Learning the best way to learn how people learn

John Rust, Georgetown University

August 21, 2023

Too many uses and meanings of the word, “learning”

- I intentionally used this word three times in three different ways in the title of this lecture.
- Bayesian learning, Machine Learning, Deep Learning, Q-learning, Deep-Q-learning, supervised learning unsupervised learning, offline learning, online learning, real-time-learning. Have I missed anything?
- Not all of these uses of “learning” are what I would actually consider to be “learning”. In many cases they describe *algorithms* to *approximate a solution or function* or to *make a prediction.* They are not “learning” in the sense we think about in everyday life.
- Real life learning includes things like “potty training” “learning to tie your shoes” “learning to drive a car” “learning a foreign language” “learning how to multiply numbers” “learning to play piano” and “learning what you want to do with your life”
- Each of these types of human learning may involve very different neural processes. For example learning to play the piano may involve very different parts of the brain and underlying mechanisms that learning a foreign language.

Human learning: vastly more complex than machine learning

- We do not have adequate models that are anywhere close to capturing the myriad of ways people actually learn and adapt.
- *Human learning is intimately connected with experimentation and decision making*
- Not clear that any ML algorithm is capable of what humans do when learning (such as what we do in our research)
    1. deciding what questions are "interesting"
    2. deciding where/how to get information data and what sort of method to use to answer the question and learn about the topic of interest
    3. developing "mental models" that provide plausible initial guesses about the costs and benefits of different courses of action
    4. formulating plausible hypotheses and designing experiments to test these hypotheses

Human learning is dynamic and involves experimentation

- Human learning is dynamic, active and involves considerable creativity including experimentation on the fly
- Captured by the phrase “learning by doing”
- There are elementary dynamic theories of Bayesian and statistical learning, including Wald’s initial work on the sequential likelihood ratio test, but this is a fairly abstract and oversimplified model of most actual learning that we engage in on a daily basis.
- But to understand such complex behavior, it makes sense to start simple and focus on studying narrower, more well defined examples of human learning.
- Laboratory experiments is a good example of this, and it perfectly reflects that “meta principle” of how people learning: starting simple but choosing well designed experiments and data gathering to move the ball forward.

Most of the brain's processing is *sub-conscious*

> *"The first thing we learn from studying our own circuitry is a simple lesson: most of what we do and think and feel is not under our conscious control. The vast jungles of neurons operate their own programs. The conscious you – the I that flickers to life when you wake up in the morning – is the smallest bit of what's transpiring in your brain. Although we are dependent on the functioning of the brain for our inner lives, it runs its own show. Most of its operations are above the security clearance of the conscious mind. The I simply has no right of entry." David Eagleman, 2011, INCOGNITO: The Secret Lives of the Brain*

- Most likely much of our learning is also subconscious, such as *muscle memory* that comes from repetition of actions until they become nearly automatic such as occurs in many sports.

Intelligence depends on *mental models of reality*

In 1956 the neuroscientist Donald MacKay proposed that the visual cortex is fundamentally a machine whose job is to generate a model of the world. He suggested that the primary visual cortex constructs an internal model that allows it to anticipate the data streaming up from the retina. The cortex sends its predictions to the thalamus, which reports on the difference between what comes in through the eyes and what was already anticipated. The thalamus sends back to the cortex only that difference information – that is, the bit that wasn't predicted away. This unpredicted information adjusts the internal model so there will be less of a mismatch in the future. In this way, the brain refines its model of the world by paying attention to its mistakes."

David Eagleman, 2011, INCOGNITO: The Secret Lives of the Brain

Theory-based causal induction

> *Inducing causal relationships from observations is a classic problem in scientific inference, statistics, and machine learning. It is also a central part of human learning, and a task that people perform remarkably well given its notorious difficulties. People can learn causal structure in various settings, from diverse forms of data: observations of the co-occurrence frequencies between causes and effects, interactions between physical objects, or patterns of spatial or temporal coincidence. These different modes of learning are typically thought of as distinct psychological processes and are rarely studied together, but at heart they present the same inductive challenge-identifying the unobservable mechanisms that generate observable relations between variables, objects, or events, given only sparse and limited data.*

from Griffiths and Tenenbaum 2009 *Pscyhological Review*

Human learning usually takes a lot of effort

- When you learn unfamiliar new things it is often exhausting and taxing, even if you are curious to learn new stuff.
- You are "rewiring your brain" when you are learning, and this process of altering neuronal connections in the brain takes a lot of energy.
- But once we are "trained" after sufficient investment and practice, the new ideas become second nature and intuitive or "hard-wired" — and quasi-automatic.
- This is captured in Daniel Kahneman's 2011 book, *Thinking: Fast and Slow* "The book's main thesis is a differentiation between two modes of thought: 'System 1' is fast, instinctive and emotional; 'System 2' is slower, more deliberative, and more logical. *Wikipedia* "System 1 thinking is a near-instantaneous process; it happens automatically, intuitively, and with little effort. It's driven by instinct and our experiences. System 2 thinking is slower and requires more effort. It is conscious and logical."

Training artificial deep nets is also energy and time-intensive

- Training a deep neural network involves searching for parameters of the network that minimize a given criterion function, such as mean squared error, or the negative of a log-likelihood function.
- This training process takes time and is energy-intensive for the computer. You can hear your laptop fan go on when you are training a big, deep neural network because it is taxing on the CPU (or GPUs). This may be similar to the "rewiring" that goes on in our own brains when we are trying to learn new stuff.
- But once this "training cost" is paid, the execution of the deep net is very very fast and almost instantaneous and requiring much less effort/energy by the computer.
- So if you find you are getting tired during my lecture, hopefully it is because I am presenting a lot of new information that your deep deep neural networks are trying to absorb.
- But with some effort and patience, hopefully your System 2 will learn new skills that it can hand over to your System 1 and become more intuitive, less confusing, and easier for you with some time and *learning by doing.*

Machine Learning vs Structural Estimation

- The title of a survey article by myself, Bertel and Fedor in 2020 titled *Machine Learning and Structural Estimation: Contrasts and Synergies*
- We started with a section titled "How does ML differ from SE and will ML put us out of work?"
- Then we tried to define the key differences between ML and SE

  ML and SE share a common interest in prediction and decision making, but the goal of ML is to enable computers to do these tasks whereas SE is focused on how humans do them. Thus ML is more practically oriented, in trying to automate tasks that previously only humans could do well, whereas SE, for reasons we discuss below, has been more academically oriented, in trying to understand and model human economic behaviour."

Example: Credit Default

- Simon talked about using big data on mortgages and deep learning methods to predict *mortgage default.*
- The focus of ML and banks is on *predicting who defaults and the probability of and loss from a default.*
- But in structural econometrics we are more interested in *understanding what causes people to default* and how economic/legal institutions and conditions and incentives affect the default decision.
- One of my former students, Yangfan Sun, used big data on 37 million 30 year mortgages acquired by Fannie Mae between 2000 and 2016 and a structural model to show how different bankruptcy laws in different states affect the likelihood of mortgage default. Some states are "non-judicial" which makes it easier for the bank to foreclose and evict a delinquent borrower, others are "judicial staties" where it there is a longer more difficult legal process to foreclose.

Using structural models for *counterfactual prediction*

- Fannie Mae charges an insurance premium to cover the costs of default called the *gaurantee fee* but this fee does not differ between judicial and non-judicial states.
- This leads to a cross-subsidization since default costs are higher for lenders in judicial than non-judicial states.
- Yangfan's goal was to quantify the magnitude of this cross subsidization and the impact on foreclosure rates if Fannie Mae were to differentiate the guarantee fee to reflect the higher costs of foreclosure in judicial states.
- Doing this requires a structural model since this exercise requires *counterfactual predictions* of how mortgage default behavior would change as a result of the change in mortgage interest rates induced by a change in guarantee fees, which would raise mortgage interest rates to borrowers in judicial states and lower them to borrowers in non-judicial states.

Can ML do counterfactual prediction?

Our article discussed the tremendous success of Alpha-zero in "self-training" using *deep Q learning* to become the world-champion in chess in just a matter of hours of wall-clock time.

> *"Can ML handle these sorts of counterfactual prediction exercises? A simple thought experiment suggests that, at least for the board game example, the answer is yes. We can imagine redoing the Q-learning exercise, but instead of under the existing rules of chess, we impose the counterfactual new rules* Π*. Then we train the algorithm the same way as it was trained under the current rules of chess, resulting ultimately in its learning (i.e., converging to) a new strategy* δ_Π *that is optimal under the new rules* Π*."*

An actual example of counterfactual prediction by ML

"RCTs Against the Machine: Can Machine Learning Prediction Methods Recover Experimental Treatment Effects?" by Prest, Wichman and Palmer, 2021.

> *"This paper compares treatment effect estimates from a randomized electricity pricing and information experiment to non-experimental effects estimated using both standard difference-in-difference methods and three commonly used ML counterfactual prediction methods (XGBoost, random forests, and LASSO) in a 'prediction-error' framework where predicted outcomes serve as counterfactuals."*
>
> *"We find that ML counterfactual prediction methods can replicate experimental treatment effects remarkably well with relatively little data on predictors."*

So, is structural econometrics already obsolete?

In our survey, we noted the angst: "In the course of trying to better understand and replicate ourselves, we may be collectively paving the way to our own ultimate demise, as Benzell et al. (2015) suggested in their paper "Robots R Us: Some Economics of Human Replacement"

> *"Will smart machines do to humans what the internal combustion engine did to horses – make them obsolete?"*

However we concluded on a less apocalyptic note.

> *"At least in the short run, we see great opportunities and no immediate threats to our existence as structural econometricians from progress on ML."*

ML is a complement and not substitute for SE

And we quoted my thesis advisor, Dan McFadden:

> *"What are the lessons here for econometricians? You should not simply dismiss learning machines and their computer operators as deplorables. You should instead think of them as your worst possible students – ignorant, arrogant, and disinterested. If you can figure out how to break through and teach them to respect and use the scientific content of economics, they may prove to be your best research associates. To keep structural econometrics vigorous and relevant, you need to continue to move aggressively to embrace the innovations in data collection and analysis created by computational advances."*

ML, SE and Artificial Intelligence

- A big area of commonality between ML and SE (and economics more generally) is the perspective of *rationality*
- A big goal of ML is to create "rational robots" that can do more and more tasks formerly done by humans, even expert, such as looking at a mammogram to see if there is evidence of tumor cells (something usually done by radiologists and medical doctors)
- A big goal of economics and SE is to model *people as rational robots.* That is, we are starting with a null hypothesis that people behave "as if" they have a well defined intertemporal utility function and learn and take actions sequentially to maximize their expected welfare.
- Thus, while we may have angst that we are being superseded by computer scientists who can build and train rational robots better and faster than we can as economists, we should also realize how much mathematics and economics have contributed to the definition and implementation of "rationality" that core ideas of some of the biggest recent successes in AI such as reinforcement learning and deep Q learning.

Computational Rationality

The title of a 2015 review article in *Science* by 3 leading cognitive scientists.

> *"AI research has its roots in the theory of computability developed in the 1930s. Efforts then highlighted the power of a basic computing system (the Turing Machine) to support the real-world mechanization of any feasible computation. The promise of such general computation and the fast-paced rise of electronic computers fueled the imagination of early computer scientists about the prospect of developing computing systems that might one day both explain and replicate aspects of human intelligence."*

> *"Early pioneers in AI reflected about uses of probability and Bayesian updating in learning, reasoning, and action. Analyses by Cox, Jaynes, and others had provided foundational arguments for probability as a sufficient measure for assessing and revising the plausibility of events in light of perceptual data."*

Fundamental impact of Von Neumann and Morgenstern

Authors of the landmark 1947 book *Theory of Games and Economic Behavior*

> *"In influential work, von Neumann and Morgenstern published results on utility theory that defined ideal, or 'rational,' actions for a decision-making agent. They presented an axiomatic formalization of preferences and derived the 'principle of maximum expected of utility' (MEU). Specifically, they showed that accepting a compact and compelling set of desiderata about preference orderings implies that ideal decisions are those actions that maximize an agent's expected utility, which is computed for each action as the average utility of the action when considering the probability of states of the world. The use of probability and MEU decision-making soon pervaded multiple disciplines, including some areas of AI research, such as projects in robotics."*

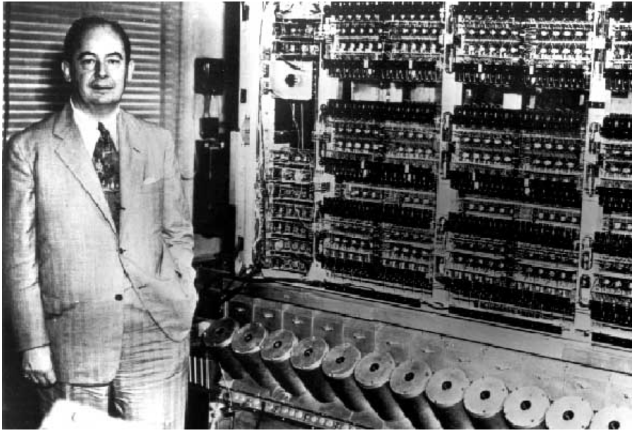
The importance of learning in AI

> *"In parallel with developments in AI, the study of human intelligence has charted a similar progression toward computational rationality. Beginning in the 1950s, psychologists proposed that humans are 'intuitive statisticians' using Bayesian decision theory to model intuitive choices under uncertainty. In the 1970s and 1980s, this hypothesis met with resistance from researchers who uncovered systematic fallacies in probabilistic reasoning and decision-making, leading some to adopt models based on informal heuristics and biases rather than normative principles of probability and utility theory. The broad success of probabilistic and decision-theoretic approaches in AI over the past two decades, however, has helped to return these ideas to the center of cognitive modeling."*

Do economists rationalize irrational behavior?

- Economists are very good at *rationalizing* human behavior, i.e. finding a rational explanation for behavior the rest of us might regard as obviously irrational. But we (as economists) need to be careful in doing that, however, because our scientific writings and statements can sometimes be taken seriously by the public.
- Even if people are not always well approximated as "rational robots" the development of tools for computing and implementing increasingly detailed and realistic models of intelligent decision making has contributed to AI, as the previous quotes from the article on "computational rationality" indicated. So economists are also contributing to AI, even though in doing that we may all be contributing to our own ultimate demise, to be superseded by more rational creatures with even bigger brains.

John von Neumann 1903-1957



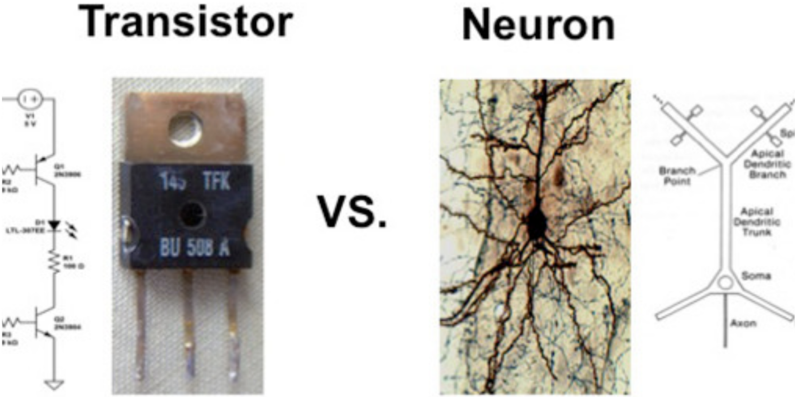
The ‘von Neumann machine”

- Name given to modern computers based on the original design by John von Neumann where
- “In March of 1953 there were 53 kilobytes of high-speed random access memory on planet Earth.” (from George Dyson (2012) *Turing’s Cathedral: The Origins of the Digital Universe*)
- How much RAM is there now? About 295 billion gigabytes and increasing by about *10 gigabytes per second!*
- There is even more long term digit storage: according to Andrew Beeson, “In 2018, the total amount of data created, captured, copied and consumed in the world was 33 zettabytes (ZB) — the equivalent of 33 trillion gigabytes. This grew to 59ZB in 2020 and is predicted to reach a mind-boggling 175ZB by 2025.”

Artificial vs human brains

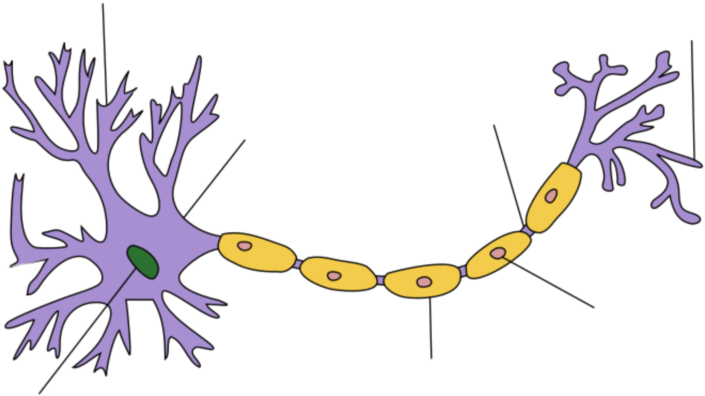
- The fundamental processing unit of a digital computer is not a neuron but an electronic switch called a *transistor.*
- The M1-Max Pro processor inside this computer has 57 billion transistors and consumes about 110 watts of power at full load.
- In comparison the neurons in the human brain fire at most about 1000 times per second, or about 360 million times more slowly than the transistors in the M1-Max chip can switch on/off, but it consumes only 10 watts of energy.
- Our brain has about 10 billion neurons, but each neuron connects on average to 1000 other neurons, so there are approximately 100 trillion synapses in the human brain. This dense communication network seems to be a key to the processing power of our brains
- And of course the average human brain is way, way smarter than the M1-Max, though the M1-Max can do things we cannot do such as add up a 200 billion numbers in less than a second.
- Our understanding of neurons and the brain is changing the way we build computers, to make them faster, more intelligent, and more energy efficient.

Transistor vs Neuron



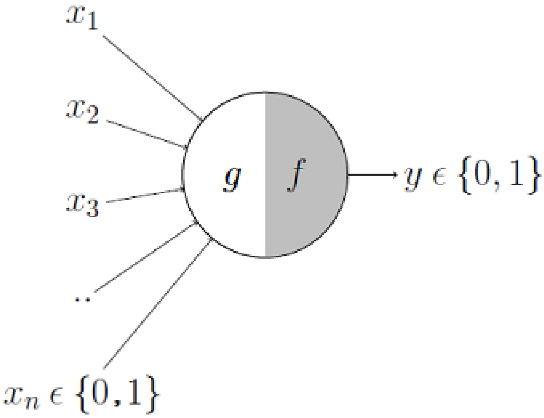
McGovern Institute for Brain Research at MIT

Schematic of a real neuron



The artificial neuron of Mcculloch and Pitts

"A Logical Calculus of the Ideas Immanent in Nervous Activity" *Bulletin of Mathematical Biology* 1943.



Artificial intelligence and the quest for "perfect rationality"

- Though the origins of *formal, conscious instances of human rationality* are part of the history of logic and mathematics going back as far as Mesopotamia 3000 BCE, much of it was counting for business account-keeping.
- We associate the logical deductive thinking via formal mathematical proofs from axioms with the ancient Greeks starting around the 600BCE, including the Pythagoreans and Euclid's *The Elements.*
- But the most important developments in math and logic that have lead to our current notion of *perfect rationality* came only recently in human history, in the 20th century.
- I will attempt to summarize the key developments and then illustrate how my own research relates to the mathematical/logical notion of perfect rationality.
- In short, the focus of most of my research has been on *testing empirically to see how well models of perfect rationality approximate actual human behavior in concrete economics contexts.*

Defining “perfect rationality”

- Stuart Russell, defines *perfect rationality* at the “capacity to generate maximally successful behaviour given the available information.” (Russell, 1997, “Rationality and Intelligence” *Artificial Intelligence*)
- This is close to my own definition, except that I allow less than perfect optimization, imperfect mental models of the world, and less the perfect ability to learn from evidence.
- We want to take care not to define perfect rationality in a way that implies *ominiscience* or god-like capabilities (though potentially “superhuman” capabilities).
- This quest might be a search for the holy grail, in the process of trying to define and create artificial intelligence and perfect rationality, humans have improved our understanding of ourselves and our own limitations, as well as machines.
- The unanswered question is whether it is possible to create true “artificial intelligence”. Is consciousness and rationality is a God-given gift, or an “emergent phenomenon” that appears at a sufficient level of complexity with millions/billions of communicating, cooperating but simpler “agents” operating in the “Society of Mind”?

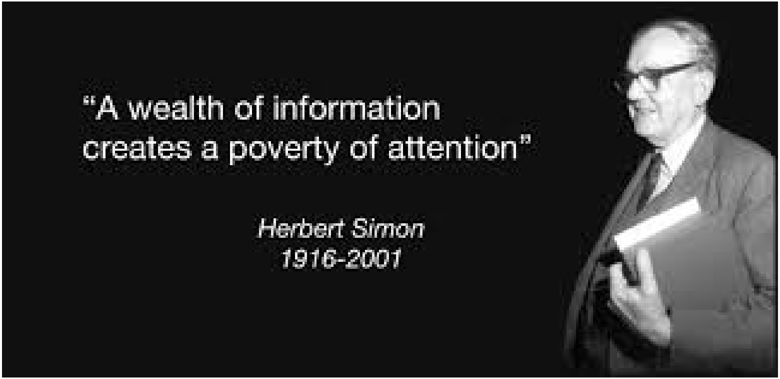
Herbert Simon: people are "boundedly rational"

Herbert Simon's definition of "bounded rationality" (*Wikipedia*)

Bounded rationality is the idea that rationality is limited by computational constraints and limited information and knowledge. In such situations, rational individuals will select a decision that is satisfactory rather than optimal. Decision-makers, in this view, act as *satisficers*, seeking a satisfactory solution using heuristics or intelligent "rules of thumb" rather than attempting to compute an optimal solution.

- Simon's notion is intended to be a more realistic alternative to the idea that people are *perfectly rational* (also known as *homo economicus*) that is so widely adopted among economists.
- "Broadly stated, the task is to replace the global rationality of economic man with the kind of rational behavior that is compatible with the access to information and the computational capacities that are actually possessed by organisms, including man, in the kinds of environments in which such organisms exist. (Simon 1955a: 99)" (quoted from Stanford Encyclopedia of Philosophy)

Herbert Simon



"A wealth of information creates a poverty of attention"

Herbert Simon
1916-2001

Implications of Bounded Rationality

> *"In the face of the combinatorial complexity of formulating and solving real-world decision-making, a school of research on heuristic models of bounded rationality blossomed in the later 1950s. Studies within this paradigm include the influential work of Simon and colleagues, who explored the value of informal, heuristic strategies that might be used by people – as well as by computer-based reasoning systems – to cut through the complexity of probabilistic inference and decision-making. The perspective of such heuristic notions of bounded rationality came to dominate a large swath of AI research."*

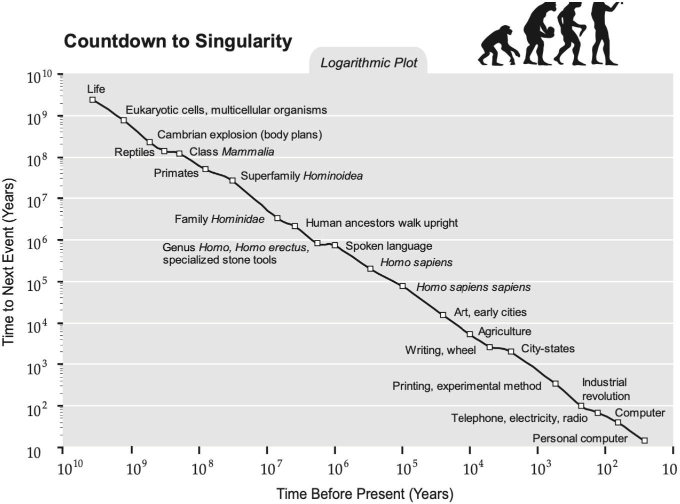
Accounting for "costs of cognition"

> *"If the brain is adapted to compute rationally with bounded resources, then 'fallacies' may arise as a natural consequence of this optimization. For example, a generic strategy for approximating Bayesian inference is by sampling hypotheses, with the sample-based approximation converging to the true posterior as more hypotheses are sampled. Evidence suggests that humans use this strategy across several domains, including causal reasoning, perception, and category learning. Sampling algorithms can also be implemented in biologically plausible neural circuits, providing a rational explanation for the intrinsic stochasticity of neurons."*

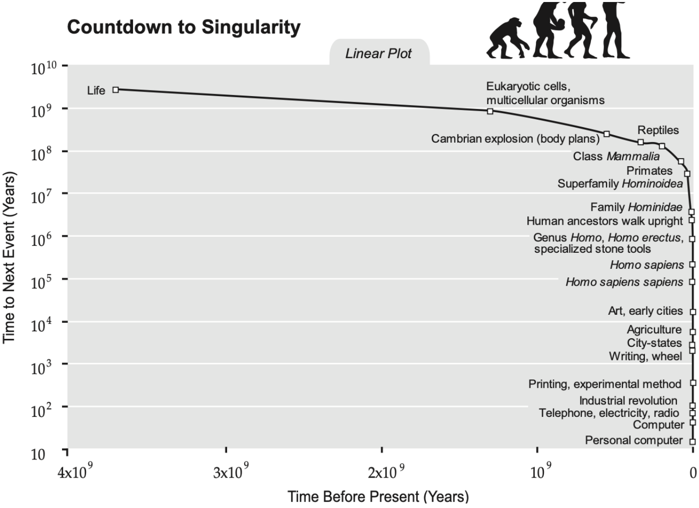
Accounting for "costs of cognition"

This argument rests crucially on the assertion that the brain is equipped with metareasoning mechanisms sensitive to the costs of cognition. Some such mechanisms may take the form of heuristic policies hardwired by evolutionary mechanisms; we call these 'heuristic' because they would be metarational only for the range of situations that evolution has anticipated. There is also evidence that humans have more adaptive metareasoning mechanisms sensitive to the costs of cognition in online computation. In recent work with the 'demand selection' task, participants are allowed to choose between two cognitive tasks that differ in cognitive demand and potential gains. Behavioral findings show that humans trade off reward and cognitive effort rationally according to a joint utility function. Brain imaging of the demand selection task has shown that activity in the lateral prefrontal cortex, a region implicated in the regulation of cognitive control, correlates with subjective reports of cognitive effort and individual differences in effort avoidance."

The Singularity is Near – loglog plot



The Singularity is Near – semilog plot



The Law of Accelerating Change

> *"Two billion years ago our ancestors were microbes; a half-billion years ago, fish; a hundred million years ago, something like mice, ten million years ago, arboreal apes; and a million years ago, proto-humans puzzling out the taming of fire. Our evolutionary lineage is marked by the mastery of change. In our time, the pace is quickening."*
>
> *Carl Sagan*

Kurzweil's techno optimistic view of future

> *"Our ability to create models – virtual realities – in our brains, combined with our modest looking thumbs, has been sufficient to usher in another form of evolution: technology. That development enabled the persistence of the accelerating pace that started with biological evolution. It will continue until the entire universe is at our fingertips."*
>
> *Ray Kurzweil, p. 487, concluding sentences of The Singularity is Near*

Hawking's techno pessimistic view of future

"While primitive forms of artificial intelligence developed so far have proved very useful, I fear the consequences of creating something that can match or surpass humans. Humans, who are limited by slow biological evolution, couldn't compete and would be superseded."

"If computers continue to obey Moore's Law, doubling their speed and memory capacity every eighteen months, the result is that computers are likely to overtake humans in intelligence at some point in the next hundred years. When an artificial intelligence (AI) becomes better than humans at AI design, so that it can recursively improve itself without human help, we may face an intelligence explosion that ultimately results in machines whose intelligence exceeds ours by more than ours exceeds that of snails. When that happens, we will need to ensure that the computers have goals aligned with ours. It's tempting to dismiss the notion of highly intelligent machines as mere science fiction, but this would be a mistake, and potentially our worst mistake ever."
Stephen Hawking (2018) Brief Answers to the Big Questions

Homo Deus

- Title of 2015 book by Yuval Harari
- This book "envisions a future in which technology replaces humanist ideals and liberal government. Dissecting the concepts of religion, immortality, and technology, Harari argues that the world of the future may be run by advanced algorithms and artificial intelligence, not human beings."

  "What will happen to society, politics and daily life when non-conscious but highly intelligent algorithms know us better than we know ourselves?"

The origins of “perfect rationality”

- A series of critical developments in the 20th century paved the way for our current concept of “perfect rationality”
- von Neumann’s work on games and the MiniMax Theorem in 1928
- The work of Bellman and others on dynamic programming in the late 1940s
- publication of “Theory of Games and Economic Behavior” by von Neumann and Morgenstern in 1944
- John Nash’s work on game theory and existence of “Nash equilibrium” in general non-cooperative games in 1951
- The pioneering effect of John von Neumann to invent the one of the first digital computer where both instructions and data are stored in memory “the von Neumann machine” in the mid 1950s
- These mathematical concepts and the physical tools provided by the new digital computer technology paved the way for many of the modern developments in economics and AI.

Dynamic programming and the "Principle of Optimality"

- Dynamic programming is a mathematical method for solving sequential decision making problems under uncertainty.
- The method is known as *backward induction*

The Principle of Optimality

An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision. Bellman, 1957 *Dynamic Programming*

- In game-theoretic language, the Principle of Optimality is equivalent to the concept of a *subgame-perfect equilibrium of a game against "nature"* and is also known as the *one shot deviation principle*
- *Bellman equation:* encodes the process of backward induction we use to solve these problems

The Bellman Equation

$$V(x) = \max_{d \in D(x)} \left[u(x,d) + \beta \int V(x') p(x'|x,d) \right]$$

or more abstractly

$$V = \Gamma(V)$$

so V is the *fixed point* of the Bellman operator Γ

Bellman's Curse of Dimensionality

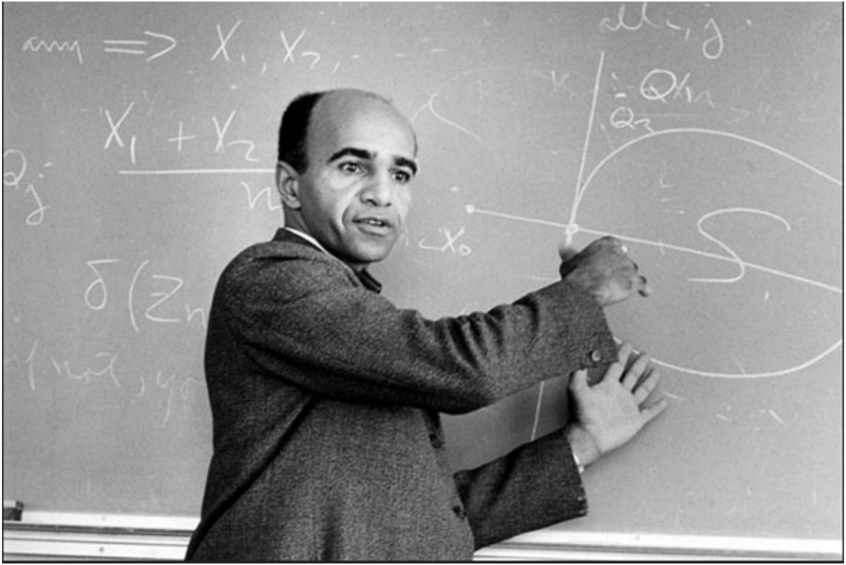
The difficulty of solving the Bellman equation increases exponentially fast in the dimension of the state variable x

Problems where x can take many possible values such as chess are way more difficult to solve than problems where x can take on only a much smaller number of possible values.

Pierre Massé 1898–1927

Kenneth A. Arrow 1921–2017

David Blackwell 1919–2010



Richard Ernest Bellman, 1920–1984

Origin of the term "Dynamic Programming"

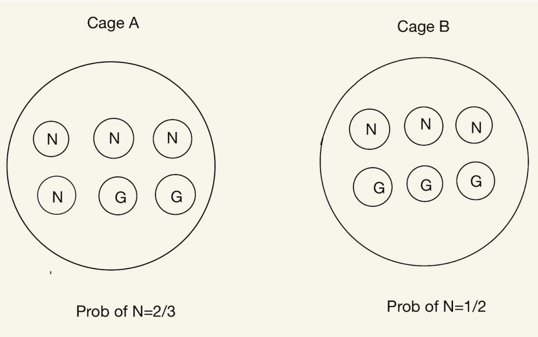
- From Bellman's autobiography, *The Eye of the Hurricane*
- "The 1950's were not good years for mathematical research. We had a very interesting gentleman in Washington named Wilson. He was Secretary of Defence, and he actually had a pathological fear and hatred of the word, research."
- "I'm not using the term lightly; I'm using it precisely. His face would suffuse, he would turn red, and he would get violent if people used the term, research, in his presence. You can imagine how he felt, then, about the term, *mathematical*."
- "Hence, I felt I had to do something to shield Wilson and the Air Force from the fact that I was really doing mathematics inside the RAND Corporation. What title, what name, could I choose?"

Origin of the term “Dynamic Programming”

- “In the first place, I was interested in planning, in decision-making, in thinking. But planning, is not a good word for various reasons.”
- “I decided therefore to use the word, ‘programming’. I wanted to get across the idea that this was dynamic, this was multistage, this was time-varying. I thought, let’s kill two birds with one stone.”
- “Let’s take a word which has an absolutely precise meaning, namely dynamic, in the classical physical sense. It also has a very interesting property as an adjective, and that is it’s impossible to use the word, dynamic, in the pejorative sense.”
- “Try thinking of some combination which will possibly give it a pejorative meaning. It’s impossible.”
- “Thus, I thought dynamic programming was a good name. It was something not even a Congressman could object to. So I used it as an umbrella for my activities.”

Are People Bayesian?

- Title of the classic 1995 article by El-Gamal and Grether in *Journal of American Statistical Association*
- Laboratory experiment with 257 subjects from 4 California universities (UCLA, Pasadena Community College, Occidental College and Cal State Univ Los Angeles)
- Subjects were shown samples of 6 balls drawn at random (with replacement) from one of two cages, A or B, and asked to predict which cage the sample came from.



Experimental design

- At the start of each trial, a 6 sided die was thrown. If it landed $\{3, 4, 5, 6\}$ cage A was selected, otherwise cage B was selected.
- Subjects were not shown the outcome of the dice throw or which cage used used to draw the sample of 6 balls, but only the *rule for selecting cage A or B based on the dice throw.*
- This induced a credible, objective *prior probability* of selecting cage A equal to $\pi_A = 2/3$, and of course cage B was selected with $\pi_B = 1 - \pi_A = 1/3$.
- The prior probability of drawing cage A, π_A, varied over the 3 values $\pi_A \in \{1/3, 1/2, 2/3\}$.
- Then 6 balls were drawn at random (with replacement) from the selected cage and subjects where shown the outcome of all 6 draws.
- Let n denote the number of balls marked N in the sample of 6 balls from the cage.
- Based on the information they were given $I = (\pi_A, n)$ subjects were asked to choose which of the two cages the sample was drawn from.

Experimental design

- All subjects were paid a flat fee just for participating in the experiment.
- However some subjects were paid a bonus $10 for selecting the correct cage in a randomly selected trial that the subject participated in.
- They refer to these as the *pay treatment* and *no pay treatment* respectively.
- A further key design choice: *no "pre-training" or other feedback to subjects while the trials occurred.*

  "In both treatments, subjects were not given any feedback on the correctness of their responses until the very end of the experiment, when their payoffs were computed."

  "The sessions lasted approximately 1 hours, and the number of decisions made by each subject ranged from 14 to 21."

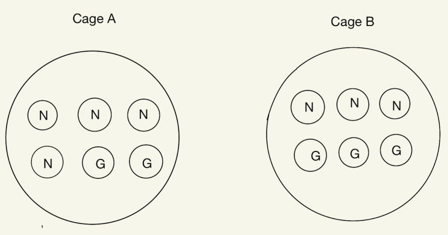
Why don't you try it?

- Suppose that $\pi_A = 2/3$ so the probability of using cage A to draw the 6 ball sample is 2/3. Since probabilties sum to 1, the probability of selecting cage B is $\pi_B = 1/3$.
- One of the cages is selected and 6 balls are drawn from it with replacement. Suppose the sample drawn is

$$(N, G, G, N, N, G)$$

- Recall that cage A has 2/3 fraction of N balls and cage B has 1/2 fraction of N balls. Which of the two cages is more likely to have been the cage from which the sample above was drawn?

https://editorialexpress.com/are_people_bayesian



El-Gamal and Grether's Model

- Assume subjects use *cutoff rules* to choose cage A or B
- **Cutoff Rule** An integer $c \in \{0, 1, \ldots, 6\}$ such that subject chooses cage A if $n > c$ otherwise choose B.
- In the experiment there are 3 priors $\pi \in \{1/3, 1/2, 2/3\}$ so let c_π denote the cutoff rule corresponding to prior π.
- "Ignoring the order of draws, there are seven possible outcomes (zero through six N's) and three priors, resulting in 21 possible decision situations. In each of these situations the subject could choose either cage A or cage B. Therefore there are in principle $2^{21} = 2,097,052$ possible decision rules."
- However there are only $8^3 = 512$ possible cutoff rules, (c_1, c_2, c_3) since each cutoff c_i for prior π_i, $i \in \{1, 2, 3\}$ can take 8 possible values $c_i \in \{-1, 0, \ldots, 6\}$.
- Note that *Bayes Rule* corresponds to the cutoffs $(c_1, c_2, c_3) = (4, 3, 2)$.

Generating a likelihood function

- Let $x_{s,t}(c) = 1$ if the choice of subject s on trial t is consistent with the cutoff rule $c = (c_1, c_2, c_3)$. Note that no single cutoff rule will generally be able to "explain" all choices of any given subject.
- To derive a non-zero likelihood for the observations, El-Gamal and Grether assumed that with probability ε the subject guesses (in effect flips a coin), whereas with probability $1 - \varepsilon$ the subject's decision is governed by the cutoff rule c.
- This implies a *non-degenerate likelihood* i.e. every observed choice will have positive probability for any cutoff rule c. Define the sufficient statistic $X_s(c) = \sum_{t=1}^{t_s} x_{st}(c)$, the number of the t_s choices by subject s that are consistent with cutoff rule c. Then the likelihood for all subjects is

$$L(c,\varepsilon) = \prod_{s=1}^{S} L(X_s(c)|\varepsilon), \text{ where } L(X_s(c)|\varepsilon) = \left(1 - \frac{\varepsilon}{2}\right)^{X_s(c)} \left(\frac{\varepsilon}{2}\right)^{1-X_s(c)}$$

and the MLE is $(\hat{c}, \hat{\varepsilon}) = \operatorname{argmax}_{c,\varepsilon} L(c,\varepsilon)$.

Allowing for subject heterogeneity: the EC algorithm

- This likelihood assumes subjects are *homogeneous* – they have the same probability ε of guessing and use the same cutoff rule c. If subjects are different can *unsupervised learning* discover their types?
- With *panel data* we can allow for heterogeneity using *fixed effects* – estimate *subject-specific* parameters $(\hat{c}^s, \hat{\varepsilon}^s) = \mathrm{argmax}_{c,\varepsilon} L(X_c^s|c,\varepsilon)$.
- However subjects participated in relatively small numbers of trials: $t_s = 19$ or 20 for most subjects. We have 4 unknown parameters (c,ε) with only 19 parameters per subject, so not many "degrees of freedom" and a potential *incidental parameters problem* leading to poor performance of fixed effects maximum likelihood.
- **Solution:** The EC (Estimation-Classification) Estimator. Suppose we restrict the number of "types" to be $K < S$. Let (c^k, ε^k) be the cutoff rule and error rate of a "type k" subject. Let the indicator $\delta_{s,k} = 1$ if subject s is "assigned" to be a type k. Then for a fixed number of types K the EC algorithm estimates the K types $(\hat{c}^1,\ldots,\hat{c}^K,\hat{\varepsilon}^1,\ldots,\hat{\varepsilon}^K)$ as follows

$$(\hat{c}^1,\ldots,\hat{c}^K,\hat{\varepsilon}^1,\ldots,\hat{\varepsilon}^K,\{\hat{\delta}_{s,k}\}) = \underset{\vec{c},\vec{\varepsilon},\{\delta_{s,k}\}}{\mathrm{argmax}}\, L(\vec{c},\vec{\varepsilon},\{\delta_{s,k}\}). \quad (1)$$

EC likelihood

- where $L(\vec{c}, \vec{\varepsilon}, \{\delta_{s,k}\})$ is given by

$$L(\vec{c}, \vec{\varepsilon}, \{\delta_{s,k}\}) = \prod_{s=1}^{S} \prod_{k=1}^{K} L(X^{s}_{c^k} | c^k, \varepsilon^k)^{\delta_{s,k}}, \qquad (2)$$

  and the maximization of $L(\vec{c}, \vec{\varepsilon}, \{\delta_{s,k}\})$ is done subject to the constraint that $\delta_{s,k} \in \{0, 1\}$ and $\sum_{k=1}^{K} \delta_{s,k} = 1$ for all subjects $s \in \{1, \dots, S\}$. *Each subject can be assigned to only one of the K types $(c^1, \dots, c^K, \varepsilon^1, \dots, \varepsilon^K)$ and $\hat{\delta}_{s,k} = 1$ denotes the type choice for subject k that has the highest likelihood $L(X^{s}_{c^k} | c^k, \varepsilon^k)$ across the $k = \{1, \dots, K\}$ types.*
- Thus the EC algorithm consists of an "outer loop" that searches over $(c^1, \dots, c^K, \varepsilon^1, \dots, \varepsilon^K)$ and an "inner loop" that for each subject s sets $\hat{\delta}_{s,k} = 1$ for the type k for which $L(X^{s}_{c^k} | c^k, \varepsilon^k)$ is the largest.
- The total number of parameters for EC is $4 * K$ vs $4 * S$ in a fixed effects estimation approach, which in effect is trying to estimate infinitely many parameters as $S \to \infty$ with $t/4$ observations per parameter, whereas it equals $St/4k$ for EC.

Allowing for subject heterogeneity: random effects

- Heckman and Singer (1984) showed a consistent estimator with unknown hetergeneity is the following *random effects* estimator, under the assumption that there are a fixed number $K < S$ types of subjects. The random effects estimator estimates the "types" $(\hat{c}^1, \ldots, \hat{c}^k, \hat{\varepsilon}^1, \ldots, \hat{\varepsilon}^K)$ along with the probabilities $(\hat{\mu}^1, \ldots, \hat{\mu}^K)$ that any given subject is one of these K types via the *mixture likelihood*

$$(\hat{c}^1, \ldots, \hat{c}^K, \hat{\varepsilon}^1, \ldots, \hat{\varepsilon}^K, \hat{\mu}^1, \ldots, \hat{\mu}^K) = \underset{\vec{c},\vec{\varepsilon},\vec{\mu}}{\operatorname{argmax}} L(\vec{c}, \vec{\varepsilon}, \vec{\mu}), \tag{3}$$

where

$$L(\vec{c}, \vec{\varepsilon}, \vec{\mu}) = \prod_{s=1}^{S} \left[\sum_{k=1}^{K} L(X^s_{c^k} | c^k, \varepsilon^k, \mu^k) \mu^k \right], \tag{4}$$

subject to the constraint that $\sum_{k=1}^{K} \mu^k = 1$.

- Note that the maximized value of the EC likelihood, $L(\vec{c}, \vec{\varepsilon}, \vec{\{\delta_{s,k}\}})$ will exceed the maximized value of the random effects likelihood $L(\vec{c}, \vec{\varepsilon}, \vec{\mu})$ since EC assigns each subject to their most likely type.

El-Gamal and Grether's empirical findings

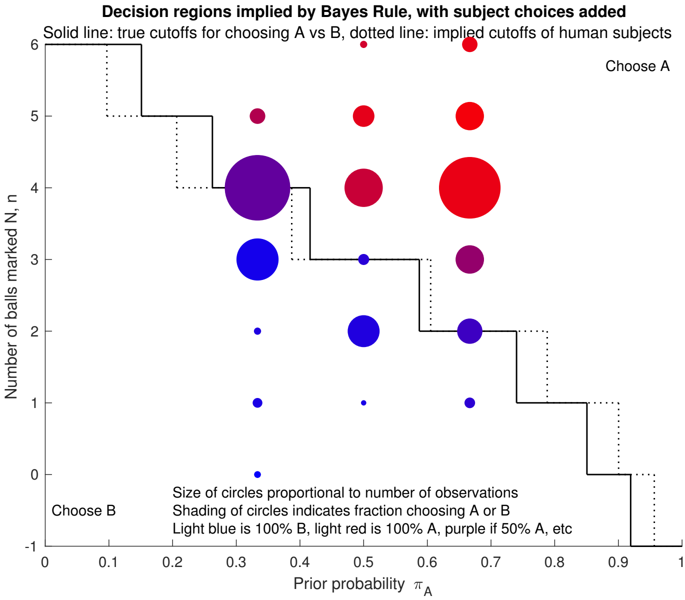
- They started out estimating a single type $k = 1$ model and found that $\hat{c} = (4, 3, 2)$. That is, subjects are making choices consistent with Bayes Rule!
- However the estimated error rate is large: $\hat{\varepsilon} = .38$.
- Error rates were lower in experiments where subjects were paid (.3 vs .45), and errors for UCLA subjects were lower than the other schools (PCC,Occidental,CSULA).
- If $k = 2$ types are allowed, EC finds that Bayes Rule $(4, 3, 2)$ is the most frequent, but "Representativeness" $(3, 3, 3)$ is the next most frequently used cutoff rule (63% vs 37%).
- If $k = 3$ types are allowed, EC finds that the 3rd most common cutoff rule is "Conservatism" $(5, 3, 1)$.
- Using likelihood ratio tests, they strongly reject the hypothesis that "subjects at different schools act in similar ways and subjects across different payment schemes act in similar ways."

El-Gamal and Grether's conclusion

> *"Hence, even though the answer to 'are experimental subjects Bayesian?' is 'no,' the answer to 'what is the most likely rule that people use?' is 'Bayes's rule.' The second most prominent rule that people use is 'representativeness,' which simply means that they ignore the prior induced by the experimenter and make a decision based solely on the likelihood ratio. The third most prominent rule that our algorithm selects on the basis of the data is "conservatism," which means that subjects give too much weight to the prior induced by the experimenter, needing more evidence to change their priors than Bayes Rule would imply."*

- Hereafter for brevity we will refer to El-Gamal and Grether by their initials, EGG.

Summary of subjects choices in EGG's study



Evidence in favor of cutoff rules

- The graph suggests that interpreting subjects as using "cutoff rules" is a reasonable way to view the experimental outcomes.
- We can view the figure as a type of *confusion matrix* that indicates where subjects make most of their mistakes (classification errors) in the experiment: *close to the cutoff line separating the "choose A" region from the "choose B" region.*
- The size of the dots conveys sample size, the color conveys the "confusion" i.e. the classification error rates: deep blue colors imply that subjects mostly chose cage B, deep red indicates that subjects mostly chose cage A. Purple colors reveal the higher rate of classification errors that occur for experiments near the boundary of the two regions, i.e. near the border where the true posterior probability for cage A, $\Pi(A|n, \pi_A) = 1/2$.
- Thus: the experiment suggests that the subjects made the most errors at the boundary between the two regions, which is precisely where we would expect them to make the most errors because the "evidence" (n, π_A) is not strong for choosing A or B.

Critique of the EGG model

- Why should a person “randomly guess” especially when the evidence makes them pretty certain that the cage from which the sample was drawn was either cage A or cage B?
- We would expect that a person would be most likely to “guess” when their *subjective posterior* is close to 1/2, i.e. when the evidence does not clearly favor cage A or cage B.
- Estimating cutoff rules does not reveal *underlying beliefs* of the subjects, i.e. their *subjective posterior belief* about the probability that the observed sample was drawn from cage A.
- There is no way that the model reflects the extra incentives for the Pay subjects (who received a bonus for choosing the correct cage) except indirectly via reduced error rates.
- Can we model the subjects a different way, and infer their *subjective posterior probabilities* using data only on their *observed (binary) choices?*
- Could we use insights from *deep learning* and *machine learning* to model how people might be making their choices?

Other ways of modeling subjects in the EGG experiments

- EGG used a *structural model* of subjects' choices that was derived from a coherent *theory* of how subjects made choices, allowing for "mistakes" and subject heterogeneity that provides a very good fit to the data. But are there other ways of modeling subject behavior?
- *Reduced-form models* are models that attempt only to *predict* what subjects do, without attempting to go deeper to try to *explain* what they do and why they do it. These approaches focus on estimation of the *Conditional Choice Probability* (CCP). With enough CCPs can be estimated by a number of different non-parametric methods.
- Deep neural networks (DNNs) can be viewed as a class of "flexible functional forms" for doing non-linear regression that can approximate unknown functions such as CCPs arbitrarily well and thus can be viewed as another type of non-parametric estimation. The focus of Deep Learning is *prediction* not so much *explanation.* It is useful to have a non-parametric "baseline" to compare structural models to. When data are *sparse* (i.e. there are many "zero cells") DNNs have attractive properties for *smoothing the data* without imposing assumptions on the data that may not be justified.

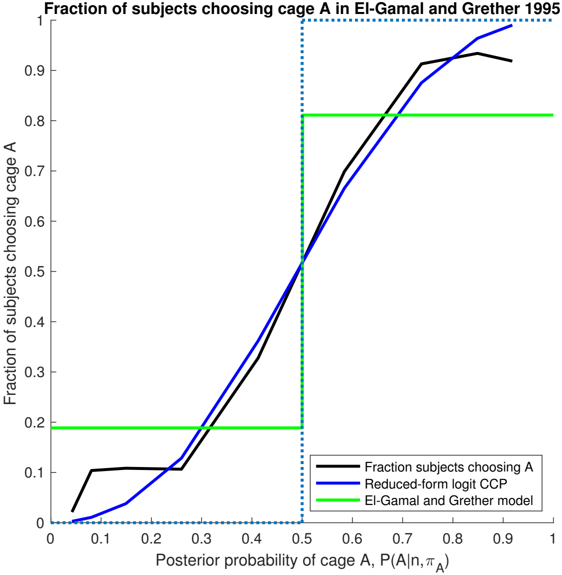
The logit model: the simplest possible neural net

- A natural alternative but *reduced-form model* of subject choices is the *binary logit model* which is a 3 parameter model of the probability that a subject chooses cage A given by

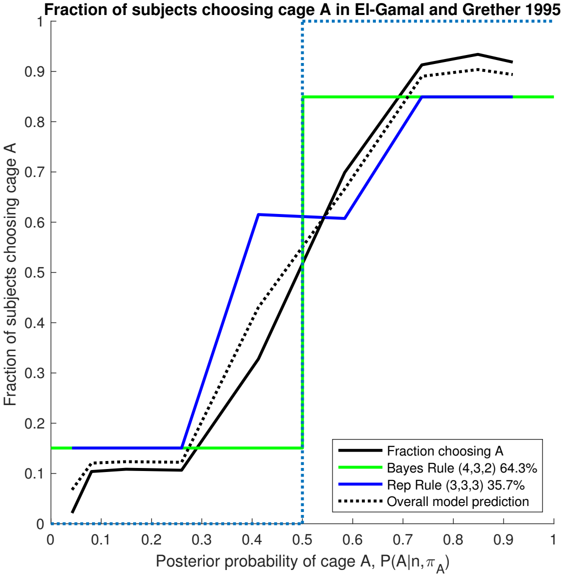
$$P(A|n, \pi_A) = \frac{1}{1 + \exp\{\beta_0 + \beta_1 n + \beta_2 \pi_A\}}. \tag{5}$$

- The logit model can be regarded as *flexible functional form for approximating the probability of choosing cage A*. But it doesn't have a direct interpretation as a structural model or theory of how people learn. But it is the simplest example of a *single layer feedforward neural network with 2 inputs, 1 output, and the "softmax" or logistic "squashing function."*
- The logit/NN model fits the data significantly better than the EGG (single type) model: the log-likelihood for the EGG model (4 parameters) is −1942.63 whereas the logit/NN model results in a log-likelihood of −1811.31. A 2 type EGG model (7 parameters) fit by EC fits even better: log-likelihood −1699.39. A 2 type logit/NN model (6 parameters) also fit by EC fits even better, as is evident from comparing predicted probabilities to non-parametric CCP.

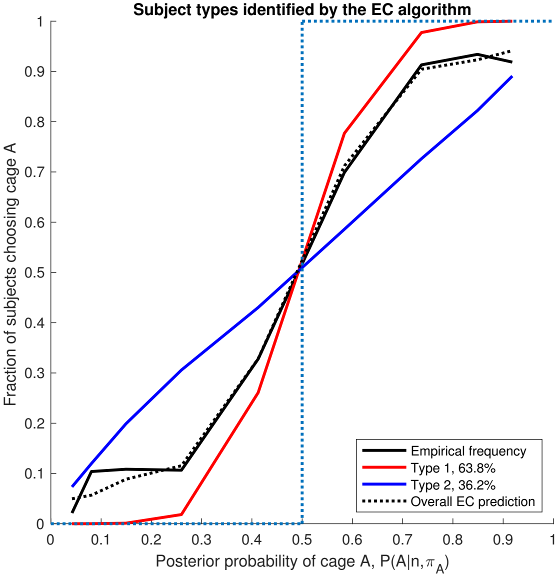
The data vs El-Gamal/Grether and binary logit models



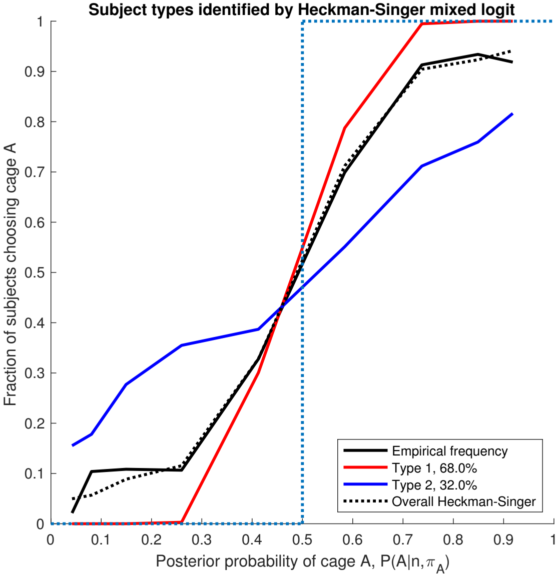
A 2nd type really improves fit of the EGG model . . .



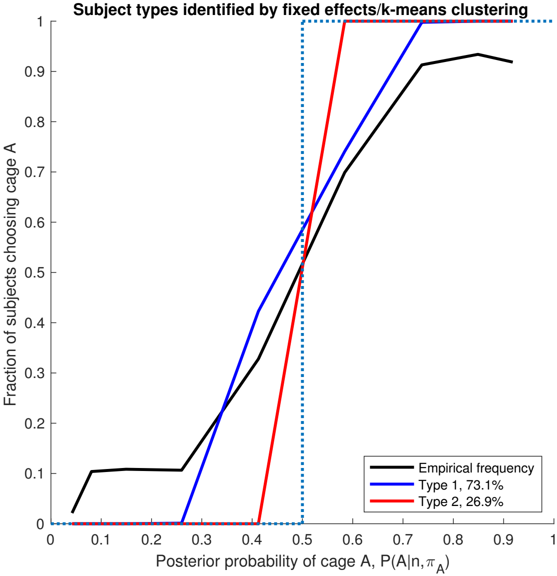
but a 2 type logit/NN model does even better



2 type model: Heckman-Singer random effects



2 type model: fixed effects with k-means clustering



Why should a rational person use Bayes Rule?

- First realize that in *IID* draws, *order does not count* so simply the number of balls N, denoted by n, is all that matters to make a decision (along with the knowledge of the prior probability π_A).
- A rational person will use an *optimal decision rule* for choosing A or B, one that maximizes their expected payoff from the choice.
- Mathematically, this can be represented by a function, $\delta(n, \pi_A) \to \{A, B\}$.
- What is the optimal decision rule? We can prove that it is optimal to behave according to *Bayes Rule* and compute the *posterior probability* $\Pi(A|n, \pi_A)$ that A is the cage used to draw the sample given the information $I = (n, \pi_A)$.

Theorem *The optimal decision rule is given by*

$$\delta(n, \pi_A) = \begin{cases} A & \text{if } \Pi(A|n, \pi_A) \geq 1/2 \\ B & \text{otherwise} \end{cases} \tag{6}$$

What is Bayes Rule?

Bayes Rule is a model of *rational learning* where based on information received a decision maker changes their *prior belief* to a *posterior belief* according to the formula

$$\Pi(A|n, \pi_A) = \frac{f(n|A)\pi_A}{f(n|A)\pi_A + f(n|B)\pi_B} \tag{7}$$

where $f(n|A)$ is the *likelihood* (or probability) of observing n balls marked N if the true cage used to draw the sample is A, and $f(n|B)$ is the likelihood if the true cage is B.

- In this case, $f(n|A)$ and $f(n|B)$, are *binomial probabilities* given by

$$\begin{aligned} f(n|A) &= \binom{6}{n} p_A^n (1-p_A)^{6-n} \\ f(n|B) &= \binom{6}{n} p_B^n (1-p_B)^{6-n} \end{aligned}$$

- where $p_A = 2/3$ and $p_B = 1/2$ are the probabilities of drawing an N in cages A and B, respectively.

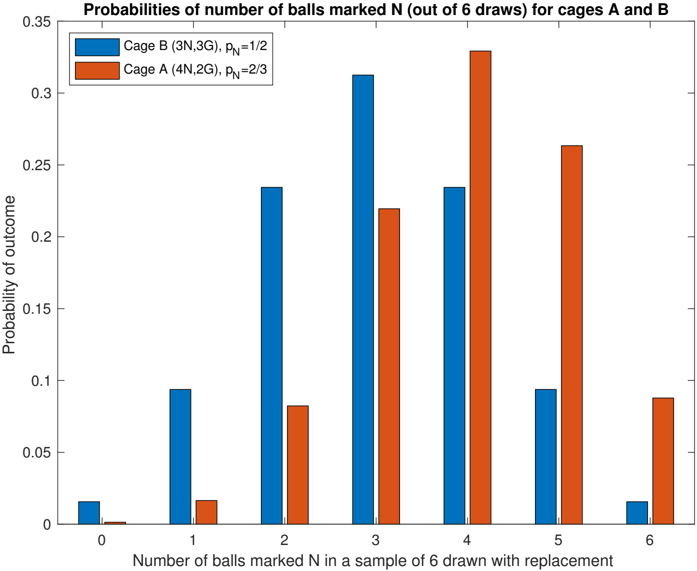
Likelihood and prior odds ratios

- Calculating the posterior probability for $n = 3$ when $\pi_A = 2/3$ we get $\Pi(A|3, 2/3) = .58414$. So a rational person chooses cage A.
- We can rearrange the formula for $\Pi(A|n, \pi_A) \geq 1/2$ equivalently as

$$\left(\frac{f(n|B)}{f(n|A)}\right)\left(\frac{1-\pi_A}{\pi_A}\right) \leq 1. \tag{8}$$

- So our decision depends both on the outcome n via the likelihood ratio $f(n|B)/f(n|A)$ (expressing how likely B is over A given n alone) and the prior odds ratio $(1-\pi_A)/\pi_A$ (expressing how much more likely B is over A *a priori*).
- In this case, $f(3|B)/f(3|A) = 1.4238$, so based on n alone, we find cage B more likely. But the prior odds ratio is $(1-\pi_A)/\pi_A = 1/2$ so the prior odds tell us that cage A is more likely to have been chosen for the draws.
- Thus we get $[f(3|B)/f(3|A)][(1-\pi_A)/\pi_A] = 1.4238 \times 1/2 = .7119$. Since this is less than 1, the combined data and prior info favor the choice of cage A over B. In essence, the prior information that $\pi_A = 2/3$ is stronger than the evidence that $n = 3$.

Binomial densities, $f(n|A)$ and $f(n|B)$



A Model of a "Subjective Bayesian" decision maker

- Let $\Pi_s(A|n, \pi_A)$ represent the *subjective posterior probability* that cage A was used to generate the sample given the information (n, π_A). We parameterize it as follows

$$\Pi_s(A|n, \pi_A, \beta) = \frac{1}{1 + \exp\left\{\beta_0 + \beta_1 \log\left(\frac{f(n|A)}{f(n|B)}\right) + \beta_2 \log\left(\frac{\pi_A}{1-\pi_A}\right)\right\}}$$

- Note that $\Pi_s(A|n, \pi_A, \beta) = \Pi(A|n, \pi_A)$ when $\beta = (0, -1, -1)$, so this model "nests the true Bayes posterior" as a special case, but when $\beta \neq (0, -1, -1)$ the subjective posterior differs from the true posterior.
- Given there is a payoff of \$10 for choosing the right cage we use a *multinomial logit model* to reflect the choice given the subjective beliefs. Thus the subject chooses cage A if

$$10\Pi_s(A|n, \pi_A, \beta) + \sigma\epsilon(A) \geq 10\Pi_s(B|n, \pi_A, \beta) + \sigma\epsilon(B) \qquad (9)$$

A Model of a "Subjective Bayesian" decision maker

- The shocks $(\epsilon(A), \epsilon(B))$ reflect random factors that the experimenter cannot observe that affect a subject's choice, including "calculational errors"
- If the shocks have a Type-1 extreme value distribution, my thesis adviser Daniel McFadden showed in the 1970s that the probability of choosing cage A will be given by the *binomial logit formula*

$$P(A|n, \pi_A, \theta) = \frac{\exp\{10\Pi_s(A|n, \pi_A, \beta)/\sigma\}}{\exp\{10\Pi_s(A|n, \pi_A, \beta)/\sigma\} + \exp\{10[1 - \Pi_s(A|n, \pi_A, \beta)]/\sigma\}}$$

where $\theta = (\sigma, \beta)$ are the unknown parameters that capture the subject's behavior. We can rewrite this as

$$P(A|n, \pi_A, \theta) = \frac{1}{1 + \exp\left\{20\left[\frac{1}{2} - \Pi_s(A|n, \pi_A, \beta)\right]/\sigma\right\}}.$$

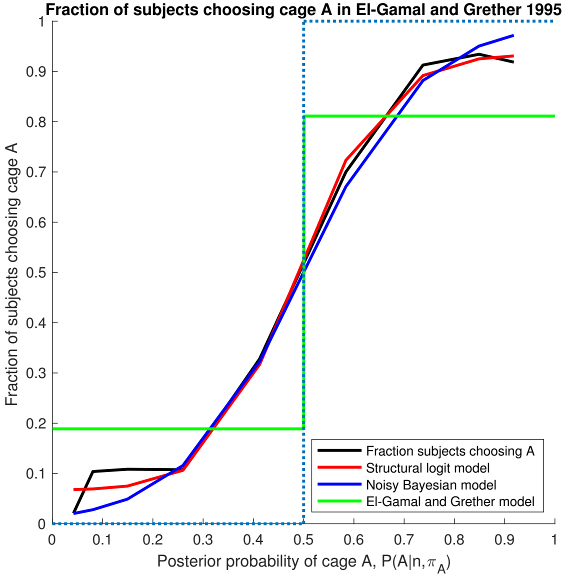
Interpretation of the structural logit model

- Notice that the structural model enables us to *infer subjective posterior probabilities* that we hypothesize subjects use when making their choice of cage A or B.
- The CCP at the 2nd "layer" of the model then compares the subjective posterior $\Pi_s(A|n, \pi_A, \beta)$ to 1/2. When $\Pi_s(A|n, \pi_A, \beta) = 1/2$ the subject is indifferent and has a 50% chance of choosing cage B.
- The subject is more likely to choose cage A when $\Pi_s(A|n, \pi_A)$ exceeds 1/2, and more likely to choose cage B when $\Pi_s(A|n, \pi_A)$ is less than 1/2. This matches the pattern of classification errors we observe in the data but is not what the EGG predicts.
- The parameter σ scales the "noise" that affects the subjects' choices beyond just evaluating the subjective posterior $\Pi_s(A|n, \pi_A)$ and comparing it to 1/2. It is comparable to the "error rate" parameter ε in the EGG except that the effect of errors in the structural logit is biggest when $\Pi_s(A|n, \pi_A)$ is close to 1/2 and lowest when it is close to 1 or 0, just as what we observe in the data.

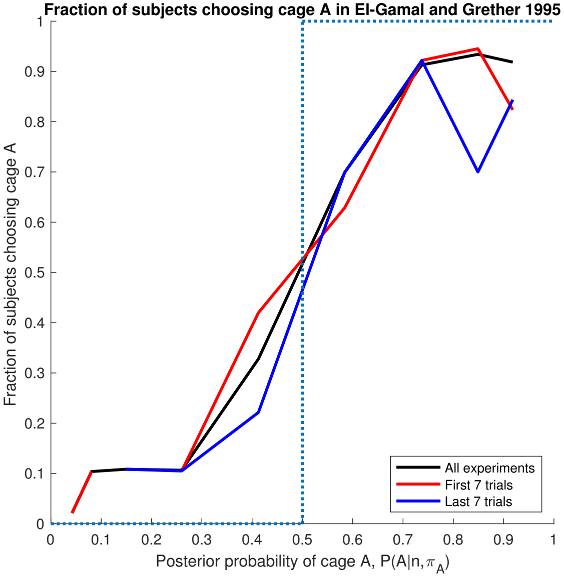
The "Noisy Bayesian" model

- Note also that increasing the \$10 reward for a correct classification works comparably to scaling *down* the σ "noise parameter", explaining the finding of lower classification error rates for subjects who were paid compared to those who weren't.
- We might interpret the extreme value shocks as a sort of random "calculational error" by subjects, which may be endogenous, and a result of *mental effort* (*apropos* Daniel Kahneman's book *Thinking: Fast and Slow*). It might reflect the fact that subjects exert more mental effort (and hence make fewer calculational errors) when incentives for making a correct decision is higher. But the effect just follows mechanically, for fixed σ because the *signal to noise ratio in the choices increases when payoff to a correct decision increases.*
- Note that the subjective posterior $\Pi_s(A|n, \pi_A, \beta)$ includes the true Bayesian posterior when $\beta = (0, -1, -1)$. We can evaluate this special case of the model as a *noisy Bayesian decision maker.* However via a likelihood-ratio test, we can strongly reject the hypothesis that subjects in the EGG experiments are "noisy Bayesians" (P-value 9.2×10^{-12}).

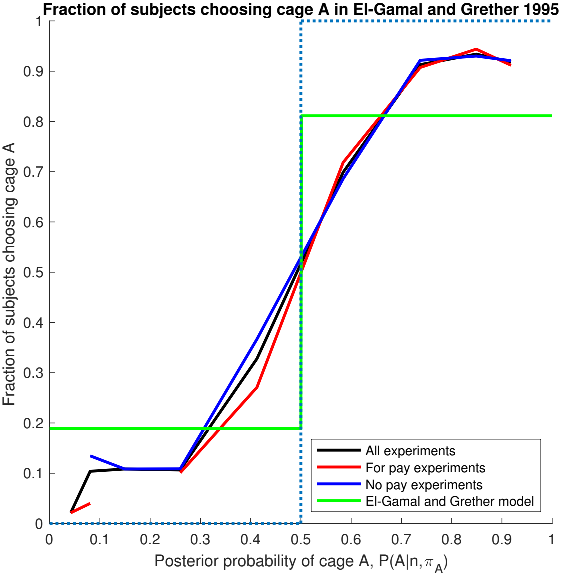
“Noisy Bayesian” vs “Subjective Bayesian” models



Experience/learning effects in the experiment



Incentive effects in the experiment



The structural model is a 2 layer NN!

The structural model is a "logit inside a logit".

1. A "layer 1" logit for the subjective posterior, $\Pi_s(A|n, \pi_A, \beta)$ that depends on the three parameters $(\beta_0, \beta_1, \beta_2)$.
2. A "layer 2" logit that determines the subject's response on whether to choose cage A or cage B by comparing the subjective posterior to 1/2 allowing for "calculational noise" (or alternatively, choosing the cage with the highest expected payoff, also reflecting calculational noise or other distractions). Thus the "output" of this two layer NN is the CCP,

   $$P(A|n, \pi_A, \sigma, \beta) = P(A|\Pi_s(A|n, \pi_A, \beta), \sigma) \tag{10}$$

   the result of using the layer 1 logit as the "input" to the layer 2 logit.

Thus the structural model can be regarded as a *2 layer feed-forward neural network model with two inputs,* (n, π_A).

A general 2 layer binary NN classifier

- We can generalize this model by adding a "bias term" to the 2nd output layer:

$$P(A|n,\pi_A) = \phi\left(\alpha_0 + \alpha_1\phi(\beta_0 + \beta_1 T(n) + \beta_2 T(\pi_A))\right), \qquad (11)$$

  where $\phi(x) = 1/(1+\exp(-x))$ is the logistic activation or "squashing function" and $T(n)$ and $T(\pi)$ are the "transformed inputs" given by $T(n) = \log(f(n|A)/f(n|B))$ (log-likelihood ratio) and $T(\pi_A) = \log(\pi_A/(1-\pi_A))$ (log-prior odds ratio).
- This NN is *shallow and thin* since it has only 2 layers (i.e. a "depth" of 2) and 1 hidden unit in each layer (i.e. a "width" of 1). It depends on only 5 parameters $(\alpha_0,\alpha_1,\beta_0,\beta_1,\beta_2)$ where (α_0,β_0) are the "bias terms" and $(\alpha_1,\beta_1,\beta_2)$ are the weights.
- Notice that the structural logit model is a special case of the two layer binary NN classifier (11) when

$$\alpha_0 = \frac{-1}{2\sigma/U},\ \beta_0 = 0,\ \alpha_1 = \frac{1}{\sigma/U},\ \beta_1 = -1,\ \beta_2 = 1 \qquad (12)$$

  where $U = 10$ is the payoff to a correct classification. The structural model depends on only 4 parameters since it restricts $\alpha_0 = -2\alpha_1$.

The value of transforming the inputs

- Why "pre-transform" the "inputs" (n, π_A) using the transformations $T(n)$ and $T(\pi_A)$?
- **Answer:** because theory suggests it: the output of this 2 layer feedforward NN includes the true Bayes Rule classifier $P(A|n, \pi_A) = I\{\Pi(A|n, \pi_A) \geq 1/2\}$ as a limiting special case of the NN classifer when β_0, $\beta_1 = \beta_2 = 1$ and $\alpha_0 = -2\alpha_1$ and $\alpha_1 \to \infty$.
- But as $(\alpha_0, \alpha_1, \beta_0, \beta_1, \beta_2)$ vary over R^5 the NN classifier can approximate many *non-Bayesian* decision rules as well. Thus, the NN can be viewed as a *flexible functional form* or *reduced form model* of the CCP.
- If we don't transform the inputs we get a standard 2 layer feedforward NN

$$P(A|n, \pi_A) = \phi\left(\alpha_0 + \alpha_1 \phi(\beta_0 + \beta_1 n + \beta_2 \pi_A)\right), \tag{13}$$

You can show that the model (13) no longer nests the true Bayesian decision rule as a limiting special case.

The value of adding more layers

- Suppose we assume that human subjects are incapable of quickly transforming the inputs $n \to T(n)$ before choosing a cage as the structural model presumes. Does this rule out their ability to behave according to Bayes Rule even in the limit for any choice of parameters $(\alpha_0, \alpha_1, \beta_0, \beta_1, \beta_2) \in R^5$?
- Yes, but we can solve the problem by *adding extra layers to the neural network.* The idea of the additional layers is to use them to *approximate the transformations*

$$T(n) = \log(f(n|A)/f(n|B)) \quad \text{and} \quad T(\pi_A) = \log(\pi_A/(1-\pi_A)).$$

- We can also improve the approximation by making the neural network *wider* by adding more hidden units at each layer.
- What is the better way to go: use a deeper by narrower NN, or a wider and shallower NN?
- This is a question of *how to design the NN architecture.*

A brief detour on the mathematics of NNs

- NNs are fundamentally a way of *approximating functions*
- It has long been known that even single layer NNs are *universal approximators* of continuous multivariate functions
- Let $f : R^d \to R$ be a continuous function. Approximate $f(x)$ by $g(x)$ given by

$$g(x_1, \dots, x_d) = \sum_{i=1}^{h} w_i^2 \phi(b_i + w_i^1 x_d). \tag{14}$$

  where $\phi : R \to R$ is a continuous activation or squashing function.
- Note that the approximation g in equation (14) is a single layer feed-foward neural network with h hidden units that depends on a total of $(d+2) * h$ parameters:
  1. $h \times d$ input weights $w^1 = \{w_{ij}^1 | i = 1, \dots, h, j = 1, \dots, d\}$
  2. h bias terms $b = (b_1, \dots, b_h)$
  3. h output weights $w^2 = \{w_i^2 | i = 1, \dots, h\}$.

Universal Approximation using single layer NNs

Universal Approximation Theorem (Hornik, Cybenko, White, etc.)

If the target function $f : R^d \to R^m$ is continuous and activation function $\phi : R \to R$ is continuous but not a polynomial, then for any $\varepsilon > 0$ and any compact subset $K \subset R^d$, there exist $m(d+2)l$ parameters (b, w^1, w^2) such that

$$\sup_{x \in K} |f(x) - g(x)| \leq \varepsilon. \tag{15}$$

- Thus, even a single layer NN can approximate any continuous function arbitrarily closely as long as it is sufficiently wide (i.e. has sufficiently many hidden units, h).
- The ability of NNs to approximate arbitrary continuous functions was known before this work in the 1980s. Some of the earliest results are *exact representation theorems* for multivariate continuous functions first proved by Kolmogorov in the 1950s.

Kolmogorov Representation Theorem

Kolmogorov Superposition Theorem (1957)

Any multivariate continuous function can be represented as a superposition of one- dimensional functions, i.e., there exists continuous univariate functions $\{\sigma_1, \ldots, \sigma_{2n}\}$ *where* $\sigma_i : R \to R$ *and* $2n^2$ *continuous functions* $\sigma_{ij} : R \to R$ *such that*

$$f(x_1, \ldots, x_d) = \sum_{i=1}^{2n} \sigma_i \left(\sum_{j=1}^{n} \sigma_{ij}(x_j) \right) \tag{16}$$

- Kolmogorov (at age 27) solved *Hilbert's 13th problem.*
- This representation resembles a single layer feedforward neural network, except that each "hidden unit" σ_{ij} only depends on a single "input" x_j rather than a linear combination of all d inputs $(x_1, \ldots, x_d)$ with bias terms.

Lorentz *et. al.* Representation Theorem

Lorentz *et. al.* Representation Theorem (1996)

For any function $f \in C([0,1]^d) \to R$ there exists a continuous univariate function $\sigma : R \to R$ and $2d+1$ continuous and strictly increasing functions $\sigma_i : R \to R$, and d positive constants $\lambda_d > 0$ satisfying $\sum_{j=1}^{d} \lambda_j \leq 1$ such that:

$$f(x_1, \ldots, x_d) = \sum_{i=1}^{2d+1} \sigma \left(\sum_{j=1}^{d} \lambda_j \sigma_i(x_j) \right), \tag{17}$$

where σ depends on f.

- Notice the Kolmogorov and Lorentz *et. al.* are not approximation results, but rather *exact representations of continuous functions* unlike infinite Taylor series and polynomial expansions that generally require infinitely many terms to represent any function.

The Curse of Dimensionality for function approximation

- In the literature on analytic computational complexity, there are results that prove that the set of continuous functions of d variables are subject to a *Curse of Dimensionality.*
- That is, the number of computer operations that would be required to find a function g that provides a uniform approximation to tolerance ε to any function f in some appropriate class $\mathcal{F}$ (e.g. all Lipschitz continuous or continuously differentiable fuctions from R^d to R) is subject to a Curse of Dimensionality.
- Let $N(\varepsilon, d, f)$ denote the total number of computer operations (function evaluations of f) to produce a function g satisfying $\|f - g\| \leq \varepsilon$, then there exists a constant $C > 0$ such that

$$\sup_{f \in \mathcal{F}} N(\varepsilon, d, f) \geq C \left(\frac{1}{\varepsilon}\right)^d . \tag{18}$$

- For example, if $\varepsilon = .001$ and $d = 10$ then the worst case lower bound on $N(\varepsilon, d, f)$ for any $f \in \mathcal{F}$ is $[10^3]^{10} = 10^{30}$, i.e. astronomically large. Thus, it is computationally infeasible to closely approximate functions of 10 variables, at least in the worst case.

Barron's Bound

Barron's Theorem (1993)

Let $\mathcal{F}$ be the class of functions $f : R^d \to R$ that have finite-mean Fourier representations, i.e.

$$f(x) = \int_{\omega \in R^d} \exp\{i\langle x, \omega\rangle\}\tilde{f}(\omega) d\omega \tag{19}$$

for some complex-valued function $\tilde{f}(\omega)$ with finite 2nd moment

$$C_f = \int_{\omega \in R^d} \sqrt{\langle \omega, \omega\rangle}\tilde{f}(\omega) d\omega < \infty. \tag{20}$$

Given a probability measure μ over $B_r = \{x \in R^d | \|x\| \leq r\}$ there exists a single layer feedfoward NN with h hidden units that satisfies the error bound

$$\int_{x \in B_r} (f(x) - g(x))^2 \mu(dx) \leq \frac{(2rC_f)^2}{h}. \tag{21}$$

Do NNs break the Curse of Dimensionality?

- Barron's bound show that to approximate functions of d variables to within mean squared error of ε we only need $h = (2Cr)^2/\varepsilon$ hidden units and thus a total of only $(d+2)h$ parameters.
- Compare this to tensor-product polynomial approximations of functions (e.g. using tensor products of univariate Chebyshev polynomials) that will have k^d total parameters, where k is the number of Chebyshev polynomials used for each variable, x_i, $\{\psi_1(x_i), \dots, \psi_k(x_i)\}$, $i = \{1, \dots, d\}$.
- To guarantee a uniform ε approximation, k must increase at rate $1/\varepsilon$, so tensor product polynomial approximation methods (or related spline, B-spline and other methods) will require a total number of parameters that explodes exponentially with d, i.e. these methods suffer from the Curse of Dimensionality.
- However without further restrictions, the answer is NO: NNs are just another method of function approximation and the general complexity lower bound for function approximation applies to *any* algorithm or approximation method.

Why can't NNs break the Curse of Dimensionality?

- While the number of parameters $(d+2)h$ in a NN approximation only grows linearly in $1/\varepsilon$ (independent of d) the total computational effort to find network parameters that *globally minimize* mean squared approximation error rises exponentially fast in h due to Theorems on complexity of function minimization by Nemirov and Yudin (1976).
- Barron's bounds are will hold only if an optimization method finds a *global minimum* but guaranteeing this in the worst case for nonlinear functions without "special structure" is also subject to the Curse of Dimensionality.
- Barron's assumptions on 2nd moment Fourier integrable function class $\mathcal{F}$ also result in "special structure" that his bound exploits.
- For more general classes of functions, such as $\mathcal{F} = C_s([0,1]^d)$, i.e. the class of s times continuously differentiable functions from $[0,1]^d \to R$, Lu, Shen, Yang and Zhang (2021) proved the following lower bound on the worst case approximation error for a DNN with H hidden units per layer and a depth of L

Lower Bound on DNN Approximation Error

Lu, Shen, Yang and Zhang lower convergence rate bound for DNNs

Consider a DNN with L layers and H hidden units per layer, i.e. a width of H and depth of L. Let $\theta(H, L)$ denote the parameters of a fully connected feedforward DNN of width H and depth L. Then we have

$$\sup_{f \in \mathcal{F}} \min_{\theta(H,L)} \|f - g(\theta(H, L), H, L)\| \geq C \left(H^2 L^2 \log(H)^3 \log(L)^3\right)^{-s/d}, \quad (22)$$

for some positive constant C.

- This lower bound shows that the convergence rate slows with dimension d similar to what we see in non-parametric regression, both in width and depth of the NN.
- Thus, to guarantee a uniform ε-approximation for any $f \in \mathcal{F}$, we need to have both L and H grow exponentially fast with d. In other words, DNNs cannot break the Curse of Dimensionality in the worst case.
- **END OF MATHEMATICAL DETOUR**

Success in using DNNs: more an "art" than a "science"?

- Theory can only tell us so much. A huge amount of how DNNs really work and the pros and cons of different methods of function approximation work is in the domain of *practical, hands-on experience* (otherwise known as *learning by doing*)
- See Adcock and Dexter (2021) "The Gap between Theory and Practice in Function Approximation with Deep Neural Networks"

  "despite the impressive empirical and theoretical results achieved in the broader DL community, there is concern that methods based on DNNs do not currently meet the usual rigorous standards for algorithms in computational science … [complexity theory] says little about its practical performance when trained by modern approaches in DL. If such techniques are to achieve widespread adoption in scientific computing, it is vital they be understood through the lens of numerical analysis, namely, (i) stability, (ii) accuracy, (iii) sample complexity, (iv) the curse of dimensionality, and (v) computational cost."

What is compressed sensing?

- In a nutshell, it is a linear approximation scheme using tensor products of univariate orthonormal polynomials for each component x_i of the function $f(x_1, \dots, x_d)$ to be approximated, using Hilbert space theoryto write

  $$f(x) = \sum_{\nu} \beta_\nu \Psi_\nu(x) \tag{23}$$

  where ν is a *multi-index*, $\nu = (\nu_1, \dots, \nu_d)$ where each ν_i is a non-negative integer, and $\beta_\nu = \int f(x) \Psi_v(x) \rho(dx)$ is the OLS coefficient of the projection of f onto the orthonormal basis $\{\Psi_\nu(x)\}$.
- The tensor products of univariate orthogonal polynomials $\Psi_\nu(x)$ is defined by

  $$\Psi_\nu(x) = \prod_{i=1}^{d} \psi_{\nu_i}(x_i), \tag{24}$$

  where $\psi_{\nu_i}(x_i)$ is a univariate orthogonal polynomial (e.g. Legendre or Chebyshev polynomial).

Function approximation as a regression problem

- As mentioned above there is an exponential blowup in the number of tensor products as d increases. For example if we restricted the maximum degree ν_i to 3, there would be 4^d possible tensor products $\Psi_\nu(x)$ as the multi-indices $\nu = (\nu_1, \ldots, \nu_d)$ each run from 0 to 3.
- Suppose we can evaluate f at N points, so f is the $N \times 1$ vector $f = (f(x_1), \ldots, f(x_N))'$. Next we choose J basis functions evaluated at the same x_i points and let the $N \times J$ matrix A have (i, j) element $\{\Psi_j(x_i)\}$. Then we seek to approximately solve the large linear system $A\beta = f$ for the coefficients $\beta = (\beta_1, \ldots, \beta_K)$ to obtain our approximation of the function f.
- Compressed sensing uses a Lasso or L_2 regularization to approximately solve the linear system $A\beta = f$ using LASSO

$$\hat{\beta} = \underset{\beta}{argmin} \|\beta\|_1 + \mu \|A\beta - f\|_2 \tag{25}$$

for an appropriate positive parameter $\mu > 0$.

Which is better: CS or DNN?

They compared the performance of DNNs of different widths and depths on a variety of test functions in various dimensions, including smooth and piecewise smooth functions to results using CS.

> *Our main conclusion from these experiments is that there is a crucial gap between the approximation theory of DNNs and their practical performance, with trained DNNs performing relatively poorly on functions for which there are strong approximation results (e.g., smooth functions) yet performing well in comparison to best-in-class methods for other functions.” But “with sufficiently careful architecture design and training, one may achieve superior performance with DNNs over CS. The hope is that, with these further efforts, DNNs may develop into effective tools for scientific computing that can consistently outperform current best-in-class approaches across a range of challenging problems.”*

Comments on model selection via cross validation

- Simon's lecture discussed how to choose a best fitting DNN architecture that balances the bias/variance tradeoff (and the dangers of "overfitting") by dividing the observations into 1) training, 2) testing, and 3) validation subsamples.
- This is a good way to go when you have "big data" but for many economic problems we don't have enough data.
- For example in the "Are People Bayesian?" study we only have 4010 subject/trial observations. DNNs can easily have far many more parameters than this.
- When we don't have the luxury of big data, it may be a good idea to use *model selection* such as using *Akaike Information Criterion* (AIC) or *Bayesian Information Criterion* (BIC) to select a "best model".
- The model selection approach includes a *penalty function for model complexity* and the goal is to appropriately balance the tradeoff between model fit and model complexity.
- In the case of maximum likelihood using DNNs or other flexibly parametrized semi-parametric methods, we have

$$AIC_k = 2k - 2L_N(\hat{\theta}) \quad \mathrm{BIC}_k = N\log(k) - L_N(\hat{\theta}), \tag{26}$$

AIC or BIC?

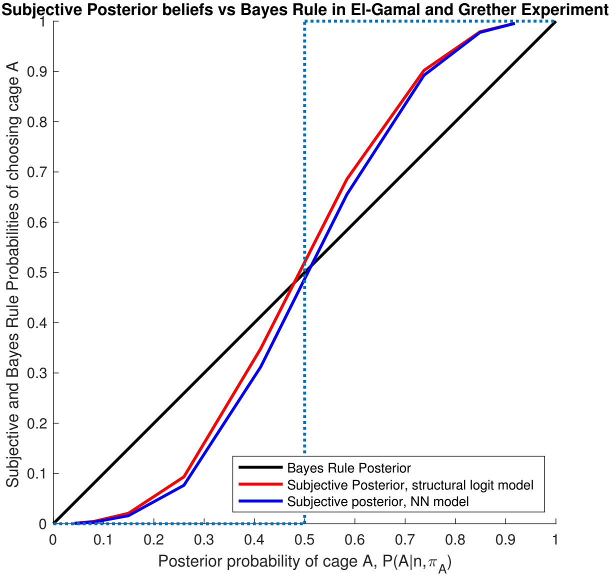
> *"A point made by several researchers is that AIC and BIC are appropriate for different tasks. In particular, BIC is argued to be appropriate for selecting the 'true model' (i.e. the process that generated the data) from the set of candidate models, whereas AIC is not appropriate. To be specific, if the 'true model' is in the set of candidates, then BIC will select the 'true model' with probability 1, as $N \to \infty$; in contrast, when selection is done via AIC, the probability can be less than 1. Proponents of AIC argue that this issue is negligible, because the 'true model' is virtually never in the candidate set. Indeed, it is a common aphorism in statistics that 'all models are wrong'; hence the 'true model' (i.e. reality) cannot be in the candidate set." (Wikipedia)*

AIC or BIC?

See also the review by Zhang, Yang and Ding (2023) "Information criteria for model selection" who conclude:

> *"AIC represents a group of information criteria that are efficient in the nonparametric scenario in the sense that the prediction performance of the selected model is asymptotically close to the best among the candidate models. However, they may fail to be consistent in choosing the most parsimonious well-specified model if it exists. It is because their penalties are too small to distinguish between two well-specified models and thus will lead to an over-selection of the model dimension, k."*

Subjective posterior beliefs reflect “overconfidence”



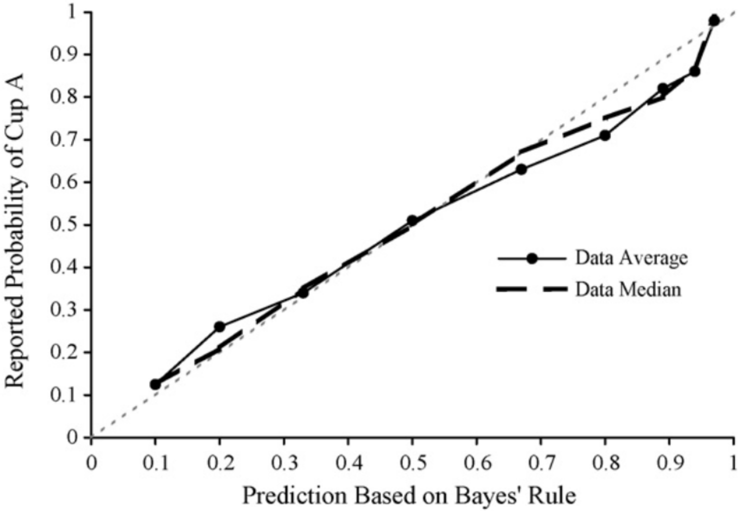
The "overconfidence effect" (from *Wikipedia*)

> *The overconfidence effect is a well-established bias in which a person's subjective confidence in their judgments is reliably greater than the objective accuracy of those judgments, especially when confidence is relatively high. Overconfidence is one example of a miscalibration of subjective probabilities. Throughout the research literature, overconfidence has been defined in three distinct ways:*

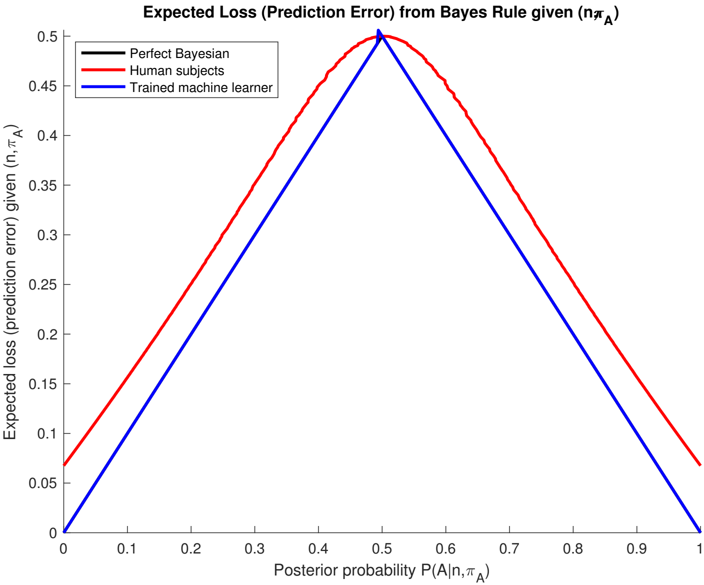
1. overestimation of one's actual performance
2. overplacement of one's performance relative to others
3. overprecision in expressing unwarranted certainty in the accuracy of one's beliefs.

> *The data show that confidence systematically exceeds accuracy, implying people are more sure that they are correct than they deserve to be. By contrast, the key finding is that confidence exceeds accuracy so long as the subject is answering hard questions about an unfamiliar topic.*

Compare to elicited posteriors (Holt and Smith, 2009)



Estimating the costs being irrational



So, are people Bayesian learners?

- Well, not exactly. Some people are more "Bayesian" than others, but almost nobody is a "perfect Bayesian decision maker"
- But I like this quote from a 1994 survey by Hutchinson and Meyer "Dynamic Decision Making: Optimal Policies and Actual Behavior in Sequential Choice Problems" in *Marketing Letters*

  "In summary, when compared to normative sequential choice models, humans are likely to perform less well because the processes of forward planning, learning, and evaluation have a number of inherent biases. From a broader perspective, however, one can argue that optimal solutions are known for a relatively small number of similar, well-specified problems whereas humans evolved to survive in a world filled with a large and diverse set of ill-specified problems. Our 'suboptimality' may be a small price to pay for the flexibility and adaptiveness of our intuitive decision proceess."