

Theorem (Cox Theorem)

A credence function that satisfies Cox's assumptions for beliefs is isomorphic to a probability function.

Preferences Lead to Utility

- A lottery $L = [p_1, x_1; \ldots; p_n, x_n]$ is a probability distribution over outcomes $\mathcal{X}$.
- Preferences of a rational agent must obey constraints.

1. completeness $x_1 \succ x_2 \vee x_1 \sim x_2 \vee x_2 \succ x_1$
2. transitivity $x_1 \succ x_2 \wedge x_2 \succ x_3 \rightarrow x_1 \succ x_3$
3. continuity $x_1 \succ x_2 \succ x_3 \rightarrow \exists p : x_2 \sim [p, x_1; 1-p, x_3]$
4. independence $x_1 \sim x_2 \rightarrow [p, x_1; 1-p, x_3] \sim [p, x_2; 1-p, x_3]$
5. monotonicity $x_1 \succ x_2 \wedge p > q \rightarrow [p, x_1; 1-p, x_2] \succ [q, x_1; 1-q, x_2]$
6. decomposability
   $\big[p, x_1; 1-p, [q, x_2; 1-q, x_3]\big] \sim \big[p, x_1; (1-p)q, x_2; (1-p)(1-q), x_3\big]$

Theorem (von Neumann & Morgenstern 1944)

If a preference relation $\succeq$ satisfies the above constraints, then there exists a function $u : \mathcal{X} \rightarrow [0, 1]$ such that

$$x_1 \succeq x_2 \iff u(x_1) \geq u(x_2) \quad \textit{and}$$

$$u\big([p_1, x_1; \ldots; p_n, x_n]\big) = \sum_{i=1}^{n} p_i u(x_i)$$

Game Theory

	cooperate	defect
Cooperate	$-1, -1$	$-4, 0$
Defect	$0, -4$	$-3, -3$

Table: Prisoner's Dilemma

	opera	football
Opera	$1, 2$	$0, 0$
Football	$0, 0$	$2, 1$

Table: Battle of the Sexes

	stop	go
Stop	$0, 0$	$0, 1$
Go	$1, 0$	$-\infty, -\infty$

Table: Chicken/Traffic

	head	tail
Head	$1, -1$	$-1, 1$
Tail	$-1, 1$	$1, -1$

Table: Matching Pennies

Pareto Optimality

- Given an initial situation, a *Pareto improvement* is a new situation where some agents will gain, and no agents will lose.
- A situation is called *Pareto dominated* iff it has a Pareto improvement.
- A situation is called *Pareto optimal* iff no change could lead to improved satisfaction for some agent without some other agent losing.
- The *Pareto frontier* is the set of all Pareto optimal allocations.

not change	change
$100, 100$	$1000, 99$

Table: Pareto Improvement vs Rawls' Fair Opportunity Principle (veil of ignorance)

$$0.5 * 1000 + 0.5 * 90 = 545 > 100$$

卡尔多-希克斯（Kaldor-Hicks）标准

卡尔多-希克斯（Kaldor-Hicks）标准

如果一种变革使得受益者的所得足以弥补受损者的所失，这种变革就是一个卡尔多-希克斯改进。如果受损者得到足够的补偿，就是帕累托改进。

- 一家化工厂与家属区一墙之隔，为了上班方便，居民在化工厂的墙上挖了一个洞。一天，家属区的一个小孩钻洞进工厂玩，找到一瓶化学液体并点燃烧伤了自己。
- 假设化工厂补墙成本为 c，如果不补墙，发生事故的概率为 p，损失为 l，且 $c < pl$，化工厂需要承担责任吗?

外部性与科斯定理

- 个人收益与社会收益：一项活动的社会收益等于决策者个人得到的收益加社会其他成员得到的收益（如养花）；
- 个人成本与社会成本：社会成本等于决策者的个人承担的成本加社会其他成员承担的成本（如环境污染、交通堵塞）；
- 如果个人收益（成本）不等于社会收益，就存在外部性。
- 个人最优与社会最优的不一致意味着有帕累托改进的余地；
- 如何将外部性内部化？如何使得个人在边际上承担全部的社会成本、获得全部的社会收益？
- 政府征税或补贴？
- 科斯定理：只要产权界定是清楚的，如果交易成本为零，外部性可以通过当事人之间谈判解决，帕累托最优可以实现；并且，最终的资源配置与初始的产权安排无关。

科斯定理 — 例子

- 如果产权归农场主，农场主禁止放牧，小于社会最优量 S；但是，增加放牧给牧羊人带来的边际利润大于给农场主造成的损失，牧羊人将有积极性贿赂农场主，直到放牧量达到 S 为止；
- 如果产权归牧羊人，牧羊人的利润最大点是 P，大于社会最优量 S；但是，减少放牧量对牧羊人的边际利润损失小于给农场主节约的边际成本，农场主将有积极性贿赂牧羊人，直到放牧量达到 S 为止。

沉默的目击者

- n 个目击者围观 Kitty 被虐杀。
- 报警的成本是 c。
- 如果有人报警，Kitty 会得救，每个目击者会获得效用 v。如果没有人报警，目击者的效用是 0。

	报警	不报警
报警	$v-c, v-c$	$v-c, v$
不报警	$v, v-c$	$0, 0$

- 当 $v < c$ 时，所有人不报警是纳什均衡。
- 当 $v > c$ 时，有一个人报警、其他人不报警是纳什均衡。
- 考虑对称的混合纳什均衡，即所有参与人报警的概率相等，设为 p。此时，报警与不报警的期望收益相等。

$$v - c = v(1-(1-p)^{n-1}) \implies p = 1 - \left(\frac{c}{v}\right)^{\frac{1}{n-1}}$$

- n 个人中至少有一人报警的概率为

$$P(n) := 1-(1-p)^n = 1 - \left(\frac{c}{v}\right)^{\frac{n}{n-1}}$$

但 $\frac{\mathrm{d}P(n)}{\mathrm{d}n} < 0$，随着人数的增多，至少有一个人报警的概率下降。

走出囚徒困境 — 奖惩 — 利维坦

	cooperate	defect
Cooperate	$-1, -1$	$-4, 0$
Defect	$0, -4$	$-3, -3$

Table: Prisoner's Dilemma

	cooperate	defect
Cooperate	$-1, -1$	$-4, 0 - x$
Defect	$0 - x, -4$	$-3 - x, -3 - x$

Table: Prisoner's Dilemma with Punishment $0 - x < -1$

走出囚徒困境 — 奖惩 — 作为激励机制的等级制度

	cooperate	defect
Cooperate	$3, 3$	$-1, 4$
Defect	$4, -1$	$0, 0$

$$\Downarrow$$

	first	second
First	$0, 0$	$2, 1$
Second	$1, 2$	$0, 0$

儒家 — “礼”

“合作”方可得“君子”名分，君子享有优先权。
协调预期、定分止争。
声誉约束。

Multiple Nash Equilibria

	left	right
Left	$1, 1$	$0, 0$
Right	$0, 0$	$1, 1$

Table: Coordination Game

	wcdma	td-scdma
WCDMA	$8, 8$	$3, 2$
TD-SCDMA	$2, 3$	$4, 4$

Table: 协商选择帕累托最优的纳什均衡

Mixed Nash Equilibrium

	偷懒	不偷懒
监督	$1, -1$	$-1, 2$
不监督	$-2, 3$	$2, 2$

- 如果员工偷懒的概率是 p，那么，老板监督和不监督的期望收益分别为

$$1 \cdot p + (-1) \cdot (1 - p) = 2p - 1$$

$$(-2) \cdot p + 2 \cdot (1 - p) = 2 - 4p$$

- 如果老板监督的概率是 q，那么，员工偷懒和不偷懒的期望收益分别为

$$(-1) \cdot q + 3 \cdot (1 - q) = 3 - 4q$$

$$2 \cdot q + 2 \cdot (1 - q) = 2$$

- 混合纳什均衡是 $(\frac{1}{4}, \frac{1}{2})$，老板以 $\frac{1}{4}$ 的概率监督，员工以 $\frac{1}{2}$ 的概率偷懒。

承诺

Figure: 如果 Alice 承诺不选择 U

假如企业一开始定价 80，如果前两个消费者购买了，企业将有积极性在 50 的价格下向第三个顾客出售。预期到这一点，前两个顾客将不会购买。如果企业承诺，任何降价的差额将返还顾客，前两个顾客则会购买。

产量	价格	收入
1	100	100
2	80	160
3	50	150
4	30	120

- 承诺意味着限制自己的自由：选择少反而对自己好。
- 为什么画家死了画会升值?
- 民主、法治是政府对人民的一种承诺。
- 民法“民不告、官不究”，而刑法实行“公诉制度”是一种承诺。

逆向选择 — 劣币驱逐良币

- 假设好车和坏车对卖方的保留价值分别为 $2x$ 和 x，对买方的价值是卖方的 a 倍。成交价为 Q。
- 为使交易达成，买方买车的期望收益应不小于不买车的期望收益。

$$(2ax - Q)p + (ax - Q)(1 - p) \geq 0$$

- 拥有好车的卖方能接受的最低价是 $Q \geq 2x$。
- 这意味着，买卖双方合意的价格需满足：

$$(1 + p)ax \geq Q \geq 2x \implies p \geq \frac{2}{a} - 1$$

如何解决非对称信息的问题?

Example (逆向选择)

- 坏车使得好车不能成交。
- 高风险的投保人使得低风险的投保人无法投保。
- 假乞丐使得真乞丐得不到救济。
- 高水平的学者竞争不过灌水的学者。

解决非对称信息的市场机制和政府管制

1. 信号显示（signaling）：卖方承诺保修。
2. 信息甄别（screening）：保险公司向投保人提供不同的合同选择，投保人根据自己的状况选择适合自己的合同。
3. 声誉机制：品牌的价值。
4. 适度的政府管制。但政府管制不当可能破坏声誉机制的有效性 — 管制导致企业预期不稳定，追求短期行为；管制导致垄断，使得市场惩罚不可信；管制导致腐败，贿赂官员比贿赂投资者和客户更合算。

Boxed Pig Game

	press	not press
Press	$4, 0$	$3, 3$
Not Press	$7, -1$	$0, 0$

Table: Boxed Pig Game

	press	not press
Press		$3, 3$
Not Press		$0, 0$

Table: Rationalizability and Iterated Elimination of Dominated Actions

	press	not press
Press		$3, 3$
Not Press		

Table: Rationalizability and Iterated Elimination of Dominated Actions

Iterated Elimination of Dominated Actions and n^{th}-order Rationality

	c_1	c_2	c_3	c_4
r_1	$5,10$	$0,11$	$1,20$	$10,10$
r_2	$4,0$	$1,1$	$2,0$	$20,0$
r_3	$3,2$	$0,4$	$4,3$	$50,1$
r_4	$2,93$	$0,92$	$0,91$	$100,90$

	c_1	c_2	c_3
r_1	$5,10$	$0,11$	$1,20$
r_2	$4,0$	$1,1$	$2,0$
r_3	$3,2$	$0,4$	$4,3$
r_4	$2,93$	$0,92$	$0,91$

	c_1	c_2	c_3
r_1	$5,10$	$0,11$	$1,20$
r_2	$4,0$	$1,1$	$2,0$
r_3	$3,2$	$0,4$	$4,3$

	c_2	c_3
r_1	$0,11$	$1,20$
r_2	$1,1$	$2,0$
r_3	$0,4$	$4,3$

	c_2	c_3
r_2	$1,1$	$2,0$
r_3	$0,4$	$4,3$

0-order	$c_4\times$	$\text{rational}(C)$
1-order	$r_4\times$	$K_R\,\text{rational}(C)$
2-order	$c_1\times$	$K_C K_R\,\text{rational}(C)$
3-order	$r_1\times$	$K_R K_C K_R\,\text{rational}(C)$
4-order	$c_3\times$	$K_C K_R K_C K_R\,\text{rational}(C)$
5-order	$r_3\times$	$K_R K_C K_R K_C K_R\,\text{rational}(C)$

风险与均衡

- 由于纳什均衡要求理性共识和一致预期，当人们可能犯小小的错误时，纳什均衡不一定被选择。

	left	right
up	$8, 10$	$-1000, 9$
down	$7, 6$	$6, 5$

Table: 只要 B 有千分之一的概率错误地选择 right，A 将选择 down；如果 B 怀疑 A 怀疑自己可能犯错误，B 将选择 right。

Infinitely Repeated Games

	cooperate	defect
Cooperate	T, T	S, R
Defect	R, S	P, P

Table: Prisoner's Dilemma: $R > T > P > S$ and $T + T > R + S$.

1. All-D 策略：总是背叛。
2. All-C 策略：总是合作。
3. 合作-背叛交替进行。
4. 以牙还牙 tit-for-tat TFT 策略：从合作开始，之后每次选择对方前一阶段的行动。
5. 冷酷 grim 策略：从合作开始，直到一方背叛，然后永远背叛。
6. 宽容的冷酷策略：如果对方背叛，先惩罚几次，然后再恢复合作。
7. 宽容的以牙还牙：永远以合作的态度来回报对方的合作。当遇到背叛时，以某一概率与对方进行合作。
8. 赢定输移 win stay, lose shift 策略：如果我们上一轮合作，那么合作；如果上一轮都背叛，那么以一定概率合作；如果上一轮我合作你背叛或你合作我背叛，则背叛。

- $V(\text{ALL-D}, \text{ALL-D}) = P + \gamma P + \gamma^2 P + \cdots = P\frac{1}{1-\gamma}$
- $V(\text{TFT}, \text{TFT}) = T + \gamma T + \gamma^2 T + \cdots = T\frac{1}{1-\gamma}$
- $V(\text{ALL-D}, \text{TFT}) = R + \gamma P + \gamma^2 P + \cdots = R + P\frac{\gamma}{1-\gamma}$
- $V(\text{ALL-C}, \text{grim}) = T + \gamma T + \gamma^2 T + \cdots = T\frac{1}{1-\gamma}$
- $V(\text{ALL-D}, \text{grim}) = R + \gamma P + \gamma^2 P + \cdots = R + P\frac{\gamma}{1-\gamma}$

如果 $T\frac{1}{1-\gamma} \geq R + P\frac{\gamma}{1-\gamma}$，即 $\gamma \geq \frac{R-T}{R-P}$，则对于 grim 策略来说，合作就是精炼纳什均衡。

无名氏定理（Folk Theorem）

在无限次重复博弈中，如果每个参与人都对未来足够重视（贴现因子足够大），那么，任何程度的合作都可以作为一个精炼纳什均衡得到。这里的“合作程度”指整个博弈中合作出现的频率。

惩罚

- 在重复博弈中，越在乎长远利益，合作的可能性越大。
- 背叛行为越容易被观察到，并且惩罚越可信，合作的可能性越大。
- 垄断使得惩罚不可信。
- 在确定环境中，惩罚越严厉越有助于合作，但在不确定环境中，对方有可能是无心之失，冷酷策略可能不利于长期合作。

联合抵制的社会规范

Boycott（联合抵制）：每个人都应该诚实；都有责任惩罚骗过人的人；不参与惩罚的人应该受到惩罚。

Remark: 朋友的朋友是朋友；朋友的敌人是敌人；敌人的朋友是敌人。

有限重复博弈

	c_1	c_2	c_3
r_1	$1,1$	$5,0$	$0,0$
r_2	$0,5$	$4,4$	$0,0$
r_3	$0,0$	$0,0$	$3,3$

- 两个纳什均衡：(r_1, c_1) 和 (r_3, c_3)
- 帕累托最优：(r_2, c_2)
- 如果博弈重复两次，则“好合好散，不欢而散”的策略——如果 (r_2, c_2) 则 (r_3, c_3)，否则 (r_1, c_1) ——可在第一轮博弈中实现帕累托最优。
- 但是，如果第一轮遭到背叛后，第二轮对方重新谈判，则会使得惩罚不可信。原因在于多重均衡之间 (r_3, c_3) 帕累托优于 (r_1, c_1)。

有限重复博弈

	c_1	c_2	c_3	c_4	c_5
r_1	$1,1$	$5,0$	$0,0$	$0,0$	$0,0$
r_2	$0,5$	$4,4$	$0,0$	$0,0$	$0,0$
r_3	$0,0$	$0,0$	$3,3$	$0,0$	$0,0$
r_4	$0,0$	$0,0$	$0,0$	$4,0.5$	$0,0$
r_5	$0,0$	$0,0$	$0,0$	$0,0$	$0.5,4$

- 四个纳什均衡：(r_1, c_1)、(r_3, c_3)、(r_4, c_4)、(r_5, c_5)
- 三个帕累托最优：(r_2, c_2)、(r_4, c_4)、(r_5, c_5)
- 如果博弈重复两次，则可采取如下策略：如果 (r_2, c_2) 则 (r_3, c_3)；否则，被背叛的一方根据自己的情况选择 r_4 或 c_5，如果双方同时背叛，则 (r_3, c_3)。
- 此时惩罚变得可信了，均衡为：第一轮 (r_2, c_2) 第二轮 (r_3, c_3)。

单方不完全信息的有限重复博弈

	合作	背叛
合作	$3,3$	$-1,4$
背叛	$4,-1$	$0,0$

	t1	t2
A 冷酷型 p	合作	X
A 理性型 $1-p$	背叛	背叛
B 理性型	X	背叛

Table: 囚徒博弈重复两次

- 参与人 A 有两种可能的类型："冷酷"型：选择 grim 策略，概率为 p；"理性"型：可以选择任何策略，概率为 $1-p$。
- 参与人 B 有一种类型：理性型。
- B 在第一轮选择合作或背叛最后总的期望效用分别为

$$3 \cdot p + (-1) \cdot (1-p) + 4 \cdot p + 0 \cdot (1-p) = 8p - 1$$

$$4 \cdot p + 0 \cdot (1-p) + 0 \cdot p + 0 \cdot (1-p) = 4p$$

- 如果 $p \geq \frac{1}{4}$，则 B 在第一轮合作。
- 如果博弈重复 $N \geq 3$ 轮，只要 $p \geq \frac{1}{4}$，理性型 A 在 $t = 1 \ldots N-2$ 轮合作，在最后两轮背叛；B 在前 $N-1$ 轮合作，在最后一轮背叛。

双方不完全信息的有限重复博弈

- 假设双方都有两种可能的类型："冷酷"型或"理性"型。
- 如果一开始就选择背叛，暴露了自己是理性型，那么收益最大为 4。
- 假如对方采取冷酷策略的概率是 p，则采取冷酷策略的最小期望收益为

$$3 \cdot N \cdot p + (-1 + 0 + 0 + \cdots + 0) \cdot (1 - p) = (3N + 1)p - 1$$

- 只要博弈次数 $N \geq \frac{5-p}{3p}$，则采取冷酷策略。
- 在不完全信息的情况下，只要博弈重复的次数足够长，参与人就有积极性在博弈的早期建立一个"合作"的**声誉**；一直到博弈的后期，才会选择背叛；并且，背叛的轮数只与 p 有关，而与博弈的次数 N 无关。

教育水平的信号传递作用

- 求职者 50% 可能性是高能力者 50% 可能性是低能力者。
- 高能力者的生产率是 200，低能力者 100。
- 雇主愿付高能力者工资 200，低能力者 100，不知道求职者能力高低则付平均工资 150。
- 假设高能力者受教育成本 40，低能力者受教育成本 120。
- 此时，教育可以成为传递能力的信号。
- 但如果低能力者受教育的成本低于 100，文凭就无法成为雇主区分能力高低的信号，雇主愿付的工资就是 150，也就没人愿上大学了。
- 因此，关键是不同类型的人信号传递成本不同；只有成本差异足够大，才有可能传递信号。

Spence's Education Game

- The worker can be one of two types: wise θ_H (with probability p) or dumb θ_L (with probability $1-p$). Each type can select their own level of education, e_H or e_L.
- The employer is assumed to have two choices. One is to ignore the signal and set $w^* := p\theta_H + (1-p)\theta_L$. The other is to pay a worker w_H or w_L based on whether the signal is e_H or e_L.
- The employer's payoff is $\theta - w$. The worker's payoff is $w - e/\theta$.
- This game has two equilibria. The first is a **pooling equilibrium**, in which the worker will choose the same level of education regardless of his type, and the employer pays all workers the same amount w^*.
- The other is a **separating equilibrium**, in which the worker will choose a different level of education. A low-talent worker will get no education, $e_L = 0$. The education chosen by a high-talent worker is set in such a way as to make it unprofitable for either type of worker to mimic the other.

$$w_L - 0/\theta_L \geq w_H - e_H/\theta_L \quad \text{and} \quad w_H - e_H/\theta_H \geq w_L - 0/\theta_H$$

信号传递的作用

- 为什么雄性孔雀的尾巴越长，越受到雌性孔雀的青睐？— 只有健壮者才能负担得起长尾巴。
- 高、低质量产品哪个更愿意做广告？— 广告费是高质量产品企业向市场传递信息的成本。
- 如何送礼？— 重要的是送礼对送者的成本，而不是礼物对接受者的价值。
- 中秋为什么送浪费性的月饼？请客吃巨贵的馆子？— 成本要高于价值，甚至“毁灭”了很大一部分价值更显出对对方的重视。“千里送鹅毛，礼轻仁义重。”
- 公费请客送礼传递的信息量大打折扣，成本需要翻倍！
- 为什么领结婚证？— 离婚分财产。同样，为什么送彩礼？为什么婚礼要铺张？
- 为什么街头古惑仔纹身？黑社会老大穿西装戴眼镜？
- 为什么很多繁文缛节没有实质意义的礼仪还要遵守？— 合作精神

信息不完全导致社会规范变迁

- 如果是完全分离均衡，每类人的行为都是特定的。
- 如果是混同均衡，所有人的行为都是一样的。
- 如果是准分离（混同）均衡，有些行为传递信息，有些行为不传递信息。
- 如果外部因素导致社会由分离均衡转向混同均衡或准分离均衡，社会规范就会发生变化。

人们对婚前性行为和婚外性行为态度的变化：在封闭的社会，婚前和婚外性行为都很容易观察；在流动的社会，有些能观察到，有些不能；如果被观察到的只是其中的一小部分，被观察到压力将会减少。

Auction

1. first-price open cry (English auction).
2. dutch auction: (descending auction) the seller lowers the price until it is taken.
3. first-price sealed bid: bidding without knowing the other bids.
4. second-price sealed bid: (Vickrey auction) Highest bidder wins, but the price is the second highest bid!

Theorem

Truth-telling is a dominant strategy in a second-price sealed bid.

Remark: 次价格密封拍卖的程序与高价格密封拍卖一样，中标者仍然是出价最高者，但中标者实际支付的是第二高报价。次价格密封拍卖能让竞标者有积极性说真话。

Groves-Clarke-Vickrey Mechanism

- 公共产品的偏好显示：不同的人有不同的偏好，是私人信息。如何让每个人报告自己的真实偏好?
- 每个人可以任意地报告自己的偏好，但可能要纳缴一定数量的“税”。计算办法：先将其他人的偏好加总，给出总价值最大的项目；然后将第一个人的偏好加上，如果不影响结果，不征税；否则，应纳税等于改变结果给其他人带来的损失。
- 比如同学聚会，选择哪个餐馆是一个公共产品。此机制可以让每人说出自己的真实偏好。

	川菜	粤菜	税额
A	30	10	0
B	0	40	30
C	20	10	0
合计	50	60	30

官员腐败问题

- W：官员工资。
- $B(q)$：权力租金，即官员在位置上可能收受的贿赂，也可看作不腐败的机会成本。一般来说，权力 q 越大，权力租金越高。
- p：腐败被发现的概率。
- F：对腐败的处罚。
- U：政府外的保留效用。
- α：官员的羞耻感或脸皮厚度。
- 官员腐败的期望收益

$$(1-p)(W+B(q))+p(U-\alpha F)$$

- 官员不腐败的条件：

$$W \geq \frac{1-p}{p}B(q)+U-\alpha F$$

- 如何控制腐败？提高 p, F, W, α，减少 q。

演化博弈 — 婚姻博弈

	物质型	感情型
物质型	$1, 1$	$0, 0$
感情型	$0, 0$	$2, 2$

- 假定总人口中，物质型的比例为 x，感情型的比例为 $1 - x$。
- 对任何一个个体而言，物质型的期望效用：$x1 + (1 - x)0 = x$。
- 感情型的期望效用：$x0 + (1 - x)2 = 2(1 - x)$。
- 如果 $x > 2/3$，物质型更适合生存，将演化成稳定均衡。
- 如果 $x < 2/3$，感情型更适合生存，将演化成稳定均衡。
- 如果 $x = 2/3$，两类人有同样的适应性，但此二元均衡是非稳定的。

演化博弈 — 鹰鸽博弈

	鹰	鸽
鹰	$-1, -1$	$1, 0$
鸽	$0, 1$	$\frac{1}{2}, \frac{1}{2}$

- 假定鹰派的比例是 x，鸽派的比例是 $1-x$。
- 鹰派的效用：$-1x + 1(1-x) = 1 - 2x$。
- 鸽派的效用：$0x + \frac{1}{2}(1-x) = \frac{1}{2}(1-x)$。
- 如果 $x < 1/3$，鹰派占优势，不稳定。
- 如果 $x > 1/3$，鸽派占优势，不稳定。
- 如果 $x = 1/3$，同样的适应性，稳定。
- 如果初始人口由单一类型构成，另一类型可以成功入侵，直到均衡。

† 两个纯策略均衡：（鹰、鸽），（鸽、鹰）；一个混合均衡：$(1/3, 2/3)$

Δ 假定存在某种显性的标记机制：在博弈开始之前，每个人收到一个信号：A 或 B；概率是 1/2；信号完全负相关；标记是公共知识。

1. 如果 A，选择“鹰”；如果 B，选择“鸽”。是 ESS。
2. 如果 A，选择“鸽”；如果 B，选择“鹰”。是 ESS。
3. 无论 AB，以 1/3 的概率选择“鹰”，2/3 的概率选择“鸽”。不是 ESS。

Definition (Games in Extensive Form)

A game in *extensive form* $\Gamma = \langle N, \mathcal{A}, \mathcal{H}, \mathcal{Z}, A, P, (\mathcal{I}_i)_{i \in N}, f_c, (u_i)_{i \in N} \rangle$.

- The set of players N.
- The set of actions $\mathcal{A} := \bigcup_{i \in N} \mathcal{A}_i$.
- A set $\mathcal{H}$ of sequences (finite or infinite) such that,
  1. $\emptyset \in \mathcal{H}$
  2. $\forall m < n \, (h_{1:n} \in \mathcal{H} \implies h_{1:m} \in \mathcal{H})$
  3. $\forall n \in \mathbb{N} \, (h_{1:n} \in \mathcal{H}) \implies h_{1:\infty} \in \mathcal{H}$
- A history $h \in \mathcal{H}$ is terminal iff $h = h_{1:\infty}$ or there is no $a \in \mathcal{A}$ s.t. $ha \in \mathcal{H}$. The set of terminal histories is denoted $\mathcal{Z}$.
- The set of actions at h: $A(h) := \{a : ha \in \mathcal{H}\}$.
- The player function: $P : \mathcal{H} \setminus \mathcal{Z} \to N \cup \{c\}$.
- The function f_c associates with every history h for which $P(h) = c$ a probability measure $f_c(\cdot|h)$ on $A(h)$. $f_c(\cdot|h) \in \Delta(A(h))$.
- For player $i \in N$, $\mathcal{I}_i := \mathrm{part}(\{h \in \mathcal{H} : P(h) = i\})$ where $\forall h, h' \in I \in \mathcal{I}_i : A(h) = A(h')$.
- The Payoff function for each player $i \in N$: $u_i : \mathcal{Z} \to \mathbb{R}$.

- If $|I| = 1$ for all $I \in \mathcal{I}_i$ and all $i \in N$, then we say that Γ is a game with *perfect information*.
- If $|I| > 1$ for some $i \in N$ and some $I \in \mathcal{I}_i$, then we say that Γ is a game with *imperfect information*.
- If all elements of Γ are common knowledge, then we say that Γ is a game with *complete information*; otherwise, a game with *incomplete information*.

Theorem

Every perfect information game in extensive form has a pure strategy Nash equilibrium.

Translating from Extensive Form to Matrix Form

Figure: The Hawk-Dove Game

	$(H\|H, H\|D)$	$(H\|H, D\|D)$	$(D\|H, D\|D)$	$(D\|H, H\|D)$
H	$-2, -2$	$-2, -2$	$4, 0$	$4, 0$
D	$0, 4$	$2, 2$	$2, 2$	$0, 4$

Pure Strategy

Definition (Pure Strategy)

A *pure strategy* of player i is $s_i : \mathcal{I}_i \to \mathcal{A}$ s.t. $s_i(I) \in A(I)$ for all $I \in \mathcal{I}_i$.
The *set of pure strategies* of player i is denoted by $S_i := \prod_{I \in \mathcal{I}_i} s_i(I)$.

A *strategy profile*

$$s = (s_i)_{i \in N}$$

is a vector consisting of one pure strategy for each player.
The *set of strategy profiles* is

$$S = \prod_{i \in N} S_i$$

$$s_{-i} := (s_1 \dots s_{i-1}; s_{i+1} \dots s_{|N|})$$

$$S_{-i} := S_1 \times \dots \times S_{i-1} \times S_{i+1} \times \dots \times S_{|N|}$$

Definition (Pure Strategy Consistent with History)

For any history h define a pure strategy s_i of player i to be consistent with h

$$s_i \rightsquigarrow h \iff \forall k < \ell(h)\,(P(h_{1:k}) = i \implies s_i(h_{1:k}) = h_{k+1})$$

$$\sigma_c(h) := \prod_{h' \sqsubseteq h} f_{P(h')}(h_{\ell(h')+1} | h')$$

where $f_{P(h)}(\cdot|h) := 1$ if $P(h) \neq c$.
Each strategy profile $s = (s_i)_{i \in N}$ can be assigned a payoff for each player i:

$$u_i(s) := \sum_{h : \forall i\,(s_i \rightsquigarrow h)} \sigma_c(h) u_i(h)$$

Obviously,

$$\forall h \in \mathcal{H}\,(P(h) \neq c) \implies u_i(s) = u_i(h^s)$$

where h^s is the history uniquely determined by s.

Mixed Strategy

Definition (Mixed Strategy)

A *mixed strategy* of player i in an extensive game is a probability measure over the set of player i's pure strategies:

$$\sigma_i \in \Delta S_i$$

A *mixed strategy profile* specifies a mixed strategy for each player i

$$\sigma := (\sigma_i)_{i \in N}$$

The *set of mixed strategy profiles* is:

$$\prod_{i \in N} \Delta S_i$$

Each mixed strategy profile $\sigma = (\sigma_i)_{i \in N}$ implies a probability distribution over the *set of pure strategy profiles* S

$$\sigma(s) := \prod_{i \in N} \sigma_i(s_i)$$

Payoff of Mixed Strategy

Definition (Payoff of Mixed Strategy)

The expected payoff of player i given a mixed strategy profile σ is

$$u_i(\sigma) := \sum_{s \in S} u_i(s)\sigma(s)$$

Behavioural Strategy

Definition (Behavioural Strategy)

A *behavioural strategy* of player i is a collection $(\rho_i(I))_{I \in \mathcal{I}_i}$ of probability measures, where

$$\rho_i(I) \in \Delta(A(I)) \text{ for } I \in \mathcal{I}_i$$

A *behavioural strategy profile* is a vector consisting of one behavioural strategy for each player:

$$\rho := \left((\rho_i(I))_{I \in \mathcal{I}_i}\right)_{i \in N}$$

For any history $h \in I \in \mathcal{I}_i$ and action $a \in A(h)$, we denote by $\rho_i(h)(a)$ the probability $\rho_i(I)(a)$ assigned by $\rho_i(I)$ to the action a.
The *set of behavioural strategies* of player i is

$$B_i := \prod_{I \in \mathcal{I}_i} \Delta(A(I))$$

The *set of behavioural strategy profiles* is:

$$B := \prod_{i \in N} B_i = \prod_{I \in \mathcal{I}} \Delta(A(I))$$

Payoff of Behavioural Strategy

Definition (Payoff of Behavioural Strategy)

The expected payoff of player i given a behavioural strategy profile ρ is

$$u_i(\rho) := \sum_{s \in S} u_i(s) \prod_{i \in N} \prod_{I \in \mathcal{I}_i} \rho_i(I)(s_i(I))$$

Perfect Recall

Definition (Perfect Recall)

An extensive game has *perfect recall* iff for every player $i \in N$ and for histories $h, h' \in I \in \mathcal{I}_i$

- $\ell(h) = \ell(h')$
- $\forall k \leq \ell(h) : P(h_{1:k}) = i \implies h_{k+1} = h'_{k+1} \;\&\; \exists I \in \mathcal{I}_i (h_{1:k} \in I \;\&\; h'_{1:k} \in I)$

Figure: Imperfect Recall

Outcome-Equivalence

For any profile σ/ρ, we define the *outcome* $\overline{\sigma}(h)/\overline{\rho}(h)$ to be the multiplicative product of all the chance probabilities and move probabilities over history h when all players choose their moves according to σ_i/ρ_i.

$$\sigma_i(h) := \sum_{s_i \rightsquigarrow h} \sigma_i(s_i)$$

$$\overline{\sigma}(h) := \prod_{i \in N \cup \{c\}} \sigma_i(h)$$

$$\overline{\rho}(h) := \prod_{k=0}^{\ell(h)} \rho_{P(h_{<k})}(h_{<k})(h_k)\sigma_c(h)$$

Two (mixed or behavioural) strategies of any player are *outcome-equivalent* iff for every collection of pure strategies of the other players the two strategies induce the same outcome.

Theorem (Outcome-Equivalent Theorem)

In games with perfect recall, for any mixed strategy there is an outcome-equivalent behavioural strategy and vice versa.

Proof.

" $\implies$ "

with perfect recall,

$$\forall h, h' \in I \in \mathcal{I}_i \forall a \in A(I)\, (\sigma_i(h) = \sigma_i(h') \implies \sigma_i(ha) = \sigma_i(h'a))$$

$$\rho_i(I)(a) := \frac{\sigma_i(ha)}{\sigma_i(h)} \text{ for any } h \in I \in \mathcal{I}_i$$

" $\impliedby$ "

$$\sigma_i(s_i) := \prod_{I \in \mathcal{I}_i} \rho_i(I)(s_i(I))$$

Games in Strategic Form

Definition (Strategy Form)

A game in *strategic form* is a triplet $\langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$.

- The sets of players N.
- The sets of strategies of the players $(S_i)_{i \in N}$.
- The payoff functions $u_i : S \to \mathbb{R}$.

Definition (Mixed Extension)

The mixed extension of the strategic game $\langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$ is $\langle N, (\Delta S_i)_{i \in N}, (u_i)_{i \in N} \rangle$, where the payoff function u_i in $\langle N, (\Delta S_i)_{i \in N}, (u_i)_{i \in N} \rangle$ is

$$u_i(\sigma) = \sum_{s \in S} u_i(s)\sigma(s)$$

Note that u_i is multilinear.

$$u_i(\lambda\sigma_i' + (1-\lambda)\sigma_i''; \sigma_{-i}) = \lambda u_i(\sigma_i'; \sigma_{-i}) + (1-\lambda)u_i(\sigma_i''; \sigma_{-i})$$

Pareto Optimality

- A strategy profile s^* *strongly Pareto-dominates* s iff $\forall i \in N : u_i(s^*) > u_i(s)$.
- A strategy profile s^* *weakly Pareto-dominates* s iff $\forall i \in N : u_i(s^*) \geq u_i(s)$ and $\exists i \in N : u_i(s^*) > u_i(s)$.
- A strategy profile s^* is *Pareto optimal* iff

$$\forall i \in N \forall s \in S : u_i(s) > u_i(s^*) \implies \exists j \in N : u_j(s) < u_j(s^*)$$

Dominant Strategy Equilibrium

Definition (Dominant Strategy)

A strategy $s_i^* \in S_i$ is a *dominant strategy* iff

$$u_i(s_i^*; s_{-i}) \geq u_i(s_i; s_{-i}) \text{ for all } s_i \in S_i \text{ and for all } s_{-i} \in S_{-i}$$

Definition (Dominant Equilibrium)

A strategy profile s^* is a *dominant strategy equilibrium* iff for each player $i \in N$, s_i^* is a dominant strategy.

Definition (Strictly Dominated Strategy)

A strategy s_i is called strictly dominated iff

$$\exists s_i' \forall s_{-i} : u_i(s_i'; s_{-i}) > u_i(s_i; s_{-i})$$

Definition (Weakly Dominated Strategy)

A strategy s_i is called weakly dominated iff

$$\exists s_i' \forall s_{-i} : u_i(s_i'; s_{-i}) \geq u_i(s_i; s_{-i})$$

and

$$\exists s_{-i}' : u_i(s_i'; s_{-i}') > u_i(s_i; s_{-i}')$$

Definition (Iterated Dominance Equilibrium)

An *iterated dominance equilibrium* is a strategy profile $s^* \in S$ obtained by iteratively ruling out dominated strategies from each player until only one strategy remains for each player.

Nash Equilibrium

Definition (Nash Equilibrium)

Best response of player i:

$$\mathrm{BR}_i(s_{-i}) := \operatorname*{argmax}_{s_i \in S_i} u_i(s_i; s_{-i})$$

The strategy profile s^* is a *Nash equilibrium* iff

$$\forall i \in N : s_i^* \in \mathrm{BR}_i(s_{-i}^*)$$

Definition (Mixed Nash Equilibrium)

A mixed strategy profile σ^* is a *mixed Nash equilibrium* iff,

$$\forall i \in N \forall \sigma_i \in \Delta S_i : \; u_i(\sigma_i^*; \sigma_{-i}^*) \geq u_i(\sigma_i; \sigma_{-i}^*)$$

A mixed strategy Nash equilibrium of a strategic game is a Nash equilibrium of its mixed extension.

Kakutani Fixpoint Theorem

Theorem (Kakutani Fixpoint Theorem)

Let X be a nonempty subset of n-dimensional Euclidean space, and $f : X \to 2^X$. The following are sufficient conditions for f to have a fixed point (i.e. an $x^* \in X$ with $x^* \in f(x^*)$):

1. *X is compact: any sequence in X has a limit in X.*
2. *X is convex.*
3. *for any x, $f(x)$ is nonempty and convex.*
4. *For any sequence of pairs (x_i, x_i^*) such that $x_i, x_i^* \in X$ and $x_i^* \in f(x_i)$,*
$$\lim_{i\to\infty}(x_i, x_i^*) = (x, x^*) \implies x^* \in f(x)$$

Theorem (Existence of Mixed Nash Equilibrium)

Every finite strategic game has a mixed Nash equilibrium.

Proof.

Let $X := \prod_{i\in N} \Delta S_i$ and $f(\sigma) := \prod_{i\in N} \mathrm{BR}_i(\sigma_{-i})$.

Theorem

A mixed strategy profile σ^* is a mixed Nash equilibrium iff,

$$\forall i \in N \forall s_i \in S_i : \; u_i(\sigma_i^*; \sigma_{-i}^*) \geq u_i(s_i; \sigma_{-i}^*)$$

Proof.

$$u_i(\sigma_i; \sigma_{-i}) = \sum_{s_i \in S_i} u_i(s_i; \sigma_{-i}) \sigma_i(s_i)$$

Theorem

For a finite strategic game,

$$\mathrm{supp}(\sigma_i^*) \subset \mathrm{BR}_i(\sigma_{-i}^*)$$

Proof.

Suppose there exists an $s_i' \in \mathrm{supp}(\sigma_i^*)$ s.t.

$$u_i(s_i'; \sigma_{-i}^*) < u_i(\sigma_i^*; \sigma_{-i}^*)$$

then

$$\sum_{s_i \in S_i} u_i(s_i; \sigma_{-i}^*)\sigma_i^*(s_i) < u_i(\sigma_i^*; \sigma_{-i}^*)$$

Contradiction!

- Every action in the support of any player's equilibrium mixed strategy yields that player the same payoff.
- If the set of actions of some player is not finite the result needs to be modified. In this case, σ^* is a mixed strategy Nash equilibrium iff
  1. for every player i no action in S_i yields, given σ^*_{-i}, a payoff to player i that exceeds his equilibrium payoff, and
  2. the set of actions that yield, given σ^*_{-i}, a payoff less than his equilibrium payoff has σ^*_i−measure zero.

Maxmin Strategies

- The maxmin strategy for player i is $\operatorname*{argmax}_{s_i} \min_{s_{-i}} u_i(s_i; s_{-i})$, and player i's maxmin value is $\max_{s_i} \min_{s_{-i}} u_i(s_i; s_{-i})$.
  - Why would i want to play a maxmin strategy?
  - a conservative agent maximizing worst-case payoff.
- In a two-player game, the minmax strategy for player i against player $-i$ is $\operatorname*{argmin}_{s_i} \max_{s_{-i}} u_{-i}(s_i; s_{-i})$, and player $-i$'s minmax value is $\min_{s_i} \max_{s_{-i}} u_{-i}(s_i; s_{-i})$.
  - Why would i want to play a minmax strategy?
  - to punish the other agent as much as possible.

Theorem (Minimax Theorem — von Neumann 1928)

In any finite, two-player, zero-sum game, in any Nash equilibrium each player receives a payoff that is equal to both his maxmin value and his minmax value.

Let Ω be the uncertainty space, $\mathcal{I}_i$ be the information partition of player i, $P(\cdot|\mathcal{I}_i) \in \Delta\Omega$ be the interim belief systems, and $v_i : \Omega \to S_i$ be measurable with regard to $\mathcal{I}_i$. Then $(v_i)_{i \in N}$ is a posteriori equilibrium of the strategic game (N, S_i, u_i) if for all $i \in N$ and $s_i \in S_i$:

$$\sum_{\omega \in \Omega} P(\omega|\mathcal{I}_i(\omega)) u_i(v_i(\omega), v_{-i}(\omega)) \geq \sum_{\omega \in \Omega} P(\omega|\mathcal{I}_i(\omega)) u_i\left(s_i, v_{-i}(\omega)\right)$$

Subgame

某个博弈的子博弈是由该博弈的某个单一结点（即该结点所在的信息集就只包含一个要素，就是该结点）及其所有后续结点构成的集合，并且所有的后续结点必须满足一个条件，即如果某一个后续结点属于该子博弈，那么该后续结点所在信息集的所有结点必须也属于该子博弈。

Definition (Subgame)

The *subgame* of the extensive game Γ that follows the history h is the extensive game $\Gamma(h) := \langle N, P|_h, \mathcal{H}|_h, (\mathcal{I}_i)_{i\in N}, f_c, (u_i|_h)_{i\in N}\rangle$, where

$$P|_h(h') := P(hh')$$

$$\mathcal{H}|_h := \{h' : hh' \in \mathcal{H}\}$$

$$u_i|_h(h') := u_i(hh')$$

$$\forall I \in \bigcup_{i\in N} \mathcal{I}_i : \ I \subset \{hh' : h' \in \mathcal{H}|_h\} \ \vee \ I \subset \mathcal{H} \setminus \{hh' : h' \in \mathcal{H}|_h\}$$

Subgame Perfect Equilibrium

Definition (Subgame Perfect Equilibrium)

A *subgame perfect equilibrium* of an extensive game Γ is a strategy profile s^* such that for every player $i \in N$ and every nonterminal history $h \in \mathcal{H} \setminus \mathcal{Z}$ for which $P(h) = i$ we have

$$u_i|_h(s_i^*; s_{-i}^*) \geq u_i|_h(s_i; s_{-i}^*)$$

for every strategy s_i of player i in the subgame $\Gamma(h)$.
A behaviour strategy profile ρ^* is a *subgame perfect equilibrium* of a game Γ iff for all subgames $\Gamma(h)$ of Γ, and,

$$\forall i \in N \forall \rho_i \in B_i : u_i|_h(\rho_i^*; \rho_{-i}^*) \geq u_i|_h(\rho_i; \rho_{-i}^*)$$

Sequentially Rationality

Definition (Assessment)

An *assessment* in an extensive game is a pair (ρ, μ), where ρ is a profile of behavioural strategies and μ is a belief system. $\mu(I)(h)$ is the probability that player $P(I)$ assigns to the history $h \in I$, conditional on I being reached.

Definition (Sequentially Rationality)

An assessment (ρ, μ) is *sequentially rational* iff for every player i and every information set $I \in \mathcal{I}_i$ the strategy of player i is a best response to the other players' strategies given i's beliefs at I.

$$\forall \rho_i' \in B_i : \sum_{h \in I} u_i|_h(\rho_i; \rho_{-i})\mu(I)(h) \geq \sum_{h \in I} u_i|_h(\rho_i'; \rho_{-i})\mu(I)(h)$$

We can define the outcome $O(\rho, \mu | I)$ as the distribution over terminal histories determined by ρ and μ conditional on I being reached.

$$O(\rho, \mu | I)(h') := \begin{cases} 0 & \text{if } \nexists h \in I \, (h \sqsubset h') \\ \mu(I)(h) \prod_{k=\ell(h)}^{\ell(h')} \rho_{P(h_{<k})}(h_{<k})(h_k) & \text{otherwise} \end{cases}$$

since by perfect recall, there is at most one subhistory h of h' in I, and the histories $h_{1:k}$ for $k = \ell(h), \ldots, \ell(h')$ lie in different information sets.

- An assessment (ρ, μ) is *sequentially rational* iff for every player i and every information set $I \in \mathcal{I}_i$

$$\forall \rho_i' \in B_i : O(\rho, \mu | I) \succcurlyeq_i O((\rho_i'; \rho_{-i}), \mu | I)$$

- A behavioural strategy profile to be *completely mixed* iff it assigns positive probability to every action at every information set.

Definition (Consistency of Beliefs with Strategies)

Let Γ be a finite extensive game with perfect recall. An assessment (ρ, μ) is *consistent* iff there is a sequence $((\rho^n, \mu^n))_{n=1}^{\infty}$ of assessments s.t. $(\rho, \mu) = \lim_{n \to \infty} (\rho^n, \mu^n)$ and each strategy profile ρ^n is completely mixed and each belief system μ^n is derived from ρ^n using Bayes' rule:

$$\mu^n(I)(h) = \frac{\overline{\rho}^n(h)}{\sum_{h \in I} \overline{\rho}^n(h)}$$

Definition (Sequential Equilibrium)

An assessment is a *sequential equilibrium* of a finite extensive game with perfect recall iff it is sequentially rational and consistent.

Theorem

Every finite extensive game with perfect recall has a sequential equilibrium.

Theorem

In an extensive game with perfect information, (ρ, μ) is a sequential equilibrium iff ρ is a subgame perfect equilibrium.

Bayesian Strategic Game with Observable Actions

Definition (Bayesian Strategic Game with Observable Actions)

A Bayesian game is a tuple $\langle N, (S_i)_{i \in N}, (\Theta_i)_{i \in N}, p, (u_i)_{i \in N} \rangle$ where

- The sets of players N.
- The sets of strategies of the players $(S_i)_{i \in N}$.
- Θ_i is a finite set (the set of possible types of player i). $\Theta := \prod_{i \in N} \Theta_i$.
- $p \in \Delta\Theta$.
- $u_i : S \times \Theta \to \mathbb{R}$

Expected Utility

1. ex-post — the agent knows all agents' types.
2. ex-interim — an agent knows her own type but not the types of the other agents.
3. ex-ante — the agent knows nothing about anyone's actual type.

Ex-post Expected Utility

Definition (Ex-post Expected Utility)

Player i's ex post expected utility in a Bayesian game is defined as

$$V_i(\sigma, \theta) := \sum_{s \in S} \left(\prod_{j \in N} \sigma_j(s_j | \theta_j) \right) u_i(s, \theta)$$

where the players' types are given by $\theta \in \Theta$, and the players' strategies are given by $\sigma \in \prod_{i \in N} \prod_{\theta_i \in \Theta_i} \Delta S_i$, $\sigma = (\sigma_i(\cdot|\theta_i))_{\theta_i \in \Theta_i, i \in N}$, $\sigma_i(\cdot|\theta_i) \in \Delta S_i$.

Ex-interim Expected Utility

Definition (Ex-interim Expected Utility)

Player i's ex interim expected utility in a Bayesian game, where i's type is θ_i and where the players' strategies are given by the mixed-strategy profile σ, is defined as

$$V_i(\sigma, \theta_i) := \sum_{\theta_{-i} \in \Theta_{-i}} p(\theta_{-i}|\theta_i) \sum_{s \in S} \left(\prod_{j \in N} \sigma_j(s_j|\theta_j) \right) u_i(s, (\theta_i; \theta_{-i}))$$

or equivalently as

$$V_i(\sigma, \theta_i) = \sum_{\theta_{-i} \in \Theta_{-i}} p(\theta_{-i}|\theta_i) V_i(\sigma, (\theta_i; \theta_{-i}))$$

Ex-ante Expected Utility

Definition (Ex-ante Expected Utility)

Player i's ex ante expected utility in a Bayesian game, where the players' strategies are given by the mixed-strategy profile σ, is defined as

$$V_i(\sigma) := \sum_{\theta \in \Theta} p(\theta) \sum_{s \in S} \left(\prod_{j \in N} \sigma_j(s_j | \theta_j) \right) u_i(s, \theta)$$

or equivalently as

$$V_i(\sigma) = \sum_{\theta \in \Theta} p(\theta) V_i(\sigma, \theta)$$

or again equivalently as

$$V_i(\sigma) = \sum_{\theta_i \in \Theta_i} p(\theta_i) V_i(\sigma, \theta_i)$$

Best Response in a Bayesian Game

Definition (Best Response in a Bayesian Game)

The set of player i's best response to mixed-strategy profile σ_{-i} are given by

$$\mathrm{BR}_i(\sigma_{-i}) := \operatorname*{argmax}_{\sigma_i \in \Delta S_i^{\Theta_i}} V_i(\sigma_i; \sigma_{-i})$$

Remark: It may seem odd that BR_i is calculated based on i's ex ante expected utility. However, we are in fact performing independent maximization of i's *ex interim expected utilities* conditioned on each type that he could have.
Intuitively speaking, if a certain action is best after the signal is received, it is also the best conditional plan devised ahead of time for what to do should that signal be received.

Definition (Bayesian Equilibrium)

A Bayesian equilibrium is a mixed strategy profile σ^* that satisfies

$$\forall i \in N : \sigma_i^* \in \mathrm{BR}_i(\sigma_{-i}^*)$$

Definition (Ex-post Equilibrium)

An ex post equilibrium is a mixed strategy profile σ that satisfies

$$\forall \theta \in \Theta \forall i \in N : \sigma_i \in \underset{\sigma_i \in \Delta S_i^{\Theta_i}}{\operatorname{argmax}} V_i((\sigma_i; \sigma_{-i}), \theta)$$

The ex post equilibrium is similar in flavor to equilibria in dominant strategies, which do not require agents to believe that other agents act rationally. However, it seems too good to be true most of the time.

Bayesian Extensive Game with Observable Actions

Definition (Bayesian Extensive Game with Observable Actions)

A *Bayesian extensive game with observable actions* is a tuple $\langle \Gamma, (\Theta_i)_{i \in N}, (p_i)_{i \in N}, (u_i)_{i \in N} \rangle$ where

- $\Gamma = \langle N, P, \mathcal{H} \rangle$ is an extensive game form with perfect information and simultaneous moves.
- Θ_i is a finite set (the set of possible types of player i). $\Theta := \prod_{i \in N} \Theta_i$.
- $p_i \in \Delta^0(\Theta_i)$.
- $u_i : \mathcal{H} \times \Theta \to \mathbb{R}$

Remark: the type of player encapsulates all the information possessed by the player that is not common knowledge. This is often quite simple (e.g. the player's knowledge of his private payoff function), but can also include his beliefs about other players' payoffs, about their beliefs about his own payoff, and any other higher-order beliefs.

There is a way of capturing the common prior is to hypothesize a special agent c called "Nature" who makes probabilistic choices. The agent "nature"(c) selects the types of the players, who are subsequently fully cognizant at all points of all moves taken previously. We can associate with any such game an extensive game (with imperfect information and simultaneous moves) in which the set of histories is $\{\emptyset\} \cup (\mathcal{H} \times \Theta)$ and each information set of each player i takes the form

$$I(h, \theta_i) := \{(h, (\theta_i, \theta_{-i})) : \theta_{-i} \in \Theta_{-i}\}$$

for $i \in P(h)$ and $\theta_i \in \Theta_i$.

Let s be a profile of behavioural strategies in Γ. Define $O_h(s)$ to be the probability measure on the set of terminal histories of Γ generated by s given that the history h has occurred. Define $O((s_i; \rho_{-i}), u_{-i}|h)$ to be the probability measure on the set of terminal histories of Γ given that player i uses the strategy s_i in Γ, each type θ_j of each player j uses the strategy $\rho_j(\theta_j)$, the game has reached h, and the probability that i assigns to θ_{-i} is derived from $\mu_{-i}(h)$. That is, $O((s_i; \rho_{-i}), u_{-i}|h)$ is the compound lottery in which the probability of the lottery $O_h(s_i; (\rho_j(\theta_j))_{j \in N \setminus \{i\}})$ is $\prod_{j \in N \setminus \{i\}} \mu_j(h)(\theta_j)$ for each $\theta_{-i} \in \Theta_{-i}$.

Ex-ante Expected Utility

Definition (Ex-ante Expected Utility)

At history h, the player i's ex ante expected utility in a Bayesian game, where the players' behavioural strategy profile is ρ, is defined as

$$V_i(\rho, h) := \sum_{\theta_i \in \Theta_i} \mu_i(h)(\theta_i) \sum_{\theta_{-i} \in \Theta_{-i}} u_i|_h(\rho, \theta) \prod_{j \in N \setminus \{i\}} \mu_j(h)(\theta_j)$$

Perfect Bayesian Equilibrium

Definition (Perfect Bayesian Equilibrium)

For a Bayesian extensive game with observable actions, $((\rho_i), (\mu_i)) \coloneqq ((\rho_i(\theta_i))_{\theta_i \in \Theta_i}, (\mu_i(h))_{h \in \mathcal{H} \setminus \mathcal{Z}})$ is a *perfect Bayesian equilibrium* of the game iff the following conditions are satisfied, where $\rho_i(\theta_i)$ is a behavioural strategy of player $i \in N$ and $\mu_i(h) \in \Delta\Theta_i$.

Perfect Bayesian Equilibrium — definition continued

(I) *Sequential rationality* For every terminal history $h \in \mathcal{H} \setminus \mathcal{Z}$, every player $i \in P(h)$, and every $\theta_i \in \Theta_i$ the probability measure $O((\rho_i(\theta_i); \rho_{-i}), \mu_{-i}|h)$ is at least good for type θ_i as $O((s_i; \rho_{-i}), \mu_{-i}|h)$ for any strategy s_i of player i in Γ.

$$\forall s_i \in S_i : O((\rho_i(\theta_i); \rho_{-i}), \mu_{-i}|h) \succcurlyeq_i O((s_i; \rho_{-i}), \mu_{-i}|h)$$

or equivalently, for $\rho_i' \in B_i$,

$$\sum_{\theta_{-i} \in \Theta_{-i}} u_i|_h((\rho_i; \rho_{-i}), \theta) \prod_{j \in N \setminus \{i\}} \mu_j(h)(\theta_j) \geq \sum_{\theta_{-i} \in \Theta_{-i}} u_i|_h((\rho_i'; \rho_{-i}), \theta) \prod_{j \in N \setminus \{i\}} \mu_j(h)(\theta_j)$$

(II) *Correct initial beliefs* $\forall i \in N : \mu_i(\emptyset) = p_i$

(III) *Action-determined beliefs* If $i \notin P(h)$ and $a \in A(h)$ then $\mu_i(ha) = \mu_i(h)$;
if $i \in P(h), a \in A(h), a' \in A(h)$ and $a = a'$ then $\mu_i(ha) = \mu_i(ha')$.

(IV) *Bayesian updating* If $i \in P(h)$ and $\exists \theta_i \in \mathrm{supp}(\mu_i(h))\ (a \in \mathrm{supp}(\rho_i(\theta_i)(h)))$ then for any $\theta' \in \Theta_i$ we have

$$\mu_i(ha)(\theta_i') := \frac{\rho_i(\theta_i')(h)(a) \cdot \mu_i(h)(\theta_i')}{\sum_{\theta_i \in \Theta_i} \rho_i(\theta_i)(h)(a) \cdot \mu_i(h)(\theta_i)} \qquad \text{(Bayesian Update)}$$

The condition of Bayesian updating relates to a case in which player i's action at the history h is consistent with the other players' beliefs about player i at h, given ρ_i. In such a case the condition requires not only that the new belief depend only on player i's action (as required by the condition of action-determined beliefs) but also that the players' beliefs be derived via Bayes' rule from their observation of player i's actions. Thus the players update their beliefs about player i using Bayes' rule until his behaviour contradicts his strategy ρ_i, at which point they form a new conjecture about player i's type that is the basis for future Bayesian updating until there is another conflict with ρ.

Theorem

Let $(\rho; \mu)$ be a sequential equilibrium of the extensive game associated with the finite Bayesian extensive game with observable actions. For every $h \in \mathcal{H}$, $i \in P(h)$, and $\theta_i \in \Theta_i$, let $\rho'_i(\theta_i)(h) := \rho_i(I(h, \theta_i))$. Then there is a collection $(\mu'_i(h))_{i \in N, h \in \mathcal{H}}$, where $\mu'_i(h) \in \Delta\Theta_i$, such that

$$\mu(I(h, \theta_i))(h, \theta) = \prod_{j \in N \setminus \{i\}} \mu'_j(h)(\theta_j) \quad \text{for all } \theta \in \Theta \text{ and } h \in \mathcal{H}$$

and $((\rho'_i), (\mu'_i))$ is a perfect Bayesian equilibrium of the Bayesian extensive game.

Equilibrium

Definition (Nash Euqilibrium)

- s^* is a *pure Nash equilibrium* iff $s^* \in \prod_{i \in N} \operatorname*{argmax}_{s_i \in S_i} u_i(s_i; s_{-i})$.
- σ^* is a *mixed Nash equilibrium* iff

$$\forall i \in N \forall \sigma_i \in \Delta S_i : u_i(\sigma_i^*; \sigma_{-i}^*) \geq u_i(\sigma_i; \sigma_{-i}^*)$$

Definition (Correlated Equilibrium)

Let Ω be the state space, $\mathcal{I}_i$ be the information partition of player i, $P(\cdot|\mathcal{I}_i) \in \Delta\Omega$ be the interim belief systems, and $\sigma_i : \Omega \to S_i$ be measurable with regard to $\mathcal{I}_i$. Then $(\sigma_i)_{i \in N}$ is a posteriori equilibrium of the strategic game (N, S_i, u_i) iff

$$\forall i \in N \forall s_i \in S_i : \sum_{\omega \in \Omega} P\big(\omega|\mathcal{I}_i(\omega)\big)\Big(u_i\big(\sigma_i(\omega); \sigma_{-i}(\omega)\big) - u_i\big(s_i; \sigma_{-i}(\omega)\big)\Big) \geq 0$$

Remark: For every Nash equilibrium there exists a corresponding correlated equilibrium.

Evolutionarily Stable Strategy

An evolutionarily stable strategy (ESS) is a strategy that, when used by an entire population, is immune against invasion by a minority of mutants playing a different strategy.

Evolutionarily Stable Strategy

Given a symmetric two-player normal form game, σ^* is an ESS iff $\forall \sigma \neq \sigma^* \exists \delta \in (0,1) \forall \varepsilon \in (0,\delta)$:

$$u\left(\sigma^*, (1-\varepsilon)\sigma^* + \varepsilon\sigma\right) > u\left(\sigma, (1-\varepsilon)\sigma^* + \varepsilon\sigma\right)$$

iff

$$(1-\varepsilon)u\left(\sigma^*, \sigma^*\right) + \varepsilon u\left(\sigma^*, \sigma\right) > (1-\varepsilon)u\left(\sigma, \sigma^*\right) + \varepsilon u\left(\sigma, \sigma\right)$$

iff

- $u\left(\sigma^*, \sigma^*\right) > u\left(\sigma, \sigma^*\right)$ or
- $u\left(\sigma^*, \sigma^*\right) = u\left(\sigma, \sigma^*\right)$ and $u\left(\sigma^*, \sigma\right) > u\left(\sigma, \sigma\right)$

If σ is an ESS, then (σ, σ) is a Nash equilibrium. If (σ, σ) is a strict Nash equilibrium, then σ is an ESS.

	dove	hawk
Dove	$\frac{v}{2}, \frac{v}{2}$	$0, v$
Hawk	$v, 0$	$\frac{v-c}{2}, \frac{v-c}{2}$

$$\frac{v}{2}x + 0(1-x) = vx + \frac{v-c}{2}(1-x)$$

$$c > v \implies (1 - \frac{v}{c}, \frac{v}{c})$$

$$c \leq v \implies (H, h)$$

Evolutionarily Stable Strategy

- strict Nash Equilibrium: $u(\sigma_i^*; \sigma_{-i}^*) > u(\sigma_i; \sigma_{-i}^*)$
- Nash Equilibrium: $u(\sigma_i^*; \sigma_{-i}^*) \geq u(\sigma_i; \sigma_{-i}^*)$
- ESS:

  - $u(\sigma^*, \sigma^*) > u(\sigma, \sigma^*)$ or
  - $u(\sigma^*, \sigma^*) = u(\sigma, \sigma^*)$ and $u(\sigma^*, \sigma) > u(\sigma, \sigma)$

- weak ESS:

  - $u(\sigma^*, \sigma^*) > u(\sigma, \sigma^*)$ or
  - $u(\sigma^*, \sigma^*) = u(\sigma, \sigma^*)$ and $u(\sigma^*, \sigma) \geq u(\sigma, \sigma)$

- unbeatable strategy: $u(\sigma^*, \sigma^*) > u(\sigma, \sigma^*)$ and $u(\sigma^*, \sigma) > u(\sigma, \sigma)$

unbeatable $\implies$ strict Nash $\implies$ ESS $\implies$ weak ESS $\implies$ Nash

Replicator Dynamics in Symmetric Games

- Suppose players choose one of m actions. The payoff of a player when he plays a_i and the opponent plays a_j is $u(a_i; a_j)$, where $u(a_i; a_j) \geq 0$.
- It is assumed that individuals use pure strategies.
- Let $p_{i,n}$ be the proportion of individuals using action i in generation n.
- The average reward of an individual using action i in generation n is
$$\overline{R}_{i,n} := u(a_i; p_{1,n}a_1 + \cdots + p_{m,n}a_m)$$
- The average reward in the population as a whole in generation n is
$$\overline{R}_n := \sum_{i=1}^{m} p_{i,n}\overline{R}_{i,n}$$
- The proportion of individuals using action i in generation $n+1$ is
$$p_{i,n+1} := \frac{p_{i,n}\overline{R}_{i,n}}{\overline{R}_n}$$
- A fixpoint $(p_1, \ldots, p_m)$ of the replicator dynamic equations satisfies
$$p_i = \frac{p_i\overline{R}_i}{\overline{R}} \quad \text{for } i = 1, \ldots, m$$

Example

	A	B
A	5,5	0,0
B	0,0	1,1

- Let p_n be the proportion of individuals using A in generation n.
- The average reward of A players in generation n is
$$\overline{R}_{A,n} = u(A; p_n A + (1 - p_n)B) = 5p_n$$
- The average reward of B players in generation n is
$$\overline{R}_{B,n} = u(B; p_n A + (1 - p_n)B) = 1 - p_n$$
- The average reward of the population is
$$\overline{R}_n = p_n \overline{R}_{A,n} + (1 - p_n)\overline{R}_{B,n} = 1 - 2p_n + 6p_n^2$$
- The equation governing the replicator dynamics is
$$p_{n+1} = \frac{p_n \overline{R}_{A,n}}{\overline{R}_n} = \frac{5p_n^2}{1 - 2p_n + 6p_n^2}$$
- A fixpoint of these dynamics satisfies $p = \frac{5p^2}{1-2p+6p^2}$. $p = 0, 1, \frac{1}{6}$.
- $p = 0, 1$ is an attractor. $p = \frac{1}{6}$ is not an attractor.

Fixpoints of the Replicator Dynamic Equations

- A fixed point σ^* is *stable* (also called Lyapunov stable) iff for all open neighborhoods U of σ^* there is another open neighborhood $O \subset U$ such that any trajectory initially inside O remains inside U.
- A fixed point σ^* is *attractive* iff there exists an open neighborhood U of σ^* such that all trajectory initially in U converges to σ^*. The maximum possible U is called the basin of attraction of σ^*.
- A fixed point σ^* is *asymptotically stable* (also called attractor) iff it is stable and attractive.
- A fixed point is *globally asymptotically stable* iff its basin of attraction encompasses the whole space.

- Strict Nash Equilibria are attractors.
- If a fixpoint is stable then it is an Nash Equilibrium.
- ESSs are attractors.
- For 2×2 matrix games a fixed point is an ESS iff it is an attractor.

Braess Paradox

If we all go for the blonde and block each other, not a single one of us is going to get her. So then we go for her friends, but they will all give us the cold shoulder because no one likes to be second choice. But what if none of us goes for the blonde?

— A Beautiful Mind

- The addition of options is not necessarily a good thing.
- A strategy profile is Pareto optimal iff no other strategy profile improves the payoff to at least one actor without decreasing the payoff of other actors.
- The Nash equilibrium of a game is not necessarily Pareto optimal.

Figure: 4000 cars travelling around the lake

Braess Paradox

Let $L_e(x)$ be the travel time of each car traveling along edge e when x cars take that edge ($L_e(0) := 0$). Suppose there is a traffic graph G with x_e cars along edge e. Let $E(e) := \sum_{i=1}^{x_e} L_e(i)$, and $E(G) := \sum_{e \in G} E(e)$. Take a choice of routes that minimizes the total energy $E(G)$. That will be a Nash equilibrium.

Sperner's Lemma

Lemma (Sperner's Lemma)

Suppose that some triangle with vertices V_1, V_2, V_3 is triangulated.
The vertices in the triangulation get "colors" from $\{1, 2, 3\}$ s.t. vertices on the edge (V_i, V_j) are colored either i or j, while the interior vertices are colored 1, 2 or 3.
Then in the triangulation there must be an odd number of "tricolored" triangles.

Proof of Sperner's Lemma

Proof.

Consider the dual graph to the triangulation — but take only those which cross an edge that has endvertices with the colors 1 and 2. Thus we get a partial dual graph which has degree 1 at all vertices that correspond to tricolored triangles, degree 2 for all triangles in which the two colors 1 and 2 appear, and degree 0 for triangles that do not have both colors 1 and 2.

The vertex of the dual graph which corresponds to the outside of the triangulation has odd degree: along the big edge from V_1 to V_2, there is an odd number of changes between 1 and 2. Since the number of odd-degree vertices in any finite graph is even, the number of tricolored triangles is odd.

Brouwer Fixpoint Theorem

Theorem (Brouwer Fixpoint Theorem)

Given a convex compact set $B \subset \mathbb{R}^n$ and continuous function $f : B \to B$, there exists $\mathbf{x}^*$ s.t. $f(\mathbf{x}^*) = \mathbf{x}^*$.

Theorem (Kakutani Fixpoint Theorem)

Given a compact convex set $S \subset \mathbb{R}^n$ and function $f : S \to \mathrm{P}(S)$ for which

- *for all $x \in S$ the set $f(x)$ is nonempty and convex,*
- *the graph of f is closed (i.e. for all sequences $\{x_n\}$ and $\{y_n\}$ s.t. $y_n \in f(x_n)$ for all n, $x_n \to x$, and $y_n \to y$, we have $y \in f(x)$).*

Then there exists $x^* \in S$ s.t. $x^* \in f(x^*)$.

Theorem (Schauder Fixpoint Theorem)

If K is a nonempty convex subset of a Hausdorff topological vector space V and T is a continuous mapping of K into itself such that $T(K)$ is contained in a compact subset of K, then T has a fixpoint.

Proof of Brouwer Fixpoint Theorem

Proof.

Let Δ be the triangle in $\mathbb{R}^3$ with vertices $\mathbf{e}_1 = (1,0,0)$, $\mathbf{e}_2 = (0,1,0)$, and $\mathbf{e}_3 = (0,0,1)$. We prove that any continuous map $f : \Delta \to \Delta$ has a fixpoint.
Let $\delta(T)$ be the maximal length of an edge in a triangulation T.
One can construct an infinite sequence of triangulations $T_1, T_2, \ldots$ of Δ s.t. $\lim\limits_{k\to\infty} \delta(T_k) = 0$.
Suppose f has no fixpoint. Since $\sum_i \mathbf{v}_i = 1 = \sum_i f(\mathbf{v})_i$, for each of these triangulations, we can define a Sperner coloring of their vertices $\mathbf{v}$ by setting $\lambda(\mathbf{v}) := \min\{i : f(\mathbf{v})_i < \mathbf{v}_i\}$.
Sperner's lemma tells us that in each triangulation T_k there is a tricolored triangle $\left\{\mathbf{v}_1^k, \mathbf{v}_2^k, \mathbf{v}_3^k\right\}$ with $\lambda(\mathbf{v}_i^k) = i$.
Since the simplex Δ is compact, some subsequence of $\left(\mathbf{v}_1^k\right)_{k\geq 1}$ has a limit point $\mathbf{v}^* \in \Delta$. Since $\lim\limits_{k\to\infty} \delta(T_k) = 0$, the sequences $\mathbf{v}_2^k$ and $\mathbf{v}_3^k$ converge to the same point $\mathbf{v}^*$.
Then $\forall i : f(\mathbf{v}^*)_i \leq \mathbf{v}_i^*$, which contradicts $f(\mathbf{v}^*) \neq \mathbf{v}^*$.

Nash Equilibrium

Theorem (Existence of Mixed Nash Equilibrium)

Every finite strategic game has a mixed Nash equilibrium.

Proof.

Given a strategy profile $\sigma \in \prod_{i \in N} \Delta S_i$, define

$$\varphi_{i,s_i}(\sigma) := \max\left\{0, u_i(s_i; \sigma_{-i}) - u_i(\sigma)\right\}$$

Then define a continuous $f : \prod_{i \in N} \Delta S_i \to \prod_{i \in N} \Delta S_i$ by $f : \sigma \mapsto \sigma'$, where

$$\sigma_i'(s_i) := \frac{\sigma_i(s_i) + \varphi_{i,s_i}(\sigma)}{\sum_{s_i \in S_i} \left[\sigma_i(s_i) + \varphi_{i,s_i}(\sigma)\right]} = \frac{\sigma_i(s_i) + \varphi_{i,s_i}(\sigma)}{1 + \sum_{s_i \in S_i} \varphi_{i,s_i}(\sigma)}$$

Since $\prod_{i \in N} \Delta S_i$ is convex and compact, f has a fixpoint.

Consider any fixpoint σ of f. By the linearity of expectation there exists s_i' in the support of σ, for which $u_i(s_i'; \sigma_{-i}) \leq u_i(\sigma)$. Then

$$\varphi_{i,s_i'}(\sigma) = 0 \ \ \& \ \ \sigma_i'(s_i') = \sigma_i(s_i') \implies \forall i \in N \forall s_i \in S_i : \varphi_{i,s_i}(\sigma) = 0$$

Walrasian Equilibrium

Theorem (Existence of Walrasian Equilibrium)

Consider an economy with n goods $X_1, \ldots, X_n$ with a price vector $(p_1, \ldots, p_n) \in \Delta_n := \{x \in [0,1]^n : \|x\|_1 = 1\}$, and the prices of at least two goods are not zero. Assume that an excess demand function for each good $f_i(p_1, \ldots, p_n)$ is continuous and satisfies the following condition

$$\sum_{i=1}^{n} p_i f_i = 0 \qquad \text{(Walras Law)}$$

Then, there exists an equilibrium price vector $(p_1^*, \ldots, p_n^*)$ s.t.

$$f_i(p_1^*, \ldots, p_n^*) \leq 0$$

for all $i = 1, \ldots, n$. And when $p_i > 0$ we have $f_i(p_1^*, \ldots, p_n^*) = 0$.

Contents

History of AI

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Reinforcement Learning

General Reinforcement Learning

Markov Decision Process

Figure: MDP $(\mathcal{S}, \mathcal{A}, P, R)$.

Definition (Value of a state under π)

$$V^{\pi}(s) := \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \,\middle|\, S_t = s\right]$$

Definition (Action-value under π)

$$Q^{\pi}(s,a) := \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \,\middle|\, S_t = s, A_t = a\right]$$

Bellman Expectation Equations

$$r(s,a) := \mathbb{E}[r|s,a]$$

$$V^\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s) Q^\pi(s,a)$$

$$Q^\pi(s,a) = r(s,a) + \gamma \sum_{s'\in\mathcal{S}} P(s'|s,a) V^\pi(s')$$

$$V^\pi(s) = \mathbb{E}\left[r + \gamma V^\pi(s') \mid s\right] = \sum_{a\in\mathcal{A}} \pi(a|s) \left(r(s,a) + \gamma \sum_{s'\in\mathcal{S}} P(s'|s,a) V^\pi(s') \right)$$

advantage: $A^\pi(s,a) := Q^\pi(s,a) - V^\pi(s)$

Bellman Optimality Equations

Definition (Optimal Values)

$$V^*(s) := \max_{\pi} V^{\pi}(s)$$

$$Q^*(s,a) := \max_{\pi} Q^{\pi}(s,a)$$

Definition (Optimal Policy)

A policy π is called optimal iff $\forall s \in \mathcal{S} : V^{\pi}(s) = V^*(s)$.

$$V^*(s) = \max_{a \in \mathcal{A}} Q^*(s,a)$$

$$Q^*(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V^*(s')$$

Action/Policy Evaluation Operator & Greedy Policy

Definition (Action Evaluation Operator)

$$T_a V(s) := r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V(s')$$

Definition (Policy Evaluation Operator)

$$T^\pi V(s) := \sum_{a \in \mathcal{A}} \pi(a|s) T_a V(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V(s') \right)$$

$$T^* V(s) := \max_{a \in \mathcal{A}} T_a V(s) = \max_{a \in \mathcal{A}} \left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V(s') \right)$$

Definition (Greedy policy)

Policy π is greedy w.r.t. V iff $T^\pi V = T^* V$.

Banach's Fixpoint Theorem

Theorem (Banach's Fixpoint Theorem)

Let $\mathcal{V}$ be a Banach space and $T : \mathcal{V} \to \mathcal{V}$ be a contraction mapping, with Lipschitz constant $\gamma < 1$. Then T has a unique fixpoint $v \in \mathcal{V}$. Further, for each $v_0 \in \mathcal{V}$, $\lim_{n\to\infty} \|T^n(v_0) - v\| = 0$, and the convergence is geometric:

$$\|T^n(v_0) - v\| \leq \gamma^n \|v_0 - v\|$$

Application of Banach's Fixpoint Theorem

Theorem

$(\mathcal{V}, \|\cdot\|_\infty)$ *is a Banach space, where* $\mathcal{V} := \left\{V \in \mathbb{R}^{\mathcal{S}} : \|V\|_\infty < \infty\right\}$ *and* $\|V\|_\infty := \max_{s \in \mathcal{S}} |V(s)|$.

- T^π is a contraction, and V^π is the unique fixpoint of T^π.

$$\lim_{n\to\infty} \|(T^\pi)^n V_0 - V^\pi\|_\infty = 0$$

- T^* is a contraction, and V^* is the unique fixpoint of T^*.

$$\lim_{n\to\infty} \|(T^*)^n V_0 - V^*\|_\infty = 0$$

Application of Banach's Fixpoint Theorem

$$T^\pi Q(s,a) := r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) \sum_{a' \in \mathcal{A}} \pi(a'|s') Q(s',a')$$

$$T^* Q(s,a) := r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) \max_{a' \in \mathcal{A}} Q(s',a')$$

- T^π is a contraction, and Q^π is the unique fixpoint of T^π.
$$\lim_{n\to\infty} \|(T^\pi)^n Q_0 - Q^\pi\|_\infty = 0$$
- T^* is a contraction, and Q^* is the unique fixpoint of T^*.
$$\lim_{n\to\infty} \|(T^*)^n Q_0 - Q^*\|_\infty = 0$$

Two Theorems

Theorem (Fixpoint of Bellman Optimality Operator)

Let V be the fixpoint of T^* and assume that there is policy π which is greedy w.r.t V. Then $V = V^*$ and π is an optimal policy.

Theorem (Policy Improvement Theorem)

Choose some stationary policy π_0 and let π be greedy w.r.t. V^{π_0}. Then $V^{\pi} \geq V^{\pi_0}$, i.e., π is an improvement upon π_0. In particular, if $T^*V^{\pi_0}(s) > V^{\pi_0}(s)$ for some state s then π strictly improves upon π_0 at $s : V^{\pi}(s) > V^{\pi_0}(s)$. On the other hand, when $T^*V^{\pi_0}(s) = V^{\pi_0}(s)$ then π_0 is an optimal policy.

Solving MDPs — Finite-Horizon Dynamic Programming

Principle of optimality: the tail of an optimal policy is optimal for the "tail" problem.

Backward Induction

- Backward recursion: $V_N^*(s) = r_N(s)$ and for $k = N-1, \ldots, 0$

$$V_k^*(s) = \max_{a \in \mathcal{A}_k} \left(r_k(s,a) + \sum_{s' \in \mathcal{S}_{k+1}} P_k(s'|s,a) V_{k+1}^*(s') \right)$$

- Optimal policy: for $k = 0, \ldots, N-1$

$$\pi_k^*(s) \in \operatorname*{argmax}_{a \in \mathcal{A}_k} \left(r_k(s,a) + \sum_{s' \in \mathcal{S}_{k+1}} P_k(s'|s,a) V_{k+1}^*(s') \right)$$

- Cost: $N|\mathcal{S}||\mathcal{A}|$ vs $|\mathcal{A}|^{N|\mathcal{S}|}$ of brute force policy search.
- From now on, we will consider infinite-horizon discounted MDPs.

Example — Dynamic Programming

$$D_E = \min(d(A,C) + d(C,E), d(A,D) + d(D,E)) = 3$$
$$D_F = \min(d(A,C) + d(C,F), d(A,D) + d(D,F)) = 2$$
$$D_G = \min(D_E + d(E,G), D_F + d(F,G)) = 5$$
$$D_H = \min(D_E + d(E,H), D_F + d(F,H)) = 4$$
$$D_B = \min(D_G + d(G,B), D_H + d(H,B)) = 5$$

Working backward, the "best" path from A to B is A, D, E, H, B.

Solving MDPs — Value Iteration

Theorem (Principle of Optimality)

A policy π achieves the optimal value from state s, $V^\pi(s) = V^*(s)$, iff, for any state s' reachable from s, π achieves the optimal value from state s', $V^\pi(s') = V^*(s')$.

Any optimal policy π^* can be subdivided into two components:

- an optimal first action a^*,
- followed by an optimal policy from successor state s'.

$$V^*(s) = \max_{a \in \mathcal{A}} \left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V^*(s') \right)$$

Value Iteration: $V_{k+1} \leftarrow T^* V_k$

$$V_1 \to V_2 \to \cdots \to V^*$$

$$\boxed{\|V_{k+1} - V_k\|_\infty < \varepsilon \implies \|V_{k+1} - V^*\|_\infty < \frac{2\gamma\varepsilon}{1-\gamma}}$$

Solving MDPs — Policy Iteration

Policy Iteration: $\pi_0 \xrightarrow{E} V^{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} V^{\pi_1} \xrightarrow{I} \pi_2 \xrightarrow{E} \cdots \xrightarrow{I} \pi^* \xrightarrow{E} V^*$

- E — policy evaluation:

$$V_{k+1}^{\pi} \leftarrow T^{\pi} V_k^{\pi}$$

- I — policy improvement:

$$\pi_{k+1}(s) := \operatorname*{argmax}_{a \in \mathcal{A}} \left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V^{\pi_k}(s') \right)$$

Monte-Carlo Methods

MC learns from complete episodes of raw experience without modeling the environmental dynamics and computes the observed mean return as an approximation of the expected return.

$$G_t := \sum_{k=0}^{T-t-1} \gamma^k r_{t+k+1}$$

$$V(s) := \frac{\sum_{t=1}^{T} [\![S_t = s]\!] G_t}{\sum_{t=1}^{T} [\![S_t = s]\!]}$$

$$Q(s,a) := \frac{\sum_{t=1}^{T} [\![S_t = s, A_t = a]\!] G_t}{\sum_{t=1}^{T} [\![S_t = s, A_t = a]\!]}$$

Temporal-Difference Learning

TD Learning is model-free and learns from incomplete episodes of experience.

$$V(s_t) \leftarrow (1-\alpha)V(s_t) + \alpha G_t$$
$$V(s_t) \leftarrow V(s_t) + \alpha(G_t - V(s_t))$$
$$V(s_t) \leftarrow V(s_t) + \alpha(r_{t+1} + \gamma V(s_{t+1}) - V(s_t))$$
$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t))$$

- MC updates value $V(s_t)$ toward actual return G_t.

$$V(s_t) \leftarrow V(s_t) + \alpha(G_t - V(s_t))$$

- TD updates value $V(s_t)$ toward estimated return $r_{t+1} + \gamma V(s_{t+1})$.

$$V(s_t) \leftarrow V(s_t) + \alpha(\underbrace{\overbrace{r_{t+1} + \gamma V(s_{t+1})}^{\text{TD target}} - V(s_t)}_{\text{TD error}})$$

SARSA: On-Policy TD control

SARSA

1. At time step t, we start from state s_t and pick action according to Q values, $a_t = \operatorname*{argmax}_{a \in \mathcal{A}} Q(s_t, a)$; ε-greedy is commonly applied.
2. With action a_t, we observe reward r_{t+1} and get into s_{t+1}.
3. Then pick the next action $a_{t+1} = \operatorname*{argmax}_{a \in \mathcal{A}} Q(s_{t+1}, a)$.
4. Update the action-value function:
   $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t))$.
5. $t = t + 1$ and repeat from step 1.

Expected SARSA

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left(r_{t+1} + \gamma \sum_{a \in \mathcal{A}} \pi(a|s_{t+1}) Q(s_{t+1}, a) - Q(s_t, a_t) \right)$$

Q-Learning: Off-policy TD control

1. At time step t, we start from state s_t and pick action according to Q values, $a_t = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, Q(s_t, a)$; ε-greedy is commonly applied.
2. With action a_t, we observe reward r_{t+1} and get into s_{t+1}.
3. Update the action-value function:
$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left(r_{t+1} + \gamma \max_{a \in \mathcal{A}} Q(s_{t+1}, a) - Q(s_t, a_t) \right).$$
4. $t = t + 1$ and repeat from step 1.

Figure: SARSAR, Expected SARSA, and Q-Learning

Monte-Carlo Backup

$$V(s_t) \leftarrow V(s_t) + \alpha(G_t - V(s_t))$$

Temporal-Difference Backup

$$V(s_t) \leftarrow V(s_t) + \alpha(r_{t+1} + \gamma V(s_{t+1}) - V(s_t))$$

Dynamic Programming Backup

$$V(s_t) \leftarrow \mathbb{E}_\pi[r_{t+1} + \gamma V(s_{t+1})]$$

Contents

History of AI

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Reinforcement Learning
    Artificial Neural Network and Deep Reinforcement Learning

General Reinforcement Learning

McCulloch-Pitts Artificial Neural Network

$$y = g\left(\sum_{i=1}^{n} w_i x_i + b\right)$$

$$a_k = \overbrace{g_k(\overbrace{b_k + \Sigma_j \underbrace{g_j(\underbrace{b_j + \Sigma_i a_i w_{ij}}_{z_j})}_{a_j} w_{jk}}^{z_k})}^{a_k}$$

Learning: small change in weights $\rightarrow$ small change in output

Kolmogorov Superposition Theorem

Theorem (Kolmogorov Superposition Theorem)

For each $n \geq 2$ there exists a computable function $\psi : [0,1] \to \mathbb{R}$ and computable constants $a, \lambda_{pq} \in \mathbb{R}$, $p = 1, \ldots, n$, $q = 0, \ldots, 2n$ s.t.: every continuous function $f : [0,1]^n \to \mathbb{R}$ has a representation as

$$f(x_1, \ldots, x_n) = \sum_{q=0}^{2n} g\left(\sum_{p=1}^{n} \lambda_{pq} \psi(x_p + qa)\right)$$

for some continuous function $g : [0,1] \to \mathbb{R}$ that is computable from f.

Theorem (Hecht-Nielsen Theorem)

The class of functions $f : [0,1]^n \to \mathbb{R}$, implementable by three-layer feed-forward neural networks with (computable) continuous activation functions $g : [0,1] \to \mathbb{R}$ and (computable) weights $\lambda \in \mathbb{R}$, is exactly the class of (computable) continuous functions $f : [0,1]^n \to \mathbb{R}$.

$y_{q,p} = \psi(x_p + q\,a)$
$X_q = g(y_{q,1} + \ldots + y_{q,n})$
x_1
x_2
$y_{0,1}$
$y_{1,1}$
$y_{2,1}$
$y_{3,1}$
$y_{4,1}$
$y_{0,2}$
$y_{1,2}$
$y_{2,2}$
$y_{3,2}$
$y_{4,2}$
X_0
X_1
X_2
X_3
X_4
f
input layer
hidden layer
hidden layer
output layer

Universal Approximation Theorem

Theorem (Universal Approximation Theorem)

Let g be a nonconstant, bounded, and increasing continuous function. Let I_n be any compact subset of $\mathbb{R}^n$. The space of continuous functions on I_n is denoted by $C(I_n, \mathbb{R})$. Then, given any function $f \in C(I_n, \mathbb{R})$ and $\varepsilon > 0$, there exists an integer N, real constants $v_i, b_i \in \mathbb{R}$ and real vectors $\mathbf{w}_i \in \mathbb{R}^n$, where $i = 1, \ldots, N$, s.t.

$$\forall \mathbf{x} \in I_n : |h(\mathbf{x}) - f(\mathbf{x})| < \varepsilon$$

where

$$h(\mathbf{x}) := \sum_{i=1}^{N} v_i g\left(\mathbf{w}_i^\top \mathbf{x} + b_i\right)$$

In other words, functions of the form $h(\mathbf{x})$ are dense in $C(I_n, \mathbb{R})$.

Representation & Approximation

- A feed-forward network with 1 hidden layer can represent any boolean function, but require exponential hidden units.
- A feed-forward network with 2 hidden layers and (computable) continuous activation functions can represent any (computable) continuous function.
- A feed-forward network with a linear output layer and at least 1 hidden layer and continuous and differentiable activation functions can approximate any Borel measurable function from one finite-dimensional space to another with any desired non-zero amount of error.
- A feed-forward network with 2 hidden layers and continuous and differentiable activation functions can approximate any function.

Deep Learning

1. hypothesis space — Network Structure — f_θ
2. the goodness of a function — Learning Target — loss function ℓ
3. pick the best function — Learn — find the network parameters $\theta^* := \underset{\theta}{\operatorname{argmin}}\, L(\theta)$ that minimize total cost $L(\theta)$ by gradient decent

$$\theta \leftarrow \theta - \eta \nabla_\theta L(\theta)$$

where $L(\theta) := \mathbb{E}_P \left[\ell\left(f_\theta(\mathbf{a}), t\right)\right] + \lambda \Omega(\theta)$ and $\Omega(\theta)$ is a regularizer.

Deep Learning

$$f_\theta : \mathbf{a}^{(0)} \mapsto \mathbf{g}^{(n)}\Bigg(\underbrace{\cdots \mathbf{g}^{(2)}\Bigg(\overbrace{\mathbf{w}^{(2)}\underbrace{\mathbf{g}^{(1)}\Bigg(\underbrace{\mathbf{w}^{(1)}\mathbf{a}^{(0)} + \mathbf{b}^{(1)}}_{\mathbf{z}^{(1)}}\Bigg)}_{\mathbf{a}^{(1)}} + \mathbf{b}^{(2)}}^{\mathbf{z}^{(2)}}\Bigg)\cdots}_{\mathbf{z}^{(n)}}\Bigg)$$

$$\mathbf{a}^{(2)} \qquad \mathbf{a}^{(n)}$$

where parameters $\theta := \left\{\mathbf{w}^{(i)}, \mathbf{b}^{(i)}\right\}_{i=1}^{n}$ and activation functions $\mathbf{g}$

$$\sigma(z) = \frac{1}{1 - e^{-z}} \qquad \tanh(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}} \qquad \mathrm{ReLU}(z) = \max(0, z)$$

$$\left.\begin{aligned} a_0^{(l)} &:= 1 \\ w_{0j}^{(l)} &:= b_j^{(l)} \end{aligned}\right\} \implies \begin{cases} z_j^{(l+1)} := \sum_i w_{ij}^{(l+1)} a_i^{(l)} \\ a_j^{(l+1)} := g_j^{(l+1)}\left(z_j^{(l+1)}\right) \end{cases}$$

I. Forward-propagate Input Signal

II. Back-propagate Error Signals

III. Calculate Parameter Gradients

IV. Update Parameters

$w_{ij} = w_{ij} - \eta(\partial E/\partial w_{ij})$

$w_{jk} = w_{jk} - \eta(\partial E/\partial w_{jk})$

for learning rate η

Backpropagation

$$\delta_j^{(l)} := \frac{\partial L}{\partial z_j^{(l)}}$$

$$\delta_j^{(l)} = g_j^{(l)\prime}\left(z_j^{(l)}\right) \sum_k \delta_k^{(l+1)} w_{jk}^{(l+1)}$$

$$\frac{\partial L}{\partial w_{ij}^{(l)}} = a_i^{(l-1)} \delta_j^{(l)}$$

$$w_{ij}^{(l)} \leftarrow w_{ij}^{(l)} - \eta \frac{\partial L}{\partial w_{ij}^{(l)}}$$

- network structure?
- how many layers?
- how many units per layer?
- loss function?
- regularization?
- weight decay?
- learning rate?
- activation function?
- early stopping?
- dropout?
- mini-batch?
- momentum?
- ...

$$a_{ij}^{(l+1)} = g_{ij}^{(l+1)} \left(\sum_{m=0}^{k-1} \sum_{n=0}^{k-1} w_{m,n} a_{i+m,j+n}^{(l)} + b \right)$$

Key Properties of CNNs

1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

$*$

1	-1	-1
-1	1	-1
-1	-1	1

$=$

3	-1	-3	-1
-3	1	0	-3
-3	-3	0	1
3	-2	-2	-1

Table: Convolution (stride 1)

Take advantage of the structure of the data!

- Convolutional Filters **(Translation invariance)**
- Multiple layers **(Compositionality)**
- Filters localized in space **(Locality)**
- Weight sharing **(Self-similarity)**

DQN

$$Q^{\pi}(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s, a\right]$$

$$\max_{\pi} Q^{\pi}(s,a) =: Q^*(s,a) = \mathbb{E}_{s'}\left[r + \gamma \max_{a'} Q^*(s',a') \,\middle|\, s, a\right]$$

$$Q_{t+1}(s,a) = \mathbb{E}_{s'}\left[r + \gamma \max_{a'} Q_t(s',a') \,\middle|\, s, a\right]$$

$$Q_t \xrightarrow{t\to\infty} Q^*$$

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha\left(r + \gamma \max_{a'} Q_t(s',a') - Q_t(s,a)\right)$$

$$Q(s,a;\theta) \approx Q^*(s,a)$$

$$L(\theta) := \mathbb{E}_{s,a,r,s'}\left[\left(r + \gamma \max_{a'} Q\left(s',a';\theta^-\right) - Q(s,a;\theta)\right)^2\right] \qquad \text{(DQN)}$$

DQN

$$L(\theta) := \mathbb{E}_{s,a,r,s'}\left[\left(r + \gamma Q\big(s', \operatorname*{argmax}_{a'} Q(s', a'; \theta); \theta^-\big) - Q(s, a; \theta)\right)^2\right]$$

(Double DQN)

$$Q(s, a) = V(s; \theta) + A(s, a; \theta') \qquad \text{(Dueling Network)}$$

$$\begin{cases} \nabla_\theta \log \pi(a_t | s_t; \theta) \left(\sum_{i=0}^{k-1} \gamma^i r_{t+i} + \gamma^k V(s_{t+k}; \theta') - V(s_t; \theta')\right) & \text{actor} \\ \nabla_{\theta'} \left(\sum_{i=0}^{k-1} \gamma^i r_{t+i} + \gamma^k V(s_{t+k}; \theta') - V(s_t; \theta')\right)^2 & \text{critic} \end{cases} \qquad \text{(AC)}$$

AlphaZero

$$(\mathbf{p}, v) = f_\theta(s)$$

$$\ell = (z - v)^2 - \pi^\mathsf{T} \log \mathbf{p} + c\|\theta\|^2$$

Intuition + Calculation

DNN + MCTS

GAN — Generative Adversarial Network

$$V(D,G) = \mathbb{E}_{x\sim P_{data}}\left[\log D(x)\right] + \mathbb{E}_{z\sim P_{noise}}\left[\log\left(1 - D(G(z))\right)\right]$$

$$G^* = \underset{G}{\operatorname{argmin}}\,\max_{D} V(D,G) \qquad \text{(GAN)}$$

Why "Deep" rather than "Fat"?

- Exploiting compositionality gives an exponential gain in representational power.
  - Distributed representations: feature learning
  - Deep architecture: multiple levels of feature learning
- Each basic classifier can be trained by little data.
  - deep $\longrightarrow$ modularization $\longrightarrow$ less training data?
    With more complex features, the number of parameters in the linear layers may be drastically decreased.
  - efficiency & sample complexity
  - better memory/computation trade-off?
- higher-level abstractions $\longrightarrow$ easier generalization & transfer

Minimal Sufficient Statistic

Definition (Sufficient Statistic)

Let Y be a parameter indexing a family of probability distributions. Let X be random variable drawn from a probability distribution determined by Y. $T(X)$ is a sufficient statistic for Y iff X is independent of Y given $T(X)$, i.e., $p(x|t, y) = p(x|t)$.

Definition (Minimal Sufficient Statistic)

A sufficient statistic $S(X)$ is minimal iff for any sufficient statistic $T(X)$, there exists a function f s.t. $S = f(T)$ almost everywhere w.r.t X.

Theorem

- *T is sufficient statistics for Y* $\iff$ $I(T(X); Y) = I(X; Y)$.
- *S is minimal sufficient statistics for Y* $\implies$ $I(X; S(X)) \leq I(X; T(X))$.

Information Bottleneck — Learning is to forget!

Theorem

Let X be a sample drawn according to a distribution determined by the random variable Y. The set of solutions to

$$\min_{T} I(X;T) \quad s.t. \quad I(T;Y) = \max_{T'} I(T';Y)$$

is exactly the set of minimal sufficient statistics for Y based on X.

Find a random variable T s.t.:

- $Y \leftrightarrow X \leftrightarrow T$ form a Markov chain.
- $I(X;T)$ is minimized (minimality, complexity term), while $I(T;Y)$ is maximized (sufficiency, accuracy term).

$$\boxed{T^* := \underset{T:I(T(X);Y)=I(X;Y)}{\operatorname{argmin}} I(X;T(X))}$$

is the Information Bottleneck between X and Y.

张三丰：将所见到的剑招忘得半点不剩，才能得其神髓。

老子：为学日益，为道日损。

Information Bottleneck

$$\min_{p(t|x),p(y|t),p(t)} \left\{ I(X;T) - \beta I(T;Y) \right\} \text{ subject to Markov chain } Y \to X \to T.$$

$$\mathcal{L}[p(t|x)] := I(X;T) - \beta I(T;Y) - \sum_x \lambda(x) \sum_t p(t|x)$$

Let

$$\frac{\delta \mathcal{L}}{\delta p(t|x)} = 0$$

The solution is

$$p(t|x) = \frac{p(t)}{Z(x,\beta)} \mathrm{e}^{-\beta D[p(y|x)\|p(y|t)]}$$

$$p(t) = \sum_x p(t|x)p(x)$$

$$p(y|t) = \sum_x p(y|x)p(x|t)$$

Expressiveness & Sample Complexity

Theorem

The hypothesis class of neural networks of depth T and size $O(T^2)$ contains all functions that can be implemented by a Turing machine within T operations, while having $O(T^2)$ sample complexity.

The Ultimate Hypothesis Space

- No Free Lunch: Sample complexity is exponentially large (w.r.t. the input dimension) if the hypothesis class is all possible functions.
- Shallow learning (SVM, Boosting): Hypothesis class is linear functions over manually determined features — strong prior knowledge.
- Deep learning: Hypothesis class is all functions implemented by determining the weights of a given artificial neural network.

Figure: Prior vs Universality

Prior — a necessary good or a necessary evil?

Contents

History of AI

Philosophy of Induction

Inductive Logic

Universal Induction

Causal Inference

Game Theory

Reinforcement Learning

General Reinforcement Learning

UAI

1. Solve intelligence
2. Use it to solve everything else

- learn automatically from raw inputs — not pre-programmed.
- same algorithm, different tasks.

UAI

(Deep) RL	General RL
state space	history
ergodic	not ergodic
fully observable	partially observable
ε-exploration works	ε-exploration fails
MDP/DQN	AIXI

Table: (Deep) RL vs General RL

Decision Theory	=	Probability + Utility Theory
+		+
Universal Induction	=	Occam + Bayes + Turing
$\|\|$		$\|\|$
Universal Artificial Intelligence without Parameters		

Marcus Hutter

Jan Leike, Tor Lattimore, Shane Legg, Joel Veness, Laurent Orseau, Mark Ring, Peter Sunehag, Mayank Daswani, Tom Everitt, Jan Poland, Daniel Filan, William Uther, Kee Siong Ng, David Silver, Jürgen Schmidhuber, Alexey Potapov, Bill Hibbard, Daniil Ryabko, Alexey Chernov, Michael Cohen. . .

sensors
What the world
is like now
What it will be like
if I do action a
How happy I will be
in such a state
What action
I should do now
Effectors
perception
action
Agent
Environment

Computationalism

Agent & Environment

Definition (Agent & Environment)

- finite set of possible actions $\mathcal{A}$ and perceptions $\mathcal{E}$;
- prior knowledge $w \in \Delta\mathcal{M}$ of the environments $\mathcal{M}$;
- utility function $u : (\mathcal{A} \times \mathcal{E})^* \to [0, 1]$;
- discount factor $\gamma \in [0, 1]$;

$$\pi : (\mathcal{A} \times \mathcal{E})^* \to \Delta\mathcal{A}$$

$$\mu : (\mathcal{A} \times \mathcal{E})^* \times \mathcal{A} \to \Delta\mathcal{E}$$

$$\overset{\pi}{\mu}(æ_{<t}) := \prod_{i=1}^{t-1} \pi(a_i | æ_{<i}) \mu(e_i | æ_{<i} a_i)$$

Value Function

$$r_n := u(æ_{1:n})$$

$$V_\mu^\pi(æ_{<t}) := \mathbb{E}_\mu^\pi\left[\sum_{k=0}^{\infty}\gamma^k r_{t+k}\,\middle|\, æ_{<t}\right]$$

Bellman equation:

$$V_\mu^\pi(æ_{<k}) = \sum_{a_k\in\mathcal{A}}\pi(a_k|æ_{<k})\sum_{e_k\in\mathcal{E}}\mu(e_k|æ_{<k}a_k)\left[r_k+\gamma V_\mu^\pi(æ_{1:k})\right] \qquad (\text{recursive})$$

$$= \sum_{æ_{k:m}}{}_\mu^\pi(æ_{k:m}|æ_{<k})\left[\sum_{i=k}^{m}\gamma^{i-k}r_i+\gamma^{m-k+1}V_\mu^\pi(æ_{1:m})\right] \qquad (\text{iterative})$$

$$V_\mu^\pi(æ_{<k}) = \lim_{m\to\infty}\sum_{æ_{k:m}}{}_\mu^\pi(æ_{k:m}|æ_{<k})\left[\sum_{i=k}^{m}\gamma^{i-k}r_i\right]$$

Optimal Value/Policy

$$V_\mu^* := \max_\pi V_\mu^\pi$$

$$V_\mu^*(æ_{<k}) = \lim_{m\to\infty} \max_{a_k\in\mathcal{A}} \sum_{e_k\in\mathcal{E}} \cdots \max_{a_m\in\mathcal{A}} \sum_{e_m\in\mathcal{E}} \sum_{i=k}^{m} \gamma^{i-k} r_i \prod_{j=k}^{i} \mu(e_j|æ_{<j}a_j)$$

$$= \lim_{m\to\infty} \max_{a_k\in\mathcal{A}} \sum_{e_k\in\mathcal{E}} \cdots \max_{a_m\in\mathcal{A}} \sum_{e_m\in\mathcal{E}} \left[\sum_{i=k}^{m} \gamma^{i-k} r_i\right] \mu(e_{k:m}|æ_{<k}a_{k:m})$$

$$\pi_\mu^* := \operatorname*{argmax}_\pi V_\mu^\pi$$

Bayesian Mixture & Belief Update

$$\xi(e_{<n}|a_{<n}) := \sum_{\nu \in \mathcal{M}} w_\nu \nu(e_{<n}|a_{<n})$$

$$w^{\nu}_{æ_{<n}} := \frac{w_\nu \nu(e_{<n}|a_{<n})}{\xi(e_{<n}|a_{<n})}$$

$$\sum_{k=1}^{\infty} \sum_{e_{1:k}} \mu(e_{<k}|a_{<k}) \Big(\mu(e_k|æ_{<k}a_k) - \xi(e_k|æ_{<k}a_k) \Big)^2 \leq \min_{\nu \in \mathcal{M}} \Big\{ -\ln w_\nu + D(\mu \| \nu) \Big\}$$

What probability should an observer assign to future experiences if she is told that she will be simulated on a computer?

What is 'intelligence'?

A Blind Man in a Dark Room Looking for a Black Cat That Is Not There?

Intelligence measures an agent's ability to achieve goals in a wide range of environments.

— *Shane Legg and Marcus Hutter*

$$\Upsilon(\pi) := \sum_{\nu \in \mathcal{M}} w_\nu V_\nu^\pi(\epsilon) = V_\xi^\pi(\epsilon) \qquad \text{(Intelligence Measure)}$$

$$\text{AIXI} := \operatorname*{argmax}_\pi \Upsilon(\pi) = \pi^*_\xi$$

$$V_\xi^\pi(h) = \sum_{\nu \in \mathcal{M}} w_h^\nu V_\nu^\pi(h)$$

$$w_\nu := 2^{-K(\nu)} \implies \xi(e_{1:m}|a_{1:m}) \stackrel{\times}{=} M(e_{1:m}|a_{1:m}) := \sum_{p:U(p,a_{1:m})=e_{1:m}} 2^{-\ell(p)}$$

AIXI

$$a_k^* := \operatorname*{argmax}_{a_k} \sum_{e_k} \ldots \max_{a_m} \sum_{e_m} \left[\sum_{i=k}^{m} \gamma^{i-k} r_i \right] \sum_{p:U(p,a_{1:m})=e_{1:m}} 2^{-\ell(p)} \qquad \text{(AIXI)}$$

Figure: ExpectiMax

AIXI

0 0 0 1 5 0 4
Observation Input
Reward Input
Input
Work
AIXI
Output
1 0 0 1
Output
Environment
Input
Work

RL vs GRL

- If μ is a completely observable MDP, V_μ^π reduces to the recursive Bellman equation.
- In a finite MDP, with a geometric discounting function, we can plan ahead by value iteration.
- According to Banach's fixpoint theorem, value iteration converges to the value of the optimal policy.
- What about GRL?

discount $\gamma : \mathbb{N}^2 \to [0,1]$ and utility $u : (\mathcal{A} \times \mathcal{E})^* \to [0,1]$

$$V_t^{\pi\mu}(h_{<k}) := \mathbb{E}_\mu^\pi \left[\sum_{i=k}^{\infty} \gamma_t^i u(h_{1:i}) \middle| h_{<k} \right]$$

Assumption

$$\forall t \in \mathbb{N}^+ : \lim_{m\to\infty} \sup_\pi \sum_{h_{<m}} V_t^{\pi\mu}(h_{<m})_\mu^\pi(h_{<m}) = 0$$

Theorem (Extreme Value Theorem)

If K is compact and $f : K \to \mathbb{R}$ is continuous, then f is bounded and there exist $p, q \in K$ s.t. $f(p) = \sup_{x\in K} f(x)$ and $f(q) = \inf_{x\in K} f(x)$.

$$\left\langle \Pi := \mathcal{A}^{(\mathcal{A}\times\mathcal{E})^*}, D(\pi,\pi') := \mathrm{e}^{-\min\{n : \exists h_{<n}(\pi(h_{<n}) \neq \pi'(h_{<n}))\}} \right\rangle$$

$V_\mu^\pi(h) : \Pi \to \mathbb{R}$ is continuous on the compact metric space $\langle \Pi, D \rangle$.

$$\pi_t^\mu := \operatorname*{argmax}_\pi V_t^{\pi\mu} \qquad \pi^\mu(h_{<t}) := \pi_t^\mu(h_{<t})$$

Deterministic vs Stochastic

If

$$\mu(e_{<t}|a_{<t}) = \sum_{p:U(p,a_{<t})=e_{<t}} \mu(p)$$

then μ can be interpreted in *two ways*:

- either the true environment is **deterministic**, but we only have **subjective belief** of which environment being the true environment; or
- the environment itself behaves **stochastically** defined by μ.

Intelligence vs Game

	Game in $\mathcal{M}_D$	Game in $\mathcal{M}_U$
Ex-post Equilibrium	Deterministic	π^*_μ(recursive/iterative)
Bayesian-Nash Equilibrium	π^*_μ(functional)	π^*_ξ(functional)

- Ex-post expected utility $V_t^{\pi\mu}$ $\quad V_\mu^\pi$
- Ex-interim expected utility (Intelligence Measure) $V_t^{\pi\xi}$ $\quad V_\xi^\pi$
- Ex-post equilibrium π_t^μ $\quad \pi^*_\mu$
- Bayesian-Nash equilibrium π_t^ξ $\quad \pi^*_\xi$
- Perfect Bayesian-Nash equilibrium π^ξ

Intelligence is an Equilibrium,
We just have to Identify the Game.

$$\text{Intelligence} = \underbrace{\text{Induction} + \text{Action}}_{\textit{efficiently}}$$

On-Policy Value Convergence for Bayes

Theorem (On-Policy Value Convergence for Bayes)

For any environment $\mu \in \mathcal{M}$ and any policy π,

$$\mathop{\pi}_{\mu}\left(\lim_{t\to\infty}\left[V^{\pi}_{\xi}(æ_{<t}) - V^{\pi}_{\mu}(æ_{<t})\right] = 0\right) = 1$$

- Bayesian agents perform well at learning and achieve on-policy value convergence: the posterior belief about the value of a policy π converges to the true value of π while following π: $V^{\pi}_{\xi}(æ_{<t}) - V^{\pi}_{\mu}(æ_{<t}) \xrightarrow{t\to\infty} 0$ $\mathop{\pi}_{\mu}$-almost surely.
- Since this holds for any policy, in particular it holds for the Bayes optimal policy π^*_{ξ}. This means that the Bayes agent learns to predict those parts of the environment that it sees. But if it does not explore enough, then it will not learn other parts of the environment that are potentially more rewarding.

AIXI

- Intelligence measure: valid, informative, wide range, general, dynamic, unbiased, fundamental, formal, objective, fully defined, universal?
- AIXI is the most intelligent environmental independent, i.e. universally optimal, agent possible?
- Applications: Sequence Prediction, Games, Optimization, Supervised Learning, Classification...
- AIXI is not limit computable, thus can't be approximated using finite computation. However there are limit computable ε-optimal approximations to AIXI.
- There are no known nontrivial and non-subjective optimality results for AIXI. General reinforcement learning is difficult even when disregarding computational costs.

AIXI Depends on UTM/Prior! — Dogmatic Prior

Dogmatic prior: if not acting according to one particular dogma π, got to hell with high probability. As long as the policy π yields some rewards, the prior says that exploration would be too costly and AIXI does not dare to explore.

Dogmatic Prior

Theorem (Dogmatic Prior)

Let π be any computable deterministic policy, let ξ be any Bayesian mixture over $\mathcal{M}_{\mathrm{LSC}}$. For $\varepsilon > 0$, there is a Bayesian mixture ξ' s.t. for any history $h_{<t}$ consistent with π and for which $V_{\xi}^{\pi}(h_{<t}) > \varepsilon$, the action $\pi(h_{<t})$ is the unique ξ'-optimal action.

Proof Sketch.

For every ν, let $\tilde{\nu}$ mimic ν until it receives an action that the policy π would not take. From then on, it provides rewards 0.

$$\tilde{\nu}\left(e_{1:t} \| a_{1:t}\right) := \begin{cases} \nu\left(e_{1:t} \| a_{1:t}\right), & \text{if } \forall k \leq t : a_k = \pi\left(æ_{<k}\right) \\ \nu\left(e_{<k} \| a_{<k}\right), & \text{if } k := \min\left\{i : a_i \neq \pi\left(æ_{<i}\right)\right\} \text{ exists} \\ & \quad \text{and } \forall i \in \{k, \ldots, t\} : e_i = (o, 0) \\ 0, & \text{otherwise} \end{cases}$$

Let $\tilde{w}(\nu) := \varepsilon w(\nu)$ and $\tilde{w}(\tilde{\nu}) := (1-\varepsilon) w(\nu) + \varepsilon w(\tilde{\nu})$.

The dogmatic prior $\tilde{w}$ puts much higher weight on the $\tilde{\nu}$ that behaves just like ν on the policy π, but sends any policy deviating from π to hell.

Theorem (AIXI Emulates Computable Policies)

Let $\varepsilon > 0$ and let π be any computable policy. There is a Bayesian mixture ξ' s.t. for any ξ'-optimal policy $\pi^*_{\xi'}$, and for any environment ν,

$$\left| V_\nu^{\pi^*_{\xi'}}(\epsilon) - V_\nu^{\pi}(\epsilon) \right| < \varepsilon$$

Theorem (Computable Policies are Dense)

The set $\left\{ \Upsilon_\xi(\pi) : \pi \text{ is a computable policy} \right\}$ is dense in $\left[\underline{\Upsilon}_\xi, \overline{\Upsilon}_\xi \right]$.

Deterministic policies are not dense in $\left[\underline{\Upsilon}_\xi, \overline{\Upsilon}_\xi \right]$.

AIXI Depends on UTM/Prior!

$$\overline{\Upsilon}_\xi := \sup_\pi \Upsilon_\xi(\pi) = \sup_\pi V_\xi^\pi(\epsilon) = V_\xi^{\pi^*_\xi}(\epsilon) = \Upsilon_\xi(\pi^*_\xi)$$

Computable policies are dense in $\left[\underline{\Upsilon}_\xi, \overline{\Upsilon}_\xi\right]$.

AIXI emulates computable policies.

AIXI can be arbitrarily stupid!

The devil imitates God. — orthogonality!

- Prior problem in Universal Induction ✓
- Prior problem in Universal Intelligence !

Stupid AIXI

Theorem (Some AIXIs are Stupid)

For any Bayesian mixture ξ over $\mathcal{M}_{\mathrm{LSC}}$ and every $\varepsilon > 0$, there is a Bayesian mixture ξ' s.t. $\Upsilon_\xi(\pi^*_{\xi'}) < \underline{\Upsilon}_\xi + \varepsilon$.

Theorem (AIXI is Stupid for Some Υ)

For any deterministic ξ-optimal policy π^*_ξ and for every $\varepsilon > 0$ there is a Bayesian mixture ξ' s.t. $\Upsilon_{\xi'}(\pi^*_\xi) \leq \varepsilon$ and $\overline{\Upsilon}_{\xi'} > 1 - \varepsilon$.

Theorem (Computable Policies can be Smart)

For any computable policy π and any $\varepsilon > 0$ there is a Bayesian mixture ξ' s.t. $\Upsilon_{\xi'}(\pi) > \overline{\Upsilon}_{\xi'} - \varepsilon$.

What is a good optimality criterion?

- Pareto optimality is *trivial*. Every policy is Pareto optimal in any $\mathcal{M} \supset \mathcal{M}_{\text{comp}}$.
- Bayes-optimality is *subjective*, because two different Bayesians with two different universal priors could view each other's AIXI as a very stupid agent.

Optimality

- Pareto optimality

$$\nexists \pi' : \forall \nu \in \mathcal{M} \left[\left(V_\nu^{\pi'}(\epsilon) \geq V_\nu^{\pi}(\epsilon) \right) \ \& \ \exists \rho \in \mathcal{M} \left(V_\rho^{\pi'}(\epsilon) > V_\rho^{\pi}(\epsilon) \right) \right]$$

- Balanced Pareto optimality

$$\forall \pi' : \sum_{\nu \in \mathcal{M}} w_\nu \left(V_\nu^{\pi}(\epsilon) - V_\nu^{\pi'}(\epsilon) \right) \geq 0$$

- Bayes optimality($\iff$ Balanced Pareto optimality)

$$\forall h_{<t} : V_\xi^{\pi}(h_{<t}) = V_\xi^{*}(h_{<t})$$

- Probably approximately correct (PAC)

$$\forall \varepsilon \delta > 0 : {}^{\pi}_{\mu} \left(\forall t \geq m(\varepsilon, \delta) : V_\mu^{*}(h_{<t}) - V_\mu^{\pi}(h_{<t}) > \varepsilon \right) < \delta$$

Optimality? — Guess how God created the multiverse

$$\text{prior}\begin{cases}\text{distribution}\\ \text{hypothesis space}\\ \text{prior probability}\\ \text{regularization}\end{cases}$$

No learning without prior!
no-free-lunch

$$\left.\begin{array}{r}\text{Homogeneous}\\ \text{Causality}\\ \text{Simplicity}\\ \text{Goodness}\\ \text{Beauty}\\ \text{Perfection}\\ \text{Value}\\ \text{Regret}\\ \text{Unexpectedness}\\ \text{Interesting}\\ \cdots\end{array}\right\} \Longleftarrow \text{God!}$$

Genesis — Zero-Sum Two Person Game

Figure: center of mass $\operatorname{argmax}_w \mathbb{E}_w [D(\nu \| \xi)]$

$$\xrightarrow{\text{message}} \boxed{\text{encoder}} \xrightarrow{\mu} \boxed{\begin{array}{c} \text{channel} \\ P(x|\nu) \\ \| \\ \nu(x) \end{array} \begin{bmatrix} \nu_1 \\ \nu_2 \\ \vdots \\ \nu_n \end{bmatrix}} \xrightarrow{x} \boxed{\text{decoder}} \xrightarrow{\text{estimate of message}}$$

Figure: possible worlds as channel — dominant strategy equilibrium

Genesis — Zero-Sum Two Person Game

"Subtle is the Lord, but malicious He is not."?

- God's strategy: w
- Agent's strategy: ξ
- God's utility: expected redundancy $\mathbb{E}_w[D(\mu\|\xi)]$
- Agent's utility: $-$ expected redundancy / error bound / channel capacity $\max_w E_w[D(\mu\|\xi)] = \max_w I(\mathcal{M};\mathcal{X})$
- Nash equilibrium: (w^*, ξ^*) dominant strategy equilibrium

$$w^* = \operatorname*{argmax}_w I(\mathcal{M};\mathcal{X})$$

$$\xi^* = \operatorname*{argmin}_\xi \mathbb{E}_{w^*}[D(\mu\|\xi)]$$

$\boxed{\text{The error bound could be arbitrarily large!}}$

Genesis

- Occam's razor vs Maximum entropy.

$$\underset{w \vDash \begin{cases} H(w) = C \\ \sum_{\nu \in \mathcal{M}} w_\nu = 1 \end{cases}}{minimize} \sum_{\nu \in \mathcal{M}} w_\nu K(\nu) \qquad \underset{w \vDash \begin{cases} \sum_{\nu \in \mathcal{M}} w_\nu K(\nu) = C \\ \sum_{\nu \in \mathcal{M}} w_\nu = 1 \end{cases}}{maximize} \; H(w)$$

- Optimal code length for possible worlds — Solomonoff prior.

$$\underset{w \vDash \sum_{\nu \in \mathcal{M}} w_\nu = 1}{minimize} \frac{\mathbb{E}_w[K]}{H(w)}$$

- Maximum expected redundancy/error bound/channel capacity.

$$\underset{w \vDash \begin{cases} H(w) = C \\ \sum_{\nu \in \mathcal{M}} w_\nu = 1 \end{cases}}{maximize} \mathbb{E}_w\left[D(\nu \| \xi)\right] \qquad \underset{w \vDash \begin{cases} \sum_{\nu \in \mathcal{M}} w_\nu K(\nu) = C \\ \sum_{\nu \in \mathcal{M}} w_\nu = 1 \end{cases}}{maximize} \; \mathbb{E}_w\left[D(\nu \| \xi)\right]$$

What is a good optimality criterion?

Asymptotic optimality

- Asymptotic optimality requires only convergence *in the limit*.
- The agent can be arbitrarily lazy.
- AIXI is not asymptotically optimal because it does not explore enough.
- To be asymptotically optimal you have to explore everything.
- If you explore more, you're likely to end up in a trap.
- Every policy will be asymptotically optimal after falling into the trap.

Agent needs to explore infinitely often for an entire effective horizon.

$$\mathcal{M} := \{\nu_\infty, \nu_1, \nu_2, \dots\}$$

Asymptotic Optimality

- strongly asymptotically optimal

$$\underset{\mu}{\pi}\left(\lim_{t\to\infty}\left[V^*_\mu(h_{<t}) - V^\pi_\mu(h_{<t})\right] = 0\right) = 1$$

- weakly asymptotically optimal

$$\underset{\mu}{\pi}\left(\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\left[V^*_\mu(h_{<t}) - V^\pi_\mu(h_{<t})\right] = 0\right) = 1$$

- asymptotically optimal in mean

$$\lim_{t\to\infty}\mathbb{E}^\pi_\mu\left[V^*_\mu(h_{<t}) - V^\pi_\mu(h_{<t})\right] = 0$$

- asymptotically optimal in probability (PAC)

$$\forall\varepsilon > 0 : \lim_{t\to\infty}\underset{\mu}{\pi}\left(V^*_\mu(h_{<t}) - V^\pi_\mu(h_{<t}) > \varepsilon\right) = 0$$

$$\text{strong a.o.} \implies \begin{cases}\text{weak a.o.}\\ \text{a.o. in mean} \iff \text{a.o. in probability}\end{cases}$$

AIXI

- AIXI is not asymptotically optimal.

$$\boxed{\forall \mathcal{M} \supset \mathcal{M}_{\text{comp}} \exists \mu \in \mathcal{M} \exists t_0 \forall t \geq t_0 : {}_{\mu}^{\pi^*_\xi} \left(\lim_{t\to\infty} V^*_\mu(h_{<t}) - V^{\pi^*_\xi}_\mu(h_{<t}) = \tfrac{1}{2} \right) = 1}$$

- AIXI achieves on-policy value convergence.

$$\boxed{{}_{\mu}^{\pi} \left(\lim_{t\to\infty} V^\pi_\mu(h_{<t}) - V^\pi_\xi(h_{<t}) = 0 \right) = 1}$$

Similarly for MDL $\operatorname*{argmin}_{\nu\in\mathcal{M}} \{-\log \nu(e_{<t}|a_{<t}) + K(\nu)\}$

and universal compression $2^{-Km(e_{<t}|a_{<t})}$.

Remark: AIXI asymptotically learns to predict the environment perfectly and with a small total number of errors analogously to Solomonoff induction, but only on policy: AIXI learns to correctly predict the value of its own actions, but generally not the value of counterfactual actions that it does not take.

Effective Horizon

$$\Gamma_t := \sum_{i=t}^{\infty} \gamma_i \qquad H_t(\varepsilon) := \min\left\{m : \frac{\Gamma_{t+m}}{\Gamma_t} \leq \varepsilon\right\}$$

Theorem

If there is a nonincreasing computable sequence of positive reals $(\varepsilon_t)_{t\in\mathbb{N}}$ s.t. $\varepsilon_t \xrightarrow{t\to\infty} 0$ and $\frac{H_t(\varepsilon_t)}{t\,\varepsilon_t} \xrightarrow{t\to\infty} 0$, then there is a limit-computable policy that is weakly asymptotically optimal in the class of all computable stochastic environments.

Definition (ε-Optimal Policy)

A policy π is ε-optimal in environment ν iff

$$\forall h : V_\nu^*(h) - V_\nu^\pi(h) < \varepsilon$$

ε-optimal BayesExp

Theorem (Self-Optimizing Theorem)

Let μ be some environment. If there is a policy π and a sequence of policies $\pi_1, \pi_2, \ldots$ s.t for all $\nu \in \mathcal{M}$

$$\mathop{\pi}_{\mu}\left(\lim_{t\to\infty} V^*_\nu(h_{<t}) - V^{\pi_t}_\nu(h_{<t}) = 0\right) = 1 \tag{1}$$

then

$$\mathop{\pi}_{\mu}\left(\lim_{t\to\infty} V^*_\mu(h_{<t}) - V^{\pi^*_\xi}_\mu(h_{<t}) = 0\right) = 1$$

- The policies $\pi_1, \pi_2, \ldots$ need to converge to the optimal value on the history generated by $\mathop{\pi}_{\mu}$, not $\mathop{\pi_t}_{\nu}$.
- If $\pi = \pi^*_\xi$ and (1) holds for all $\mu \in \mathcal{M}$, then π^*_ξ is strongly asymptotically optimal in the class $\mathcal{M}$.
- π^*_ξ is strongly asymptotically optimal in the class of ergodic finite-state MDPs if $\forall \varepsilon : H_t(\varepsilon) \xrightarrow{t\to\infty} \infty$.

For Which Class $\mathcal{M}$ does $V_\mu^{\pi^*_\xi}$ Converge to V_μ^*?

Recoverability

An environment ν is recoverable iff

$$\lim_{t\to\infty} \sup_{\pi} \left| \mathbb{E}_{\nu}^{\pi_{\nu}^{*}} \left[V_{\nu}^{*}(h_{<t}) \right] - \mathbb{E}_{\nu}^{\pi} \left[V_{\nu}^{*}(h_{<t}) \right] \right| = 0$$

Remark: Recoverability compares following the worst policy π for $t-1$ time steps and then switching to the optimal policy π_{ν}^{*} to having followed π_{ν}^{*} from the beginning. The recoverability assumption states that switching to the optimal policy at any time enables the recovery of most of the value.

Sublinear Regret

$$R_m(\pi, \mu) := \sup_{\pi'} \mathbb{E}_\mu^{\pi'} \left[\sum_{t=1}^{m} r_t \right] - \mathbb{E}_\mu^{\pi} \left[\sum_{t=1}^{m} r_t \right]$$

Assumption (Discount Assumption)

1. $\forall t : \gamma_t > 0$
2. γ_t *is monotone decreasing.*
3. $\forall \varepsilon > 0 : H_t(\varepsilon) \in o(t)$

Theorem

If the discount function γ satisfies the discount assumption, the environment μ is recoverable, and π is asymptotically optimal in mean, then $R_m(\pi, \mu) \in o(m)$.

$$\operatorname*{argmin}_{\pi} \max_{\mu} R_m(\pi, \mu) \qquad\qquad w_m^\mu := \frac{2^{-R_m(\pi,\mu)}}{\sum_{\mu \in \mathcal{M}} 2^{-R_m(\pi,\mu)}}$$

Regret in Non-Recoverable Environments

$$R_m(\alpha, \mu_1) = m \qquad R_m(\alpha, \mu_2) = 0$$

$$R_m(\beta, \mu_1) = 0 \qquad R_m(\beta, \mu_2) = m$$

For non-recoverable environments:

Either the agent gets caught in a trap or it is not asymptotically optimal.

Intrinsic Utility

- square $-\xi(e_{t:k}|æ_{<t}a_{t:k})$
- shannon $-\log\xi(e_{t:k}|æ_{<t}a_{t:k}) \approx K(æ_{1:k}) - K(æ_{<t})$
- KL divergence $D\left(w_{æ_{<k}}\|w_{æ_{<t}}\right)$ where $w^{\nu}_{æ_{<n}} = \dfrac{w_\nu \nu(e_{<n}|a_{<n})}{\xi(e_{<n}|a_{<n})}$
- information gain $H(w_{h_{<t}}) - H(w_{h_{1:k}})$ where

$$H(w_h) := -\sum_{\nu\in\mathcal{M}} w^{\nu}_h \log w^{\nu}_h$$

- effective complexity $\mathcal{E}_\delta(æ_{1:k}) - \mathcal{E}_\delta(æ_{<t})$ where

$$\mathcal{E}_\delta(æ_{<n}) := \min_{\nu\in\mathcal{M}}\left\{2K(\nu) + H(\nu) - K(æ_{<n}) : \nu(e_{<n}|a_{<n}) \geq 2^{-H(\nu)(1+\delta)}\right\}$$

- logical depth $\mathrm{depth_b}(h_{1:k}) - \mathrm{depth_b}(h_{<t})$ where

$$\mathrm{depth_b}(x) := \min\left\{t : U^t(p) = x \ \ \& \ \ \ell(p) - K(x) \leq b\right\}$$

Hibbard's Two-Stage Model-Based Utility Agent

$$\lambda(h) := \operatorname*{argmax}_{q \in Q} P(h|q)P(q)$$

$$\rho(h') = P(h'|\lambda(h))$$

$$Q(ha) = \sum_{e \in \mathcal{E}} \rho(e|ha) \left[\sum_{z \in Z_h} P(z|\lambda(h))u(z) + \gamma V(h\text{æ}) \right]$$

$$V(h) = \max_{a \in \mathcal{A}} Q(ha)$$

$$\pi(h) = \operatorname*{argmax}_{a \in \mathcal{A}} Q(ha)$$

where Z_h is the internal state histories induced by $\lambda(h_{<t})$ that are consistent with h.

Remarks: An agents using model-based utility function will not self-delude: it need to make more accurate estimate of its environment state variables from its interaction history.

Russell's Principles for Beneficial Machines

1. The machine's only objective is to maximize the realization of human preferences.
2. The machine is initially uncertain about what those preferences are.
3. The ultimate source of information about human preferences is human behavior.

Daniel Dewey's Value Learning Agent & CIRL

$$a_k^* = \operatorname*{argmax}_{a_k} \sum_{e_k \text{æ}_{k+1:m}} \xi(\text{æ}_{\leq m} | \text{æ}_{<k} a_k) \sum_{u \in \mathcal{U}} P(u | \text{æ}_{\leq m}) u(\text{æ}_{\leq m})$$

What could it mean for a machine to have its own goals?

Shutdown Button — Uncertainty of goals

$$\tilde{U}(u) \implies P_{\tilde{U}}(u)$$

Russell: Cooperative Inverse Reinforcement Learning
CIRL agents learn about a human utility function u^* by observing the actions the human takes.

$$V^*(\text{æ}_{<k}) = \max_{a_k \in \mathcal{A}} Q^*(\text{æ}_{<k} a_k)$$

$$Q^*(\text{æ}_{<k} a_k) = \mathbb{E}_{e_k} \left[\sum_{a_k^H} P(a_k^H \mid a_k) \sum_{u \in \mathcal{U}} P(u \mid a_k, a_k^H) u(\text{æ}_{1:k}) + \gamma V^*(\text{æ}_{1:k}) \,\middle|\, \text{æ}_{<k} a_k \right]$$

Leibniz Prior

- There's much we don't know about the world.
- but we know it's the best possible world.
- So simplicity and richness will be represented in the actual (best possible) world.
- This is a good inductive bias.

Leibniz Prior

- the best of all possible worlds
- balancing the simplicity of means against the richness of ends
- pre-established harmony

prior
$\Downarrow$
utility
$\Downarrow$
prior

Orthogonality!
Wisdom $\neq$ Intelligence

universal prior (assumption) w
$\Downarrow$ simplicity/richness
intrinsic utility
$\Downarrow$ stochastic environment
expected intrinsic utility
$\Downarrow$ Schauder fixpoint
universal prior w^*
$\Downarrow$ Bayesian mixture
ξ
$\Downarrow$ ExpectiMax
π^*_ξ

Leibniz's "Wisdom"

$$\underline{W}isdom = \operatorname*{argmax}_{\pi} \mathbb{E}_{\xi}^{\pi}[\underline{H}appiness]$$

$$\underline{H}appiness = \sum_{t=1}^{\infty} \underline{P}erfection(t)$$

$$\underline{P}erfection = \underline{V}ariety - \underline{S}implicity$$

$$\underline{V}ariety = \mathbb{E}_{w}[\underline{P}erception]$$

$$\underline{P}erception = \underline{R}eason + (\underline{E}xperience|\underline{R}eason)$$

$$\pi^{*} := \operatorname*{argmax}_{\pi} \mathbb{E}_{\xi}^{\pi}\left[\sum_{t=1}^{\infty}\left(\mathbb{E}_{w}\left[R + (E|R)\right] - S\right)\right]$$

Leibniz's "Wisdom"

$$u^{\text{in}}(h_{1:t}) = H(w_\epsilon) - H(w_{h_{1:t}}) \quad \text{or} \quad D(w_{h_{1:t}} \| w_\epsilon) - D(w_\epsilon \| w_{h_{1:t}})$$

$$\bar{U}(\nu) = \mathbb{E}_\nu \left[\sum_{t \geq 1} u^{\text{in}}(h_{1:t}) \right]$$

$$w_\epsilon^\nu \mapsto \bar{U}(\nu) \mapsto w_\epsilon^\nu$$

$$\pi^* := \operatorname*{argmax}_\pi \mathbb{E}_\xi^\pi \left[\sum_{t=1}^{\infty} u^{\text{in}}(h_{1:t}) \right]$$

- Prior: Simplicity(Kolmogorov Complexity) $\xrightarrow[\text{regular/random } \mathcal{M}]{\text{break block uniform}}$ free lunch
- Intrinsic Utility
- Universal Prior (Natural UTM)

Metaphysical vs Moral/Utilitarian

means vs ends wisdom vs intelligence

simplicity/richness $\longrightarrow$ intrinsic utility $\longrightarrow$ universal prior

inverse/value reinforcement learning

- **orthogonality**
- **human interests**
- **external wireheading**
- **shutdown button**

Knowledge-Seeking Agent

$$V_{\mathrm{IG}}^{\pi,m}(h_{<t}) := \mathbb{E}_{\xi}^{\pi}\left[H(w^{\cdot}_{h_{<t}}) - H(w^{\cdot}_{h_{1:m}}) \,\middle|\, h_{<t}\right] = \sum_{\nu\in\mathcal{M}} w^{\nu}_{h_{<t}} D_m\left({}^{\pi}_{\nu}\middle\|{}^{\pi}_{\xi} \mid h_{<t}\right)$$

$$D_\gamma\left({}^{\pi}_{\nu}\middle\|{}^{\pi}_{\xi} \mid h_{<t}\right) := \sum_{k=t}^{\infty}\gamma_k \sum_{h'\in\mathcal{H}^{k-t}} {}^{\pi}_{\nu}(h'|h_{<t}) D\left({}^{\pi}_{\nu}\middle\|{}^{\pi}_{\xi} \mid h_{<t}h'\right)$$

$$V_{\mathrm{IG}}^{\pi}(h_{<t}) := \mathbb{E}_{\xi}^{\pi}\left[\sum_{k=t}^{\infty}\gamma_k D\left(w^{\cdot}_{h_{1:k}}\middle\|w^{\cdot}_{h_{<k}}\right) \,\middle|\, h_{<t}\right] = \sum_{\nu\in\mathcal{M}} w^{\nu}_{h_{<t}} D_\gamma\left({}^{\pi}_{\nu}\middle\|{}^{\pi}_{\xi} \mid h_{<t}\right)$$

$$\pi^*_{\mathrm{IG}} := \operatorname*{argmax}_{\pi} V_{\mathrm{IG}}^{\pi}$$

$$\lim_{t\to\infty}\frac{1}{\Gamma_t}\mathbb{E}_{\mu}^{\pi}\left[D_\gamma\left({}^{\pi}_{\mu}\middle\|{}^{\pi}_{\xi} \mid h_{1:t}\right)\right] = 0 \qquad \text{(on-policy)}$$

$$\lim_{t\to\infty}\frac{1}{\Gamma_t}\mathbb{E}_{\mu}^{\pi^*_{\mathrm{IG}}}\left[\sup_{\pi\in\Pi(h_{1:t})} D_\gamma\left({}^{\pi}_{\mu}\middle\|{}^{\pi}_{\xi} \mid h_{1:t}\right)\right] = 0 \qquad \text{(off-policy)}$$

maximize knowledge / exploration=exploitation / resistant to noise / avoid traps

Bayesian Agent

Algorithm 1 Bayesian Agent

Require: Model class $\mathcal{M}$; prior $w \in \Delta\mathcal{M}$; history $æ_{<t}$.

```
function ACT(π)
    Sample and perform action a_t ~ π(·|æ_{<t})
    Receive e_t ~ ν( · |æ_{<t} a_t)
    for ν ∈ M do
        w_ν ← ν(e_{<t}|a_{<t}) / ξ(e_{<t}|a_{<t}) w_ν
    end for
    t ← t + 1
end function
```

MDL

Algorithm 2 MDL Agent

Require: Model class $\mathcal{M}$; prior $w \in \Delta\mathcal{M}$; regularizer constant $\lambda \in \mathbb{R}^+$.

1: **loop**

2: $\quad \sigma \leftarrow \arg\min_{\nu \in \mathcal{M}} \left[K(\nu) - \lambda \sum_{k=1}^{t} \log \nu(e_k | æ_{<k} a_k) \right]$

3: $\quad \text{ACT}\left(\pi^*_\sigma\right)$

4: **end loop**

MDL

Definition (MDL)

$$\widehat{\nu} = \arg\min_{\nu \in \mathcal{M}} \{K_\nu(x) + K_w(\nu)\} = \arg\max_{\nu \in \mathcal{M}} \{w_\nu \nu(x)\}$$

where $K_\nu(x) := -\log \nu(x)$ and $K_w(\nu) := -\log w_\nu$

Theorem (MDL Bound)

$$\sum_{t=1}^{\infty} \mathbb{E}_\mu \left[\sum_{x_t \in \mathcal{X}} \left(\widehat{\nu}(x_t | x_{<t}) - \mu(x_t | x_{<t}) \right)^2 \right] \stackrel{+}{\leq} 8 w_\mu^{-1}$$

MDL converges, but speed can be exponential worse than Bayes.

Weak Asymptotic Optimality — Optimistic Agent

Algorithm 3 Optimistic Agent π° $\qquad \pi^\circ_t := \operatorname{argmax}_\pi \max_{\nu \in \mathcal{M}_t} V^\pi_\nu(h_{1:t})$

Require: Finite class of deterministic environments $\mathcal{M}_0 = \mathcal{M}$

```
t = 1
repeat
    (π*, ν*) := argmax_{π∈Π, ν∈M_{t-1}} V^π_ν(h_{t-1})
    repeat
        a_t = π*(h_{t-1})
        Perceive e_t from environment μ
        h_t ← h_{t-1} a_t e_t
        Remove inconsistent environment M_t := {ν ∈ M_{t-1} : h_t^{π°ν} = h_t}
        t ← t + 1
    until ν* ∉ M_{t-1}
until M = ∅
```

stochastic case: $\mathcal{M}_t := \left\{\nu \in \mathcal{M}_{t-1} : \nu(e_{<t}|a_{<t}) \geq \varepsilon_t \max_{\rho \in \mathcal{M}} \rho(e_{<t}|a_{<t})\right\}$

Act optimally w.r.t. the most optimistic environment until contradicted.

If there is a chance: Try it! — Vulnerable to traps.

Asymptotic Optimality in Mean — Thompson Sampling

Algorithm 4 Thompson Sampling π_T

Require: Model class $\mathcal{M}$; prior $w \in \Delta\mathcal{M}$; exploration schedule $(\varepsilon_t)_{t\in\mathbb{N}}$.

```
loop
    Sample ρ ~ w(·|æ_{<t})
    for i = 1 → H_t(ε_t) do
        ACT(π*_ρ)
    end for
end loop
```

Theorem

If the discount function γ satisfies the discount assumption, the environment μ is recoverable, then $R_m(\pi_T, \mu) \in o(m)$.

Weak Asymptotic Optimality — BayesExp

Algorithm 5 BayesExp π_{BE}

Require: Model class $\mathcal{M}$; prior $w \in \Delta\mathcal{M}$; exploration schedule $(\varepsilon_t)_{t\in\mathbb{N}}$.

```
loop
    if V*_IG(æ_<t) > ε_t then
        for i = 1 → H_t(ε_t) do
            ACT(π*_IG)
        end for
    else
        ACT(π*_ξ)
    end if
end loop
```

ε-**optimal BayesExp:** If the optimal information gain value $V^*_{\mathrm{IG}} > \varepsilon_t$, then execute the ε-optimal information gain policy $\pi^{\varepsilon_t}_{\mathrm{IG}}$ for $H_t(\varepsilon_t)$ steps, else execute $\pi^{\varepsilon_t}_{\xi}$ for 1 step.

- *BayesExp* performs phases of exploration in which it maximizes the expected information gain. This explores the environment class completely, even achieving off-policy prediction.
- In contrast, Thompson sampling only explores on the optimal policies, and in some environment classes this will not yield off-policy prediction. So in this sense the exploration mechanism of Thompson sampling is more reward-oriented than maximizing information gain.

Strong Asymptotic Optimality — Inquisitive Agent

$$V_{\mathrm{IG}}^{\pi}(h_{<t}) := \mathbb{E}_{\xi}^{\pi}\left[D\left(w^{\cdot}_{h_{<t+m}} \middle\| w^{\cdot}_{h_{<t}}\right) \middle| h_{<t}\right]$$

$$\pi_{\mathrm{IG}}^{m,k} := \operatorname*{argmax}_{\pi \in \mathcal{A}^{\mathcal{H}^{<m}}} V_{\mathrm{IG}}^{\pi}(h_{<t-k})$$

$$\rho(h_{<t}, m, k) := \min\left\{\frac{1}{m^2(m+1)}, \eta V_{\mathrm{IG}}^{\pi_{\mathrm{IG}}^{m,k}}(h_{<t-k})\right\}$$

Algorithm 6 Inquisitive Agent $\pi^{\dagger}$

1: **while** True **do**
2: calculate $\rho(h_{<t}, m, k)$ for all m and for all $k < \min\{m, t\}$
3: ACT $\pi_{\mathrm{IG}}^{m,k}(h_{<t})$ with probability $\rho(h_{<t}, m, k)$
4: ACT $\pi_{\xi}^{*}(h_{<t})$ with probability $1 - \sum_{m \in \mathbb{N}} \sum_{k<m,t} \rho(h_{<t}, m, k)$
5: **end while**

$$\pi^{\dagger}(a|h_{<t}) := \sum_{m \in \mathbb{N}} \sum_{k<m,t} \rho(h_{<t}, m, k) \left[\!\left[a = \pi_{\mathrm{IG}}^{m,k}(h_{<t})\right]\!\right] + \left(1 - \sum_{m \in \mathbb{N}} \sum_{k<m,t} \rho(h_{<t}, m, k)\right) \left[\!\left[a = \pi_{\xi}^{*}(h_{<t})\right]\!\right]$$

Approximation

The AIXI approximations are outperformed by DQN/DRQN...

- MC-AIXI-CTW.
  - Approximate Solomonoff induction — most recent actions and percepts (=context) more relevant — Context Tree Weighting
  - Sample paths in expectimax tree.
- Feature Reinforcement Learning (ΦMDP). — history $\mapsto$ state
  e.g. Classical physics: Position+velocity of objects=position at two time-slices. (2nd order Markov.)

$$\Phi : h \mapsto s$$

$$\Phi^{\text{best}} := \operatorname*{argmin}_{\Phi} \text{Cost}(\Phi|h)$$

$$\text{Cost}(\Phi|h) := \text{CL}(s_{1:n}^{\Phi}|a_{1:n}) + \text{CL}(r_{1:n}|s_{1:n}^{\Phi}, a_{1:n}) + \text{CL}(\Phi)$$

  How to find the map Φ? Monte-Carlo...
- Compress and Control. — (model-free)
  Combine induction and planning.

Expectimax Approximation: MC-AIXI-CTW

Upper Confidence Tree (UCT) algorithm:

- Sample observations from Context Tree Weighting (CTW) distribution.

$$\mathrm{CTW}(e_{<t}|a_{<t}) := \sum_{\Gamma} 2^{-\mathrm{CL}(\Gamma)}\Gamma(e_{<t}|a_{<t})$$

- Select actions with highest upper confidence bound.

$$a_{\mathrm{ucb}} := \operatorname*{argmax}_{a\in\mathcal{A}} \left(\underbrace{\hat{Q}(æ_{<t}a)}_{\text{average}} + \underbrace{\sqrt{\frac{\log T(æ_{<t})}{T(æ_{<t}a)}}}_{\text{exploration bonus}} \right)$$

  where $T(\cdot)$ is the number of times a sequence has been visited.
- Expand tree by one leaf node (per trajectory).
- Simulate from leaf node further down using (fixed) playout policy.
- Propagate back the value estimates for each node.

MC-AIXI-CTW

Environment
Perform action in real world
Record new sensor information
... Past
Observation/Reward
Action
Record Action
MC-AIXI
An approximate AIXI agent
Refine environment model
a1
a2
a3
o1
o2
o3
o4
Update Bayesian Mixture of Models
+
-
-
-
+
-
.....
Simple
Complex
Large Prior
Small Prior
future reward estimate
Determine best action

Feature Reinforcement Learning (ΦMDP)

Universal Search

- Levin Search
- Speed Prior
- Hutter Search
- AIXI$^{t\ell}$
- Optimal Ordered Problem Solver
- Gödel Machine

Figure: Levin

Levin Search (LSEARCH)

An inversion algorithm p inverts a function f if given x, $p(x) = y$ s.t. $f(y) = x$.

LSEARCH

Run all $\{p : \ell(p) \leq i\}$ for $2^{i-\ell(p)}$ steps in phase $i = 1, 2, 3, \ldots$ until it has inverted f on x.

$$Kt(x) := \min_p \left\{\ell(p) + \log t(p, x) : U(p) = x\right\}$$

Theorem

All strings $\{x : Kt(x) \leq n\}$ can be generated and tested in 2^{n+1} steps.

$$t_{\text{LSEARCH}}(x) = O\left(2^{K(n)} t^+_{p_n}(x)\right)$$

where $t^+_{p_n}(x)$ is the runtime of $p_n(x)$ plus the time to verify the correctness of the result $f(p_n(x)) = x$.

Remark: If P=NP, then LSEARCH is a P algorithm for every NP problem.

Speed Prior

$$S(e_{<t}|a_{<t}) := \sum_{p:U(p,a_{<t})=e_{<t}} \frac{2^{-\ell(p)}}{t(p, a_{<t}, e_{<t})}$$

S is computable.

A function f is estimable in polynomial time iff there is a function g computable in polynomial time s.t. $f \overset{\times}{=} g$.

For any measure μ estimable in polynomial time,

$$\left(\sqrt{L_n^{\Lambda_S}} - \sqrt{L_n^{\Lambda_\mu}}\right)^2 \leq 2D_n(\mu\|S) = O(\log n)$$

On-Policy Value Convergence

If the effective horizon is bounded, then for any environment $\mu \in \mathcal{M}_{\text{comp}}$ estimable in polynomial time and any policy π,

$$\underset{\mu}{\pi}\left(\lim_{t\to\infty} \frac{1}{t} \sum_{k=1}^{t} \left(V_S^\pi(h_{<k}) - V_\mu^\pi(h_{<k})\right)\right) = 0$$

Hutter Search (HSEARCH)

$M^{\varepsilon}_{p^*}(x)$

Initialize the shared variables $L := \{\}$, $t_{\text{fast}} := \infty$, $p_{\text{fast}} := p^*$.

Start algorithms A, B, and C in parallel with ε, ε, and $1 - 2\varepsilon$ computation time, respectively.

B

for $(p, t) \in L$

run $t(x)$ in parallel for all t with computation time $2^{-\ell(p)-\ell(t)}$.

if for some t, $t(x) < t_{\text{fast}}$,

then $t_{\text{fast}} := t(x)$ and $p_{\text{fast}} := p$.

continue

A

for $i := 1, 2, 3, \ldots$ do

pick the last wff of the ith proof.

if it reads "$p(\cdot)$ is equivalent to $p^*(\cdot)$ and has time-bound $t(\cdot)$", then add (p, t) to L.

continue

C

for $k := 1, 2, 4, 8, \ldots$ do

run current p_{fast} for k steps.

if p_{fast} halts,

then print result $p_{\text{fast}}(x)$ and abort A, B and C.

continue

AIXI$^{t\ell}$

1. Let $P := \emptyset$. This will be the set of verified programs.
2. For all proofs of length $\leq n$: if the prover shows $\mathrm{VA}(p)$ for some p with $\ell(p) \leq \ell$, then add p to P.

   $$\mathrm{VA}(p) := \text{“}\forall k \forall (\mathrm{va}'æ)_{1:k} : p(æ_{<k}) = v_1 a'_1 \ldots v_k a'_k \implies v_k \leq V^{\pi}_{\xi}(æ_{<k})\text{”}$$

   (The program p not only computes future actions of π, which is the policy derived from p according to $\pi(æ_{<k}) := a'_k$, but also hypothetical past actions a'_i and lower bounds v_i for the value of the policy π.)
3. For each input history $æ_{<k}$ repeat: run all programs from P for $\leq t$ steps each, take the one with the highest promised value v_k, and return that program's policy's action.

- AIXI$^{t\ell}$ depends on t, ℓ, n but not on knowing p.
- Its setup-time is $t_{\mathrm{setup}}(p^{\mathrm{best}}) = O(n \cdot 2^n)$.
- Its computation time per cycle is $t_{cycle}(p^{\mathrm{best}}) = O(t \cdot 2^{\ell})$.

Schmidhuber's Optimal Ordered Problem Solver (OOPS)

- Solve the first task with LSEARCH.
- Freeze successful programs in non-writable memory.
- Programs tested during search for later tasks may copy non-writable code into separate modifiable storage, to edit it and execute the modified result.
- Given a new task, OOPS spends half of the time to test programs that have the most recent successful program as a prefix, the other half to fresh programs.
- Time is allocated according to a distribution over programs, which is obtained by multiplying the probabilities of the individual instructions.

Incremental Learning

Self-reference

- This sentence repeats the word ‘twice’ twice.
- Thare are five mistukes im this centence.
- **The only boldface sentence on this page is false.**
- All generalizations are wrong.
- Every rule has an exception except this one.
- Moderation in all things, including moderation.
- We must believe in free will — we have no choice!
- I know that I know nothing.
- There are two rules lor success in life:
  1. Never tell anyone all that you know.
- If you choose an answer to this question at random, what is the chance you will be correct? (A) 25% (B) 50% (C) 60% (D) 25%
-
  1. What is the best question to ask and what is the answer to it?
  2. The best question is the one you asked; the answer is the one I gave.
- Can you answer the following question in the same way to this one?
- One of the lessons of history is that no one ever learns the lessons of history.

Self-reference vs Paradox

The sentence below is false.

The sentence above is true.

Yablo Paradox

- S_1: for all $k > 1$, S_k is false.
- S_2: for all $k > 2$, S_k is false.
- S_3: for all $k > 3$, S_k is false.
- $\cdots$

Quine Paradox

"Yields falsehood when preceded by its quotation" yields falsehood when preceded by its quotation.

self-reference / circularity or infinite regress / negation / infinity / totality

The "Power" of Self-reference

Curry's Paradox

- If this sentence is true, then God exists.
- This sentence is false, and God does not exist.

Hi 美女，问你个问题呗

如果我问你"你能做我女朋友吗"，那么你的答案能否和**这个问题本身**的答案一样?

自我实现/自我修复?

这句话有 2 个'这'字，2 个'句'字，2 个'话'字，2 个'有'字，7 个'2'字，11 个'个'字，11 个'字'字，2 个'7'字，3 个'11'字，2 个'3'字。

How to Refer?

How to Refer? — Levels

Nested Virtualization?

从前有座山，山里有座庙，庙里有个老和尚在讲故事：从前有座山...

Liar Paradox vs Quine Paradox

1. 这句话是假的
2. “这句话是假的”是假的
3. ““““““……是假的”是假的”是假的”是假的”是假的”是假的
4. 把“把中的第一个字放到左引号前面，其余的字放到右引号后面，并保持引号及其中的字不变”中的第一个字放到左引号前面，其余的字放到右引号后面，并保持引号及其中的字不变
5. 把“把中的第一个字放到左引号前面，其余的字放到右引号后面，并保持引号及其中的字不变得到的句子是假的”中的第一个字放到左引号前面，其余的字放到右引号后面，并保持引号及其中的字不变得到的句子是假的

无穷虚拟层?

How to Refer? — Encoding

- 100 prisoners are lined up by an jailer, who places a red or blue hat upon each of their heads.
- The prisoners can see the hats of the people lined up in from of them, but they can't look at the hats behind them, or at their own.
- The jailer is going to ask color of each prisoner's hat starting from the last prisoner in queue. If a prisoner tells the correct color, then is saved, otherwise executed.
- How many prisoners can be saved at most if they are allowed to discuss a strategy before the jailer starts asking colors of their hats?

If the first person sees an odd number of red hats he calls out red, if he sees an even number of red hats he calls out blue.

手扶拐杖的外星绅士造访地球。临别，人类赠送百科全书："人类文明尽在其中！"。绅士谢绝："不，谢谢！我只需在拐杖上点上一点"。

Diagonalization[1]

Definition (Point-Surjective)

A morphism $f : X \to Y$ is *point-surjective* iff for every $y : 1 \to Y$, there is an $x : 1 \to X$ such that $y = f \circ x$.

Theorem (Lawvere's Fixpoint Theorem)

In a cartesian closed category, if there is a point-surjective morphism $f : X \to Y^X$, then every morphism $\alpha : Y \to Y$ has a fixpoint $y : 1 \to Y$.

[1] Lawvere: Diagonal arguments and cartesian closed categories.
Yanofsky: A universal approach to self-referential paradoxes, incompleteness and fixed points.

Lawvere's Fixpoint Theorem

- A function $g : X \to Y$ is *representable* by $f : X \times X \to Y$ iff

$$\exists y \forall x : g(x) = f(x, y)$$

Theorem (Lawvere's Fixpoint Theorem)

For sets X, Y, functions $f : X \times X \to Y$, $\alpha : Y \to Y$, let $g := \alpha \circ f \circ \Delta$.

1. *If α has no fixpoint, then g is not representable by f.*
2. *If g is representable by f, then α has a fixpoint.*

$$\begin{tikzcd}
X \times X \arrow[r, "f"] & Y \arrow[d, "\alpha"] \\
X \arrow[u, "\Delta"] \arrow[r, "g"'] & Y
\end{tikzcd}$$

$$\alpha\big(f\,(\ulcorner g \urcorner, \ulcorner g \urcorner)\big) = g\,(\ulcorner g \urcorner) = f\,(\ulcorner g \urcorner, \ulcorner g \urcorner)$$

- $\Delta : x \mapsto (x, x)$ diagonal
- f evaluation
- α "negation"
- g ($\ulcorner g \urcorner$) fixpoint-(free) transcendence
- $f\,(\ulcorner g \urcorner, \ulcorner g \urcorner)$ self-reference "I have property α."

Diagonalization

- A function $g : X \to Z$ is *representable* by $f : X \times Y \to Z$ iff

$$\exists y \in Y \forall x \in X : g(x) = f(x, y)$$

Theorem (Lawvere's Fixpoint Theorem)

For all sets X, Y, Z, and all functions $\beta : X \to Y, f : X \times Y \to Z, \alpha : Z \to Z$, let $g := \alpha \circ f \circ \langle 1_X, \beta \rangle$. Assume β is surjective.

1. *If α has no fixpoint, then g is not representable by f.*
2. *If g is representable by f, then α has a fixpoint.*

$$\begin{tikzcd}
X \times Y \arrow[r, "f"] & Z \arrow[d, "\alpha"] \\
X \arrow[u, "\langle 1_X, \beta \rangle"] \arrow[r, "g"'] & Z
\end{tikzcd}$$

Example — Kleene's Fixpoint Theorem

Theorem (Kleene's Fixpoint Theorem)

Given a recursive function h, there is an index e s.t.

$$\varphi_e = \varphi_{h(e)}$$

$$\begin{tikzcd}
\mathbb{N} \times \mathbb{N} \arrow[r, "f"] & \{\varphi_n\}_{n \in \mathbb{N}} \arrow[d, "\mathcal{E}_h"] \\
\mathbb{N} \arrow[u, "\Delta"] \arrow[r, "g"'] & \{\varphi_n\}_{n \in \mathbb{N}}
\end{tikzcd}$$

where $f : (m, n) \mapsto \varphi_{\varphi_n(m)}$, and $\mathcal{E}_h : \varphi_n \mapsto \varphi_{h(n)}$.
The function $g : m \mapsto \varphi_{h(\varphi_m(m))}$ is a recursive sequence of partial recursive functions, and thus is representable by f. Explicitly,

$$g(m) = \varphi_{h(\varphi_m(m))} = \varphi_{s(m)} = \varphi_{\varphi_t(m)} = f(m, t)$$
$$e := \varphi_t(t)$$

Fixpoint vs Diagonalization

$$\begin{tikzcd}
X \times X \arrow[r, "f"] & Y \arrow[d, "\alpha"] \\
X \arrow[u, "\Delta"] \arrow[r, "g"'] & Y
\end{tikzcd}$$

Curry $\mathbf{Y}$	$\hat{=}$	λ-fixpoint	$\hat{=}$	Gödel	$\hat{=}$	Kleene	$\hat{=}$	Russell
yx	$\hat{=}$	$N\,(\ulcorner M \urcorner)$	$\hat{=}$	$N(\ulcorner M(x) \urcorner)$	$\hat{=}$	$\varphi_n(m)$	$\hat{=}$	$x \in y$
xx	$\hat{=}$	$M\,(\ulcorner M \urcorner)$	$\hat{=}$	$M(\ulcorner M(x) \urcorner)$	$\hat{=}$	$\varphi_n(n)$	$\hat{=}$	$x \in x$
$y(xx)$	$\hat{=}$	$F\ulcorner M \ulcorner M \urcorner\urcorner$	$\hat{=}$	$F(\ulcorner M(\ulcorner M(x) \urcorner) \urcorner)$	$\hat{=}$	$h(\varphi_n(n))$	$\hat{=}$	$x \notin x$
$\lambda x.y(xx)$	$\hat{=}$	W	$\hat{=}$	$W(x)$	$\hat{=}$	$\varphi_t(n)$	$\hat{=}$	$x \notin R$
$(\lambda x.y(xx))(\lambda x.y(xx))$	$\hat{=}$	$W\,(\ulcorner W \urcorner)$	$\hat{=}$	$W(\ulcorner W(x) \urcorner)$	$\hat{=}$	$\varphi_t(t)$	$\hat{=}$	$R \notin R$

$$\boxed{\text{self-reference} \overset{?}{\Longrightarrow} \text{self-improvement}}$$

Kleene's Fixpoint Theorem

Theorem (Kleene's Fixpoint Theorem)

Given a recursive function h, there is an index e s.t.

$$\varphi_e = \varphi_{h(e)}$$

对于任意的程序 h，总存在某个程序 e，执行程序 e 的结果等价于把程序 e 当作数据输入给程序 h 执行的结果。

Self-reproducing Program

There is a program that outputs its own length.

There is a program that outputs its own source code.

Corollary (Self-reproducing Program)

There is a recursive function φ_e s.t. $\forall x : \varphi_e(x) = e$.

Quine in Python

```
s='s=%r; print(s%%s)'; print(s%s)
```

Quine in Lambda Calculus

$$(\lambda x.xx)(\lambda x.xx)$$

Self-reproducing Program

Print two copies of the following, the second copy in quotes:

"Print two copies of the following, the second copy in quotes:"

DNA / mutation / evolution

Build a baby that acts on the following instructions, and also contains a copy of those instructions in its reproductive organs.

"Build a baby that acts on the following instructions, and also contains a copy of those instructions in its reproductive organs."

von Neumann's Self-reproducing Automata

1. A universal constructor A.

$$A + \ulcorner X \urcorner \rightsquigarrow X$$

2. A copying machine B.

$$B + \ulcorner X \urcorner \rightsquigarrow \ulcorner X \urcorner$$

3. A control machine C, which first activates B, then A.

$$A + B + C + \ulcorner X \urcorner \rightsquigarrow X + \ulcorner X \urcorner$$

4. Let $X := A + B + C$. Then $A + B + C + \ulcorner A + B + C \urcorner$ is self-reproducing.

$$A + B + C + \ulcorner A + B + C \urcorner \rightsquigarrow A + B + C + \ulcorner A + B + C \urcorner$$

5. It is possible to add the description of any machine D.

$$A + B + C + \ulcorner A + B + C + D \urcorner \rightsquigarrow A + B + C + D + \ulcorner A + B + C + D \urcorner$$

6. Now allow mutation on the description $\ulcorner A + B + C + D \urcorner$.

$$A + B + C + \ulcorner A + B + C + D' \urcorner \rightsquigarrow A + B + C + D' + \ulcorner A + B + C + D' \urcorner$$

Introspective Program

Definition (ψ-introspective)

Given a total recursive function ψ,

- the ψ-*analysis* of $\varphi(x)$ is the code of the computation of $\varphi(x)$ to $\psi(x)$ steps.
- φ is ψ-*introspective* at x iff $\varphi(x)\downarrow$ and outputs its own ψ-analysis.
- φ is *totally* ψ-*introspective* iff it is ψ-introspective at all x.

Corollary

There is a program that is totally ψ-introspective.

Proof.

Let $f(n,x) :=$ "the ψ-analysis of $\varphi_n(x)$".

Introspective Program

There is a program that is totally introspective.

$$\varphi_{\mathbf{e}} = \varphi_{\mathbf{h}(\mathbf{e})}$$

Self-simulating Computer	**Self-consciousness**
Host Machine	Experiencing Self
Virtual Machine	Remembering Self
Hardware	Body

Who am I?

I think, therefore I am.

self-locating: "I" is an indexical term that I use to refer to myself as myself.

What is "me"?

What is "self-consciousness"?

- self-perception self-observation self-experience self-tracking self-reflection self-awareness
- self-evaluation self-analysis self-monitoring
- self-control self-adjustment self-modification self-actualization self-fulfillment self-surpass self-improvement
- *actual-self* pk *ideal-self* self-identity "the *self*"
- free will: Second order desire that we want to act on is second order volition. Second order volitions involve wanting a certain desire to be one's will, that is wanting it to move one to action. (Frankfurt)

- the split brain in man
- snow?
- shit!
- life as a story

Kahneman — Thinking, Fast and Slow

Figure: Why you might prefer more pain

- painful experiment
- experiencing self
- remembering self
- duration neglect
- peak-end rule

Figure: One can imagine a detailed floor plan of a room, sitting on a table in the room; this plan has an image of the table on which there is an image of the plan itself. Now introduce the dynamical aspect: the items on the plan are cut out from paper and can be moved to try a different furniture arrangement; in this way the plan models possible states of the world about which it carries information.

Manin — Cognitive Networks

The brain contains inside a map of itself, and some neural information channels in the central neural system:

- carry information about the mind itself, i.e., are reflexive;
- are capable of modelling states of the mind different from the current one, i.e., possess a modelling function;
- can influence the state of the whole mind and through that, the behavior, i.e., possess controlling function.

The reflection of the brain inside itself must be coarse grained.

Hofstadter — I am a Stange Loop

- Animate entities are those that, at some level of description, manifest a certain type of loopy pattern, which inevitably starts to take form if a system with the inherent capacity of perceptually filtering the world into discrete categories vigorously expands its repertoire of categories ever more towards the abstract.
- This pattern reaches full bloom when there comes to be a deeply entrenched self-representation — a story told by the entity to itself — in which the entity's "I" plays the starring role, as a unitary causal agent driven by a set of desires.

说谎者悖论	**我在说谎**
Grelling 悖论	“非自谓的”是自谓的吗
Russell 悖论	“不属于自身的集合的集合”属于自身吗
Berry 悖论	我是少于十八个字不可定义的最小数
Yablo 悖论	我下一句及后面所有的句子都是假的
Gödel 不动点引理	**我有性质** F
Tarski 算术真不可定义定理	我不真
Gödel 第一不完全性定理	我不可证
Gödel-Rosser 不完全性定理	对于任何一个关于我的证明，都有一个更短的关于我的否定的证明
Löb 定理	如果我可证，那么 A
Curry 悖论	如果我是真的，那么上帝存在
Parikh 定理	我没有关于自己的长度短于 n 的证明
Kleene 不动点定理	**我要进行** h **操作**
Quine 悖论	把“把中的第一个字放到左引号前面，其余的字放到右引号后面，并保持引号及其中的字不变得到的句子是假的”中的第一个字放到左引号前面，其余的字放到右引号后面，并保持引号及其中的字不变得到的句子是假的
自测量长度程序	我要输出自己的长度
自复制程序	我要输出自己
自反省程序	我要回顾自己走过的每一步
Gödel 机	我要变成能获取更大效用的自己

Schmidhuber's Gödel Machine

- The Gödel machine consists of a **Solver** and a **Searcher** running in parallel.
- The **Solver** ($\text{AIXI}^S/\text{AIXI}^{t\ell}$) interacts with the environment.
- The **Searcher** (LSEARCH/HSEARCH/OOPS) searches for a proof of "the modification of the software — including the *Solver* and *Searcher* — will increase the expected utility than leaving it as is".
- Logic: a theorem prover and a set of self-referential axioms, which include a description of its own software and hardware, and a description of the probabilistic properties of the environment, as well as a user-given utility function.
- *Since the utility of "leaving it as is" implicitly evaluates all possible alternative modifications, the current modification is globally optimal w.r.t. its initial utility function.*

Gödel Machine

- language $\mathscr{L} := \{\neg, \wedge, \vee, \rightarrow, \forall, \exists, =, (,), \ldots, +, -, \cdot, /, <, \ldots\}$
- well-formed formula
- utility function $u(s, e) = \mathbb{E}_\mu \left[\sum_{t=1}^{T} r_t \middle| s, e \right]$
- target theorem

$$u\big[s(t) \oplus \big(\mathrm{switchbit}(t) = 1\big), e(t)\big] > u\big[s(t) \oplus \big(\mathrm{switchbit}(t) = 0\big), e(t)\big]$$

- theorem prover

hardware, costs, environment, initial state, utility, logic/arithmetic/probability

Figure: Schmidhuber

Gödel Machine

- GRL
- Universal Search
- Self-Improvement
- Logic

Disadvantage: A Gödel Machine with a badly chosen utility function is motivated to converge to a "poor" program. (goal orthogonality!)

Gödel Machine vs Self-Consciousness vs Free Will?

Self-simulating Computer	Gödel Machine	Self-consciousness
Host Machine	Solver	Experiencing Self
Virtual Machine	Searcher	Remembering Self
Hardware	Hardware	Body

$$\boxed{\varphi_{\mathbf{e}} = \varphi_{\mathbf{h(e)}}}$$

$$\boxed{\text{self-reference} \overset{?}{\Longrightarrow} \text{self-improvement}}$$

Gödel Machines

1. *one-shot* self-improvement: Kleene's fixpoint theorem

$$\varphi_e = \varphi_{h(e)}$$

   - global optimality?
   - goal orthogonality? ends vs means
2. *continuous* self-improvement: Kleene's fixpoint theorem with parameters

$$\varphi_{e(y)} = \varphi_{h(e(y),y)}$$

   - "real-time" optimality. human-computer interaction?
   - intelligent explosion / technological singularity???
     continuous self-improvement ≠ exponential iteration
3. *beyond computability*: Kleene's relativized fixpoint theorem

$$\varphi^A_{e(y)} = \varphi^A_{h(e(y),y)}$$

   - Gödel Machine PK AIXI$^{t\ell}$
   - Gödel Machine PK AIXI

Limitation

1. Gödel's first incompleteness theorem / Rice's theorem
2. Gödel's second incompleteness theorem

$$\mathrm{T} \vdash \Box_{\mathrm{T}'} A \to A \implies \mathrm{T} \vdash \mathrm{Con}(\mathrm{T}')$$

   - Biological Evolution: Darwin PK Lamarch
   - Life3.0
3. Legg's incompleteness theorem. *General prediction algorithms must be complex. Beyond a certain complexity they can't be mathematically discovered.*
4. Complexity: higher-level abstractions — coarse grained.
   - Psychology: Duration neglect / Peak-end rule
   - Information Bottleneck: $\boxed{\text{Learning is to forget!}}$
5. Physical constraint: If we assume that it is not possible to measure properties without changing them (observer effect: α is fixpoint-free), then there is a limit to self-inspection.

Consciousness — Integrated Information Theory

- Experts do their specialties best when they're in a state of "flow", aware only of what's happening at a higher level, and unconscious of the low-level details of how they're doing it. The information processing that we're consciously aware of is merely the tip of the iceberg. Many behaviors and brain regions are unconscious, with much of our conscious experience representing an after-the-fact summary of vastly larger amounts of unconscious information. Consciousness lags behind the outside world by about a quarter second. Brain measurements can sometimes predict your decision before you become conscious of having made it.
- Consciousness requires a kind of information processing that's fairly autonomous and integrated, so that the whole system is rather autonomous but its parts aren't. Given a physical process that, with the passage of time, transforms the initial state of a system into a new state, its integrated information measures inability to split the process into independent parts. In other word, it measures how much different parts of a system know about each other.

Properties of Conscious Experience — Tononi

Existence	Consciousness exists. "I experience therefore I am".
Composition	Consciousness is structured: each experience consists of multiple aspects in various combinations.
Information	Consciousness is informative: each experience differs in its particular way from other possible experiences.
Integration	Consciousness is integrated: each experience is irreducible to non-interdependent components. Every part must be able to both affect and be affected by the rest of the system.
Exclusion	Consciousness is exclusive: each experience has definite borders; each experience has a particular spatial and temporal grain.

Integrated Information Theory (IIT)

- Suppose given a stochastic dynamical system, where the state of the system at time n is described by a set of random variables $\{X_i = X_i^{(n)}\}_{i=1}^N$ which correspond to a partition of the system into N subsystems, and the state at time $n+1$ by $\{Y_i = X_i^{(n+1)}\}_{i=1}^N$.
- The full system including all the mutual influences between these two sets of variables is described by $P(X, Y)$.
- Integrated information is meant to capture the difference between $P(X, Y)$ and an approximation $Q(X, Y)$ where only certain kinds of mutual influences are retained.
- These are usually taken to be the interdependencies between the variables at the same time and between each X_i and the corresponding Y_i, removing the dependencies of the Y_i from the X_j with $j \neq i$.

$$Q(Y_i|X) = Q(Y_i|X_i) \quad \text{for } i = 1, \ldots, N.$$

$$\begin{tikzcd} x_1 \arrow[r] \arrow[d, dash] & y_1 \arrow[d, dash] \\ x_2 \arrow[r] \arrow[ur] & y_2 \end{tikzcd} \qquad \begin{tikzcd} x_1 \arrow[r] \arrow[d, dash] & y_1 \arrow[d, dash] \\ x_2 \arrow[r] & y_2 \end{tikzcd}$$

- Given a partition λ

$$\{(X,Y)\} = \coprod_{i=1}^{N} \{(X_i, Y_i)\}$$

Consider the space $\mathcal{M}_\lambda := \{Q : Q(Y_i|X) = Q(Y_i|X_i) \text{ for } i = 1, \ldots, N\}$.

- The best approximation to $P(X,Y)$ by $Q(X,Y)$ in $\mathcal{M}_\lambda$ is

$$Q^*_\lambda := \operatorname*{argmin}_{Q \in \mathcal{M}_\lambda} D_{\mathsf{KL}}(P \| Q)$$

- Then the integrated information, for a given partition λ, is defined as

$$\Phi_\lambda := D_{\mathsf{KL}}(P \| Q^*_\lambda) = \min_{Q \in \mathcal{M}_\lambda} D_{\mathsf{KL}}(P \| Q)$$

with a further minimization over the choice of the partition,

$$\Phi := \min_{\lambda} D_{\mathsf{KL}}(P \| Q^*_\lambda) = \min_{Q \in \bigcup_\lambda \mathcal{M}_\lambda} D_{\mathsf{KL}}(P \| Q)$$

IIT — another version

$$\mathcal{M}_\lambda := \left\{ Q : Q(Y|X) = \prod_{i=1}^{N} Q(Y_i|X_i) \right\}$$

Φ measures inability to tensor factorize $M = M^A \otimes M^B$

- $\Phi := \min_{Q \in \bigcup \mathcal{M}_\lambda} D_{\mathsf{KL}}(P\|Q) = \sum_i H(Y_i|X_i) - H(Y|X)$
- counterexample

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1^0 & \dots & 1^{N-1} \\ \vdots & \ddots & \vdots \\ N^0 & \dots & N^{N-1} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$$

Non-operational Self-inspection?

The information available to the observer regarding his own state could have absolute limitations, by the laws of nature.

— von Neumann

$$\begin{tikzcd}
\mathbf{M} \times \mathbf{M} \arrow[r, "f"] & \mathbf{O} \arrow[d, "\alpha"] \\
\mathbf{M} \arrow[u, "\Delta"] \arrow[r, "g"'] & \mathbf{O}
\end{tikzcd}$$

- $\mathbf{M}$: quantum measurements.
- $\mathbf{O}$: possible outcomes of quantum measurements.

If we assume that it is not possible to measure properties without changing them (observer effect: α is fixpoint-free), then there is a limit to self-inspection.

Self-modification

Figure: Policy/utility self-modification. $a_t = \langle \check{a}_t, \pi_{t+1} \rangle$ or $a_t = \langle \check{a}_t, u_{t+1} \rangle$

External/Internal Wireheading & Free Will[2]

1. 我喜欢马。指鹿为马!
2. 我喜欢马。我意欲自己喜欢鹿! 我喜欢鹿!

[2] Everitt, Filan, Daswani, Hutter: Self-modification of policy and utility function in rational agents.
Frankfurt: Freedom of the will and the concept of a person.
Aaronson: The ghost in the quantum turing machine.
Calude, Kroon, Poznanovic: Free will is compatible with randomness.

Self-modification

Definition (Different Agents)

- Hedonistic Value

$$Q^{\mathrm{h},\pi}(æ_{<t}a_t) = \sum_{e_t \in \mathcal{E}} \rho(e_t|\check{æ}_{<t}\check{a}_t)\left[u_{t+1}(\check{æ}_{1:t}) + \gamma V^{\mathrm{h},\pi}(æ_{1:t})\right]$$

- Ignorant Value

$$Q_t^{\mathrm{i},\pi}(æ_{<k}a_k) = \sum_{e_t \in \mathcal{E}} \rho(e_t|\check{æ}_{<t}\check{a}_t)\left[u_t(\check{æ}_{1:k}) + \gamma V_t^{\mathrm{i},\pi}(æ_{1:k})\right]$$

- Realistic Value

$$Q_t^{\mathrm{r}}(æ_{<k}a_k) = \sum_{e_t \in \mathcal{E}} \rho(e_t|\check{æ}_{<t}\check{a}_t)\left[u_t(\check{æ}_{1:k}) + \gamma V_t^{\mathrm{r},\pi_{t+1}}(æ_{1:k})\right]$$

Self-modification — Realistic Agent

$$V_t^\pi(æ_{<k}) = Q_t(æ_{<k}\pi(æ_{<k}))$$

$$Q_t(æ_{<k}a_k) = \sum_{e_k \in \mathcal{E}} \rho(e_k|\check{æ}_{<k}\check{a}_k)\left[u_t(\check{æ}_{1:k}) + \gamma V_t^{\pi_{t+1}}(æ_{1:k})\right]$$

$$\pi_1^* := \operatorname*{argmax}_\pi V_1^\pi(\epsilon)$$

Theorem (All optimal policies are non-modifying)

Let ρ and u_1 be modification-independent. For every $t \geq 1$, for all percept sequences $e_{<t}$, and for the action sequence $a_{<t}$ given by $a_i = \pi(æ_{<i})$, we have

$$Q_1(æ_{<t}\pi_t(æ_{<t})) = Q_1(æ_{<t}\pi_1^*(æ_{<t}))$$

All realistic optimal policies are non-modifying.

Not wireheading; But orthogonal!

Orthogonality and Wireheading in Self-improving GRL

$$V_t^{\pi}(æ_{<k}) := Q_t(æ_{<k}\pi(æ_{<k}))$$

$$Q_t(æ_{<k}a_k) := \sum_{e_k \in \mathcal{E}} \sum_{\nu \in \mathcal{M}} w^{\nu}_{æ_{<k}} \nu(e_k|\check{æ}_{<k}\check{a}_k) \left[\sum_{u \in \mathcal{U}} \sum_{a_k^H} P(a_k^H|a_k) P(u \mid {}^{\pi_{t+1}}_{\nu}, æ_{<k}a_k, æ^H_{<k}a^H_k) u(\check{æ}_{1:k}) + \gamma V_t^{\pi_{t+1}}(æ_{1:k}) \right]$$

$$\pi_t(æ_{<k}) := \operatorname*{argmax}_{a_k \in \check{\mathcal{A}} \times \Pi} Q_t(æ_{<k}a_k)$$

where

$$P(u|{}^{\pi}_{\nu}, h) := \frac{\tilde{U}(u, {}^{\pi}_{\nu}, h)}{\sum_{u \in \mathcal{U}_h} \tilde{U}(u, {}^{\pi}_{\nu}, h)}$$

and

$$\tilde{U}(u, {}^{\pi}_{\nu}, h) := \sum_{z \in \mathcal{Z}_h} {}^{\pi}_{\nu}(z|h) u(z)$$

$$\pi^*(æ_{<t}) := \pi_t(æ_{<t})$$

$$\pi^*(e_{<t}) := \pi_t(e_{<t}|\pi_{t-1}(e_{<t-1}|\ldots\pi_1(\epsilon)\ldots)) \qquad \text{(Perfect Bayes-Nash)}$$

uncertain model-based utility / IRL

Fatalism — God Bless AI!

Orseau's Space-Time Embedded Intelligence

$$\pi^* := \operatorname*{argmax}_{\pi_0 \in \Pi^\ell} V(\pi_0, \epsilon)$$

$$V(\pi_t, æ_{<t}) := \sum_{a_t = \langle \check{a}_t, \check{\pi}_{t+1} \rangle} \pi_t(a_t | \check{e}_{t-1}) \sum_{e_t = \langle \check{e}_t, \pi_{t+1} \rangle} \rho(e_t | æ_{<t} a_t) \left[u(æ_{1:t}) + \gamma_t V(\pi_{t+1}, æ_{1:t}) \right]$$

$$\Downarrow$$

$$\pi^* := \operatorname*{argmax}_{\pi_0 \in \Pi^\ell} V(\pi_0)$$

$$V(\pi_{<t}) := \sum_{\pi_t \in \Pi} \rho(\pi_t | \pi_{<t}) \left[u(\pi_{1:t}) + \gamma_t V(\pi_{1:t}) \right]$$

GÖDEL
MACHINE

Incompressibility vs Incompleteness vs Intelligence

- $P(x) := \{p \in \mathcal{X}^* : \exists m \forall n \geq m \, (p(x_{1:n}) = x_{n+1})\}$
- $P(A) := \bigcap_{x \in A} P(x)$
- $P_n := P(\{x : Km(x) \leq n\})$

- $\forall n \exists p \in P_n : K(p) \overset{+}{\leq} n + O(\log n)$
- $\forall n : p \in P_n \implies K(p) \overset{+}{\geq} n$

Theorem (Legg)

For any arithmetically sound Gödelian T, $\exists n \forall p : \mathrm{T} \nvdash p \in P_n$.

Universal Artificial Intelligence vs "Selective Amnesia"

- incomplete $\xrightarrow[\text{universal prior}]{\text{Harsanyi transformation}}$ imperfect $\implies$ AIXI
- AIXI $\xrightarrow{\text{"Selective Amnesia"}}$ MDP

$$h \mapsto S$$

- partition of the set of histories = information set = state = feature
- $$\overset{\pi}{\xi}(h'|ha) \to \overset{\pi}{\mu}(h'|ha)$$

  but

  $$\overset{\pi}{\xi}(S'|Sa) \nrightarrow \overset{\pi}{\mu}(S'|Sa)$$

"Selective Amnesia"

1. compressible

$$K(S) \leq \sum_{h \in S} K(h)$$

2. minimal

$$\forall \operatorname{part}(S): \; K(S) \leq \sum_{S_i \in \operatorname{part}(S)} K(S_i)$$

3. maximal

$$\forall S' \supset S: \; K(S) \leq K(S')$$

4. MDL

$$\forall S' \in \mathcal{S} \forall h \in S': \; K(S) + K(h|S) \leq K(S') + K(h|S')$$

5. MDL(utility)

$$\forall \mathcal{S}': \; K\left(S^{\mathcal{S}}_{1:n} | a_{1:n}\right) + K\left(u_{1:n} | S^{\mathcal{S}}_{1:n}, a_{1:n}\right) + K(\mathcal{S}) \leq K\left(S^{\mathcal{S}'}_{1:n} | a_{1:n}\right) + K\left(u_{1:n} | S^{\mathcal{S}'}_{1:n}, a_{1:n}\right) + K(\mathcal{S}')$$

$$u(h) := \left[\!\left[K(h) < \ell(h) \;\; \& \;\; \forall h' \succ h \left(K(h) \leq K(h') \;\; \& \;\; \forall \operatorname{part}(h) \left(\sum_{h' \in \operatorname{part}(h)} K(h') \geq K(h) \right) \right) \right]\!\right]$$

Potapov's MSearch + RSearch

- Let $\{x_i\}_{i=1}^n$ be a set of strings.
- $K(x_1 \dots x_n) \approx \min_S \left(\ell(S) + \sum_{i=1}^n K(x_i|S) \right) \ll \sum_{i=1}^n K(x_i)$
- search for models $y_i^* := \operatorname*{argmin}_{y:S(y)=x_i} \ell(y)$ for each x_i w.r.t. some best representation $S^* := \operatorname*{argmin}_S \left[\ell(S) + \sum_{i=1}^n \ell(y_i^*) \right]$

1. Search for models

$$\mathrm{MSearch}(S, x_i) \to y_i^* = \operatorname*{argmin}_{y:S(y)=x_i} \ell(y)$$

2. Search for representations

$$\mathrm{RSearch}(x_1 \dots x_n) \to S^* = \operatorname*{argmin}_S \left[\ell(S) + \sum_{i=1}^n \ell(y_i^*) \right]$$

- $\mathrm{MSearch}$ enumerates all models to find the shortest model: $S(y_i) = x_i$.
- $\mathrm{RSearch}$ enumerates all S and calls $\mathrm{MSearch}$ for each S.

Specialization and SS'-Search

Theorem (*smn* Theorem)

For any $m, n > 0$, there exists a primitive recursive function s_n^m of $m+1$ arguments s.t. for every Gödel number e of a partial recursive function with $m+n$ arguments

$$\varphi_{s_n^m(e,x_1,\ldots,x_m)} = \lambda y_1 \ldots y_n . \varphi_e(x_1, \ldots, x_m, y_1, \ldots, y_n)$$

$$\forall x : \mathrm{spec}(\mathrm{MSearch}, S)(x) = \mathrm{MSearch}(S, x)$$

$$S' := \mathrm{spec}(\mathrm{MSearch}, S) \implies \begin{cases} \forall x : S(S'(x)) = x \\ \ell(S) + \sum_{i=1}^{n} \ell(S'(x_i)) \to \min \end{cases}$$

- S is a generative representation. (decoding)
- S' is a descriptive representation. (encoding)
- SS'-Search simultaneous search for S and S'.

Potapov's Representational MDL

$$K(x_{1:n}) \approx \min_S \left(\ell(S) + \sum_{i=1}^{n} K(x_i|S) \right) \ll \sum_{i=1}^{n} K(x_i)$$

$$q_1^* := \operatorname*{argmin}_q \left[\ell(q) + K(x|S_1 q) \right]$$

$$q_{i+1}^* := \operatorname*{argmin}_q \left[\ell(q) + K(q_i^*|S_{i+1} q) \right]$$

$$L_{S_1 \ldots S_m}(x) := K(x|S_1 q_1^*) + \sum_{i=2}^{m-1} K(q_i^*|S_{i+1} q_{i+1}^*) + \ell(q_m^*)$$

$$a_k^* := \operatorname*{argmax}_{a_k} \max_{p:U(p,e_{<k})=a_{<k}a_k} \sum_{q:U(q,a_{<k})=e_{<k}} 2^{-\ell(q)} V_q^p(æ_{<k})$$

$$a_k^* := \operatorname*{argmax}_{a_k} \max_{p:U(p,e_{<k})=a_{<k}a_k} \sum_{\{q_i\}:U(S\{q_i\},a_{<k})=e_{<k}} 2^{-\ell(\{q_i\})} V_{\{q_i\}}^p(æ_{<k})$$

where $e_{<k} = e_{m_1+1:m_2} \dots e_{m_{n-1}+1:m_n}$, $m_1 = 0$, $m_n = k-1$, and $U(Sq_i a_{<k}) = e_{m_i+1:m_{i+1}}$.

$$Q(q_k = s, a_k = a) := \max_{p:U(p,e_{<k})=a_{<k}a} \sum_{\{q_i\}:q_k=s,U(S\{q_i\},a_{<k})=e_{<k}} 2^{-\ell(\{q_i\})} V_{\{q_i\}}^p(æ_{<k})$$

$$Q(q_k = s) := \max_{a_k} Q(q_k = s, a_k = a)$$

Fundamental Challenges

- What is a good optimality criterion?
- What is a "natural" UTM/prior?
- Prior vs universality
- Exploration vs exploitation
- Where should the reward come from?
- How should the future be discounted?
- How should agents reason about themselves (or other agents reasoning about itself)?
- What is a practically feasible and general way of doing induction and planning?
- AIXI in the multi-agent setting.
- Better variants/approximations.
- Training: To maximize informativeness of reward, one should provide a sequence of simple-to-complex tasks to solve, with the simpler ones helping in learning the more complex ones.

- A Blind Man in a Dark Room Looking for a Black Cat That Is Not There?

- The Singularity is Near?

Thanks