

Learning how to learn about how people (and machines) learn

Tianshi Mu, Pranjal Rowat, John Rust and Chengjun Zhang,
Georgetown University

April 23, 2024

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, statistical 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 instances of human learning probably 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 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 a “meta principle” of how we learn: start simple but choose well designed experiments and data gathering to test our hypotheses.
- Thus a good place to start is to see if human learning is consistent with Bayes Rule.

David Grether, 1938-2021



Daniel Kahneman, 1934-2024



Kahneman and Tversky: People are not Bayesian

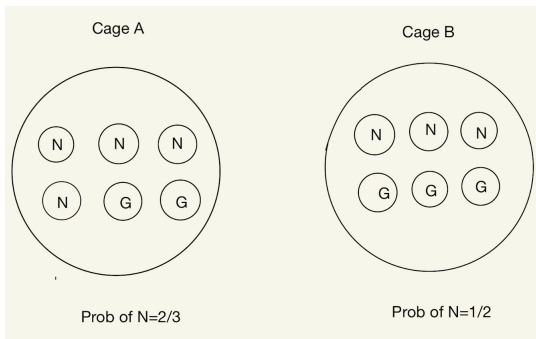
- “Judgment under Uncertainty: Heuristics and Biases” *Science* 1974.
- Reported an experiment where subjects were shown brief personality descriptions of several individuals sampled at random from 100 professionals, engineers and lawyers.
- Subjects were asked to report, for each description, the probability that the described person was an engineer.
- In the first experiment subjects were told that the pool contained 70 engineers and 30 lawyers. In the second, 30 engineers and 70 lawyers.
- “In a sharp violation of Bayes Rule, the subjects in the two conditions produced essentially the same probability judgements.”
- “The subjects used prior probabilities correctly when they had no other information.”

Grether's critique of Kahneman and Tversky

- “Recent Psychological Studies of Behavior under Uncertainty” *AER* 1978.
- Kahneman and Tversky's experiment “has features that make the applicability of the findings to economic decisions doubtful.”
- “the thumbnail descriptions were not a random sample from the alleged population”
- “there is the difficulty of controlling the information given when verbal descriptions or situations are presented.”
- “Both of these difficulties could be taken care of by the use of actual balls in urns or book-bag poker-chip set ups.”
- “Grether (1978) noted that the descriptions in this experiment were hypothetical and wondered whether the subjects truly believed the stated population proportions. He also noted that the nature of the monetary payoffs was unclear since subjects were told that they would be paid a bonus to the extent that their estimates were close to those submitted by an ‘expert panel.’” (Holt and Smith, 2009)

Are People Bayesian?

- Title of the classic 1995 article by El-Gamal and Grether in *Journal of American Statistical Association*
- Experiment with 257 subjects from 4 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

- In each trial a 6 sided die was thrown. If it landed $\{1, \dots, k\}$ cage A was selected, otherwise cage B was selected where $k \in \{2, 3, 4\}$.
- This results in a *credible* prior probability of cage A, $\pi_A \in \{1/3, 1/2, 2/3\}$.
- Subjects were not shown the outcome of the dice throw determining the cage used to draw 6 balls with replacement: they were only shown the results of the 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, A or B, 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 should people use Bayes Rule?

- Because we're *greedy*
- A rational person should 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, $\delta(n, \pi_A) \rightarrow \{A, B\}$ that maximizes the probability of selecting the correct cage from which the sample was drawn.
- An optimal decision rule can be expressed in terms of Bayes Rule

Theorem *The optimal decision rule is given by*

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

- However we will show that *you can behave like a Bayesian even if you don't believe like Bayesian*

El-Gamal and Grether's structural model of subjects' choices

- Assume subjects use *cutoff rules* to choose cage A or B
- **Cutoff Rule** An integer $c \in \{0, 1, \dots, 6\}$ 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 $c_i \in \{-1, 0, \dots, 6\}$.
- 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 $(\hat{c}^s, \hat{\varepsilon}^s) = \operatorname{argmax}_{c, \varepsilon} L(X_c^s | c, \varepsilon)$.
- 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 $(\hat{c}^1, \dots, \hat{c}^K, \hat{\varepsilon}^1, \dots, \hat{\varepsilon}^K)$ as follows

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

EC likelihood

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

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

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 $L(X_{c^k}^s | c^k, \varepsilon^k)$ 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 $L(X_{c^k}^s | c^k, \varepsilon^k)$ 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 $S \rightarrow \infty$ with $t/4$ observations per parameter, whereas it equals $St/4k$ for EC.

El-Gamal and Grether's empirical findings

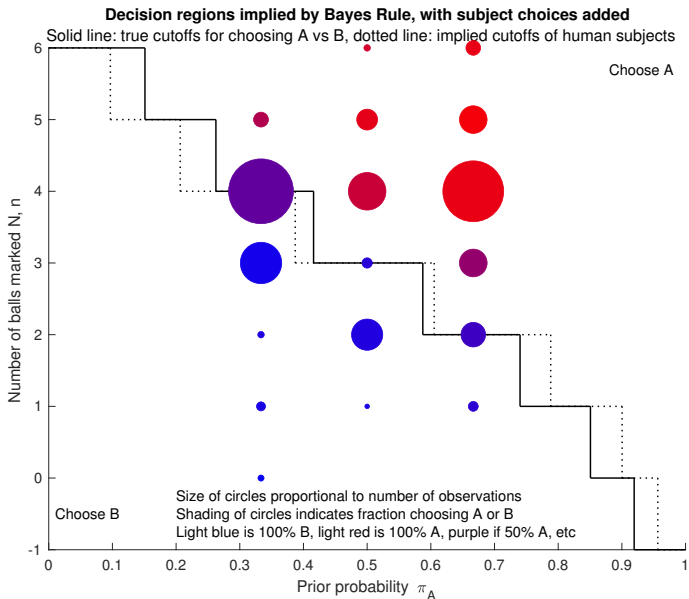
- They started out estimating a single type $k = 1$ model and found that $\hat{c} = (4, 3, 2)$. That is, subjects are making choices consistent with Bayes Rule!
- However the estimated error rate is large: $\hat{\varepsilon} = .38$.
- Error rates were lower in experiments where subjects were paid (.3 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

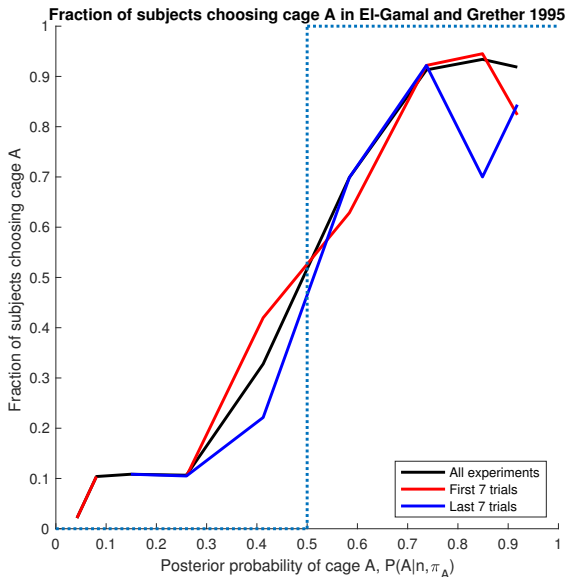
"Hence, even though the answer to 'are experimental subjects Bayesian?' is 'no,' the answer to 'what is the most likely rule that people use?' is 'Bayes's rule.' The second most prominent rule that people use is 'representativeness,' which simply means that they ignore the prior induced by the experimenter 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, EG.

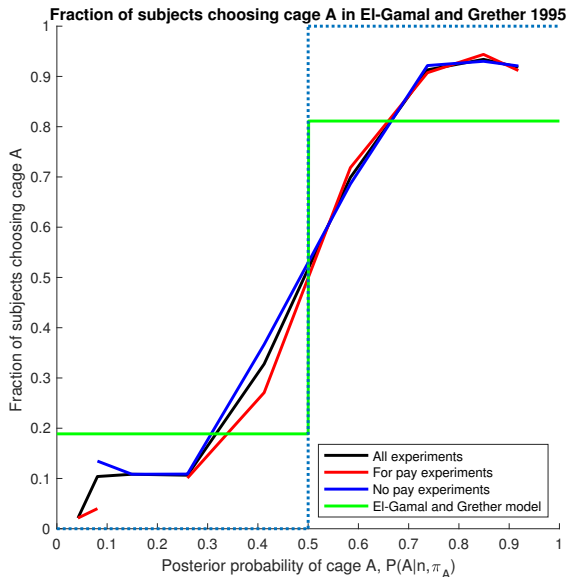
Summary of subjects' choices in EG's study



Experience/learning effects in the experiment



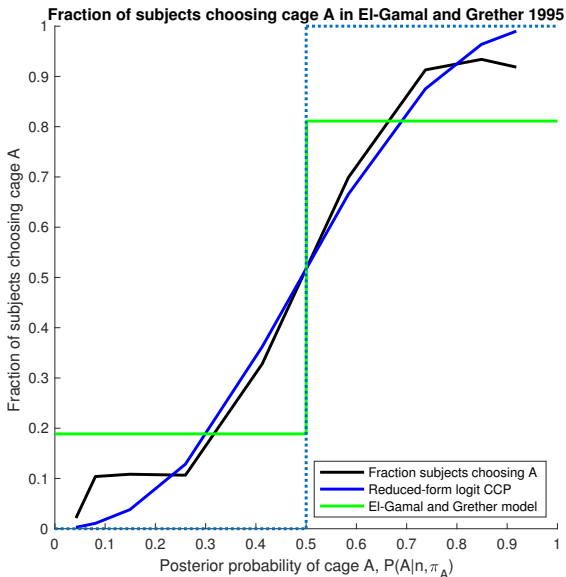
Incentive effects in the experiment



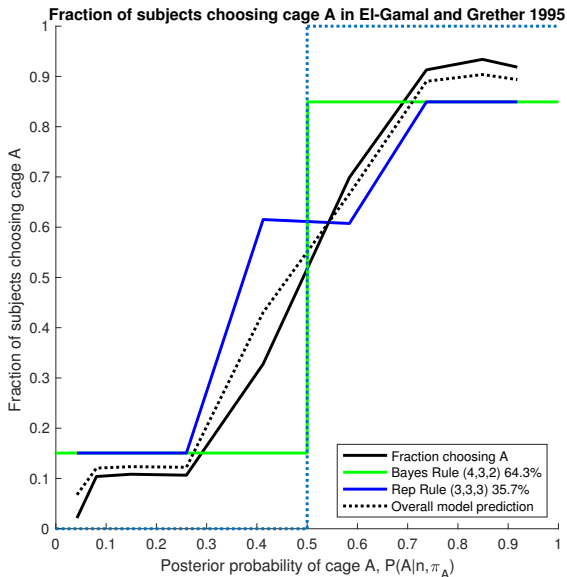
Evidence in favor of cutoff rules

- 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.

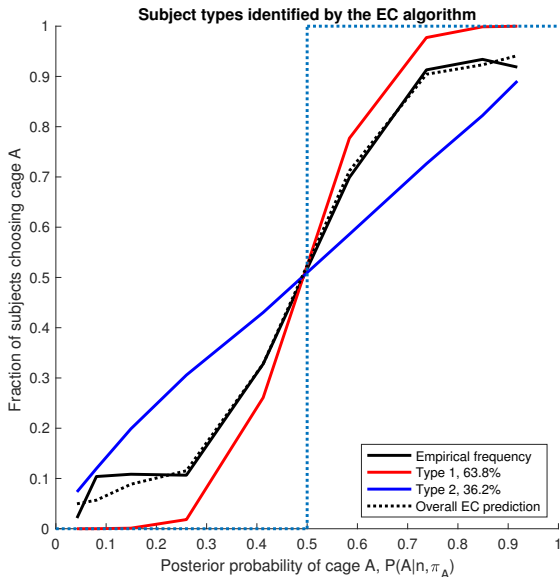
The data vs El-Gamal/Grether and binary logit models



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



But a 2 type logit model fits even better



Critique of the EG 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 EG experiments

- EG 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

$$P(A|n, \pi_A) = \frac{1}{1 + \exp\{\beta_0 + \beta_1 n + \beta_2 \pi_A\}}. \quad (3)$$

- 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 EG (single type) model: the log-likelihood for the EG model (4 parameters) is -1942.63 whereas the logit/NN model results in a log-likelihood of -1811.31 . A 2 type EG 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.

A Structural logit model of a “Subjective Bayesian”

- 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

$$10\Pi_s(A|n, \pi_A, \beta) + \sigma\epsilon(A) \geq 10\Pi_s(B|n, \pi_A, \beta) + \sigma\epsilon(B) \quad (4)$$

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

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

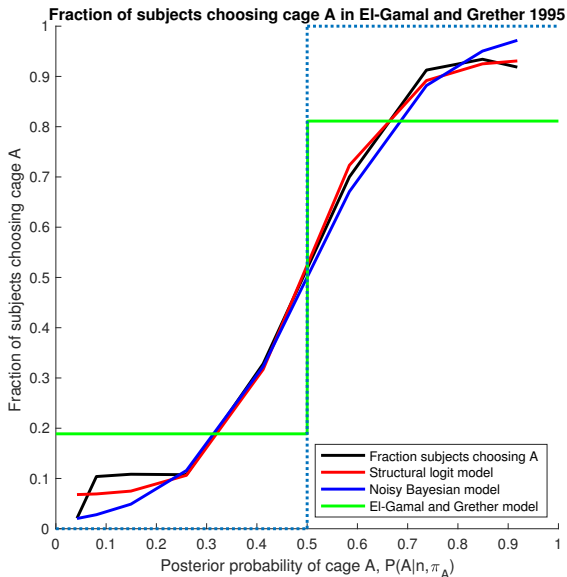
Interpretation of the structural logit model

- 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 EG 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 EG 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

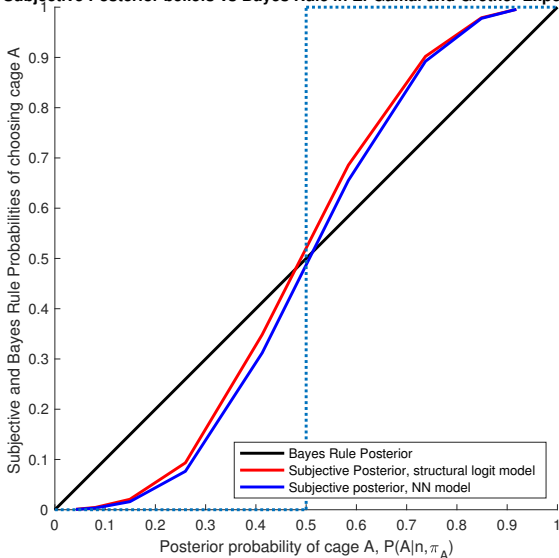
- Note also that increasing the \$10 reward for a correct classification works comparably to scaling *down* the σ “noise parameter”, explaining the finding of lower classification error rates for subjects who were paid compared to those who weren’t.
- We might interpret the extreme value shocks as a 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 EG experiments are “noisy Bayesians” (P -value 9.2×10^{-12}).

“Noisy Bayesian” vs “Subjective Bayesian” models



Inferred posterior beliefs, 1 type subjective Bayesian model

Subjective Posterior beliefs vs Bayes Rule in El-Gamal and Grether Experiment



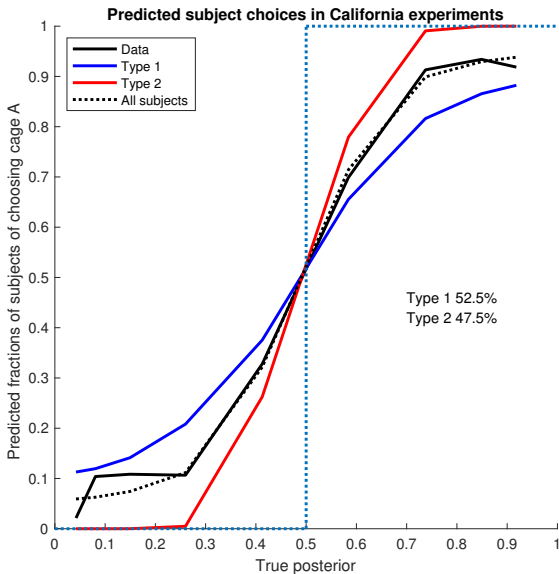
Estimating a 2 type structural logit model by EC algorithm

- This specification has 8 parameters, 4 for each of the two types identified by the EC algorithm.
- Recall that for the 1 type model $\theta = (\beta_0, \beta_1, \beta_2, \sigma)$ where $\beta = (\beta_0, \beta_1, \beta_2)$ parameterizes the subjective posterior, $\Pi(A|n, \pi_A, \beta)$ and σ is the scale value of the extreme value “noise term” in the upper level logit probability determining the subject’s actual choice.
- Under H_0 : (“Perfect Bayesian”) $\beta_0 = 0, \beta_1 = \beta_2 = -1$ and $\sigma = 0$.

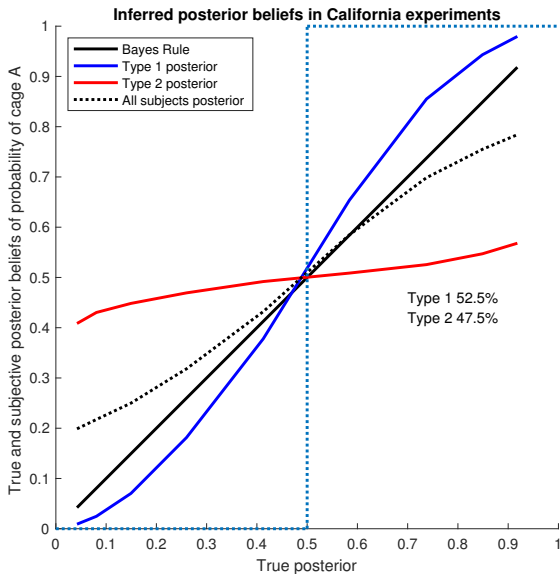
Parameter	Type 1 estimate	Type 1 std error	Type 2 estimate	Type 2 std error
β_0 ($H_0 : 0$)	-.1277	.0731	.0030	.0247
β_1 ($H_0 : -1$)	-1.400	.2928	-.1462	1.833
β_2 ($H_0 : -1$)	-1.715	.3298	-.0890	.7201
σ ($H_0 : 0$)	.4918	.0621	.0132	.1067

- Type 1 subjects are *over-confident* and noisier
- Type 2 subjects are *insensitive* to data and prior and less noisy

Estimated CCPs, 2 type subjective Bayesian model



Inferred posterior beliefs, 2 type subjective Bayesian model



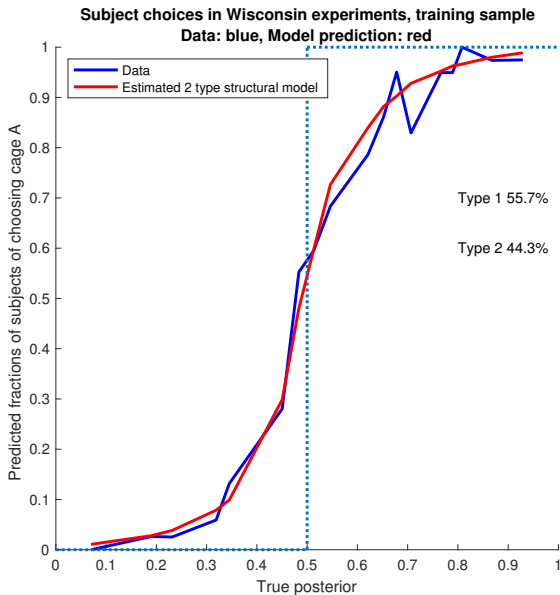
Similar conclusions from an experiment at Wisconsin

- I will discuss this experiment further below, but it exposed subjects to two different experimental designs on successive days.
- **Design 1:** Subjects were shown the results of 6 draws with replacement from cages containing 6 balls each, cage A with 4 N and 2 G balls, and cage B with 3 N and 3 G balls
- **Design 2:** Subjects were shown the results of 7 draws with replacement from cages containing 10 balls each, cage A with 4 N and 6 G balls, and cage B with 6 N and 4 G balls

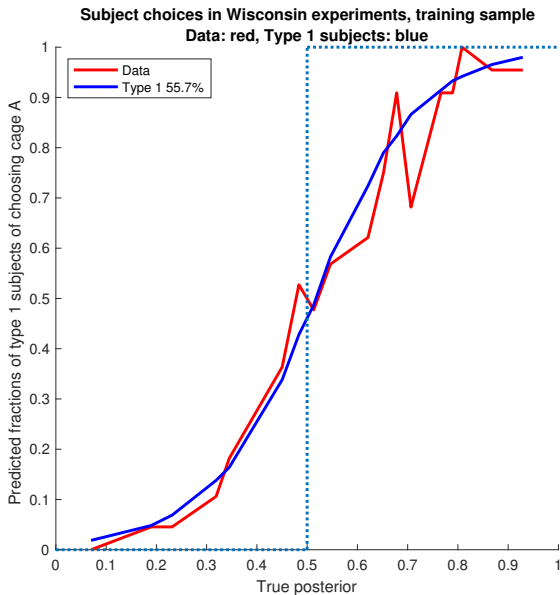
Parameter	Type 1 estimate	Type 1 std error	Type 2 estimate	Type 2 std error
β_0 ($H_0 : 0$)	.0693	.0503	-.0027	14.808
β_1 ($H_0 : -1$)	-1.068	.4785	-.0340	183.25
β_2 ($H_0 : -1$)	-1.044	.4643	-.0334	180.08
σ ($H_0 : 0$)	.2235	.0758	.0020	11.239

- Type 1 subjects are *over-confident* and noisier
- Type 2 subjects are *insensitive* to data and prior and less noisy

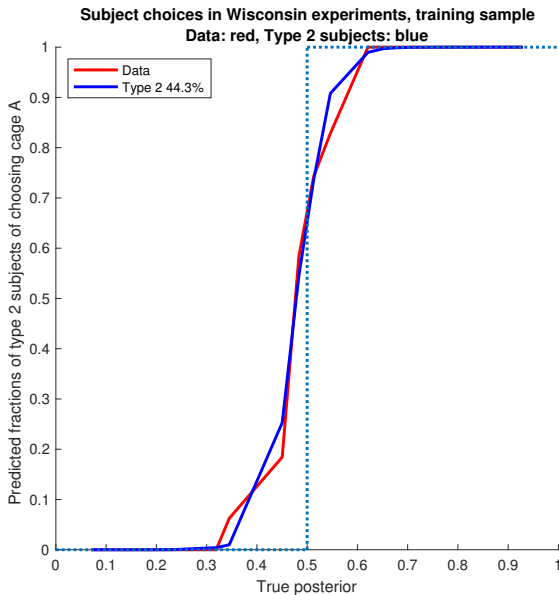
Predicted vs actual: 2 type subjective Bayesian model



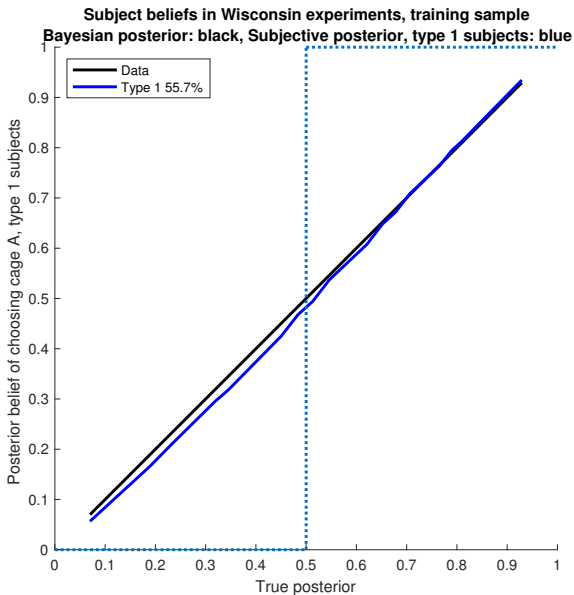
Predicted vs actual choices: type 1 subjects



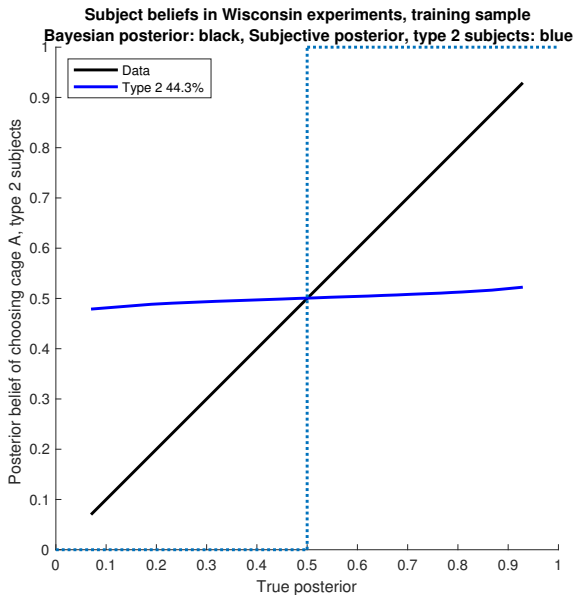
Predicted vs actual choices: type 2 subjects



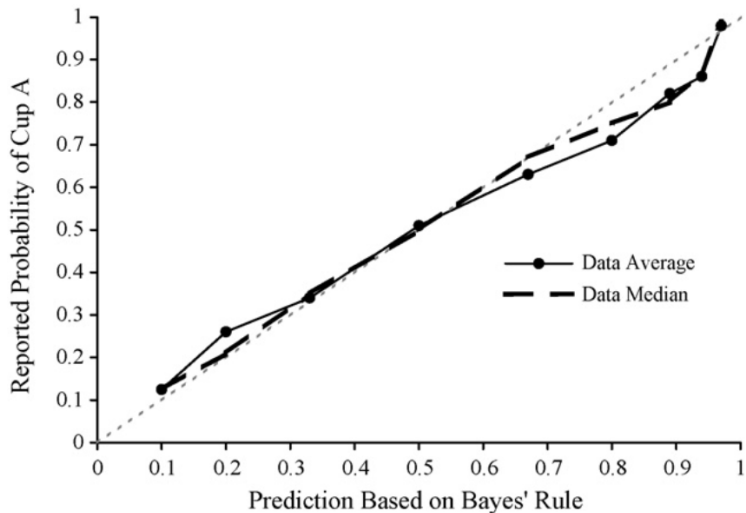
Inferred posterior beliefs: type 1 subjects



Predicted vs actual choices: type 2 subjects



Compare to elicited posteriors (Holt and Smith, 2009)



Holt and Smith used BDM mechanism to elicit beliefs

- Conducted two experiments with design similar to El-Gamal and Grether 1995 but instead of choosing which cage the sample came from (a binary response), the subjects directly reported their belief that the sample came from cage A (a continuous response).
- First experiment was done in person at UVA with 22 subjects who did 29 trials, and the second web-based experiment with 29 subjects who did 120 trials.
- To incentivize reported probabilities they used the *Becker-DeGroot-Marschak mechanism* (BDM).
- Subjects were asked to report a probability p_r and receive a reward of R from the following compound lottery involving a separate uniform random number \tilde{U} .
- if $\tilde{U} < p_r$ then the subject receives R if the sample came from cage A, and \$0 otherwise.
- if $\tilde{U} \geq p_r$ then the subject receives R with probability \tilde{U} , and \$0 otherwise.

Reanalysis of Holt and Smith, 2009 using EC algorithm

- Let $P_r(A|n, \pi_A)$ be the *reported* posterior belief by a subject in a trial where the prior of drawing from cup A is π_A and the observed number of red balls is n .
- Estimate a two type regression model of subjects' noisy reports by EC

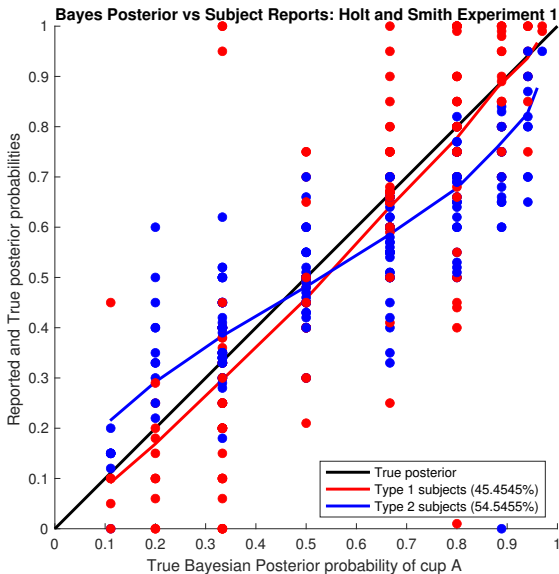
$$\log \left(\frac{P_R(A|n, \pi_A)}{1 - P_r(A|n, \pi_A)} \right) = \beta_0 + \beta_1 \log \left(\frac{f(n|A)}{f(n|B)} \right) + \beta_2 \log \left(\frac{\pi_A}{1 - \pi_A} \right) + \epsilon$$

where $\text{var}(\epsilon) = \sigma^2$.

- Under H_0 : ("Perfect Bayesian") $\beta_0 = 0, \beta_1 = \beta_2 = -1$ and $\sigma^2 = 0$.

Parameter	Type 1 estimate	Type 1 std error	Type 2 estimate	Type 2 std error
β_0 ($H_0 : 0$)	-.1618	.0818	-.0844	.0319
β_1 ($H_0 : -1$)	-.9976	.2083	-.6327	.0780
β_2 ($H_0 : -1$)	-1.0296	.0702	-.5818	.0246
σ ($H_0 : 0$)	.8974	.0473	.3962	.0182

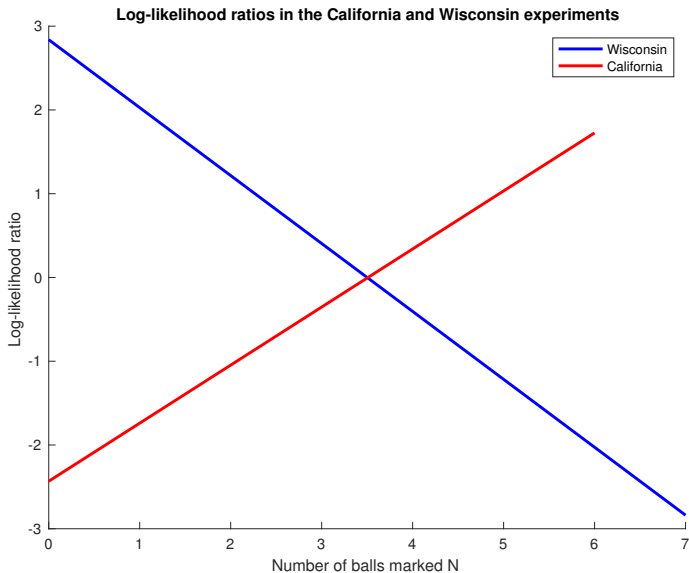
EC estimates from Holt and Smith experiment 1



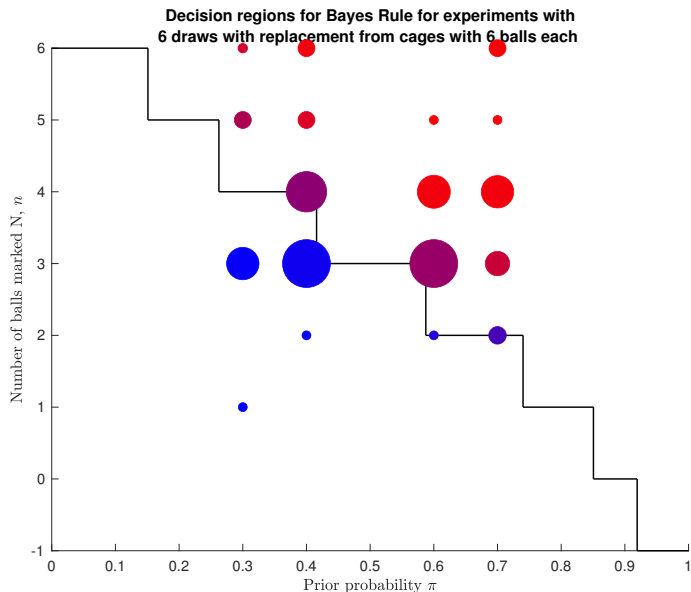
Changing decision rules: the Wisconsin experiments

- in 1999 El-Gamal and Grether conducted further experiments on learning and decision making at the University of Wisconsin involving 81 subjects.
- Subjects participated in two different experimental designs over two successive days:
- **California design:** 6 balls drawn with replacement from either Cage A (4 N balls and 2 G balls) or Cage B (3 N balls and 3 G balls)
- **Wisconsin design:** 7 balls drawn with replacement from either Cage A (4 N balls and 6 G balls) or Cage B (6 N balls and 4 G balls)
- Cage A was selected based on the draw of a 10 sided die, and π_A was chosen from the 4 values $\pi_A \in \{.3, .4, .6, .7\}$.
- Under the California design $P(N|A) = .6667$ and $P(N|B) = .5$, while under the Wisconsin design $P(N|A) = .4$ and $P(N|B) = .6$.
- This flips the likelihood ratios, so observing more N balls makes cage A more likely under the California design, but less likely under the Wisconsin design.

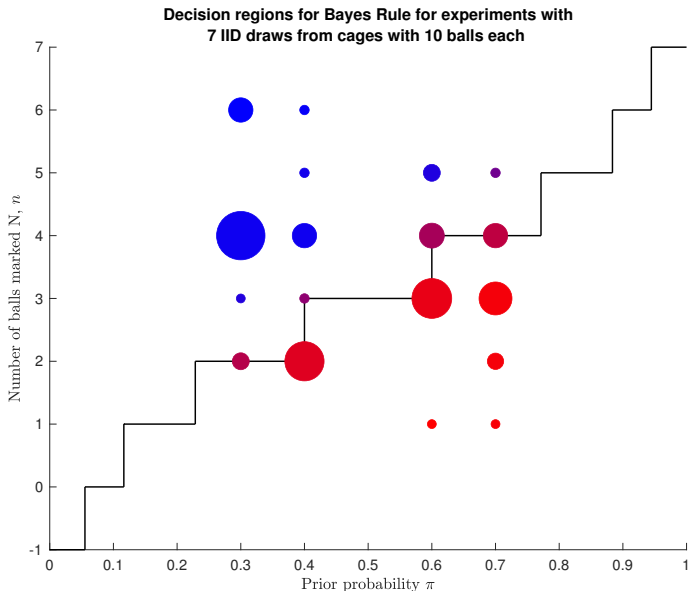
Likelihood ratios in Wisconsin experiments



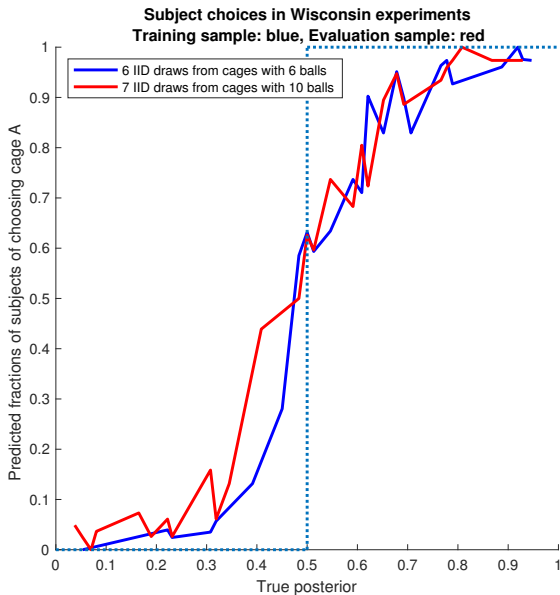
Subject choices under the California design



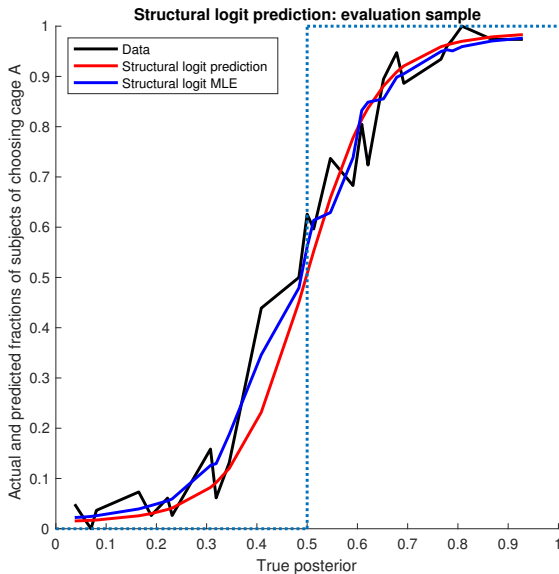
Subject choices under the Wisconsin design



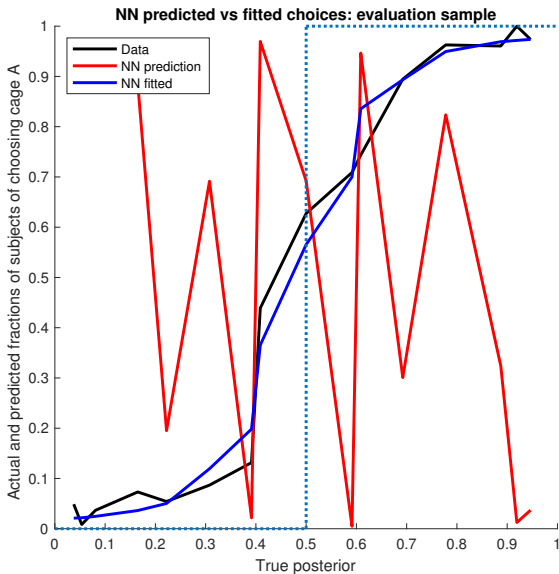
Subject choices in the Wisconsin experiments



The structural logit predicts the change in decision rules



But a reduced-form neural net specification does not



Why does the structural do better than a reduced-form NN?

The structural model can also be viewed as a NN: 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,

$$P(A|n, \pi_A, \sigma, \beta) = P(A|\Pi_s(A|n, \pi_A, \beta), \sigma) \quad (5)$$

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)* .

Generalizing: a 2 layer binary NN classifier

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

$$P(A|n, \pi_A) = \phi(\alpha_0 + \alpha_1 \phi(\beta_0 + \beta_1 T(n) + \beta_2 T(\pi_A))), \quad (6)$$

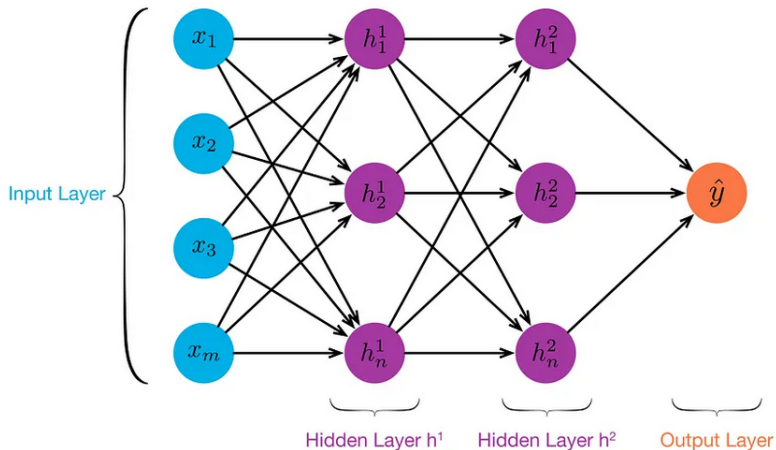
where $\phi(x) = 1/(1 + \exp(-x))$ is the logistic activation or “squashing function” and $T(n)$ and $T(\pi)$ are the “transformed inputs” given by $T(n) = \log(f(n|A)/f(n|B))$ (log-likelihood ratio) and $T(\pi_A) = \log(\pi_A/(1 - \pi_A))$ (log-prior odds ratio).

- 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 (6) when

$$\alpha_0 = \frac{-1}{2\sigma/U}, \beta_0 = 0, \alpha_1 = \frac{1}{\sigma/U}, \beta_1 = -1, \beta_2 = 1 \quad (7)$$

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

Example classification net for learning Bayes Rule



The value of “pre-transforming” the inputs to a NN

- 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 classifier when $\beta_0, \beta_1 = \beta_2 = 1$ and $\alpha_0 = -2\alpha_1$ and $\alpha_1 \rightarrow \infty$.
- 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

$$P(A|n, \pi_A) = \phi(\alpha_0 + \alpha_1 \phi(\beta_0 + \beta_1 n + \beta_2 \pi_A)), \quad (8)$$

You can show that the model (8) no longer nests the true Bayesian decision rule as a limiting special case.

Decisions actually depend on 5 pieces of information

- So far we implicitly treated subjects' decision rules as requiring only 2 pieces of information: n (number of balls marked N) and π_A (prior probability of choosing cage A).
- However the Wisconsin experiments convincingly demonstrate that human subjects are paying attention to *5 pieces of information* ($n, \pi_A, K, p(N|A), p(N|B)$) where K is the number of balls drawn from each cage with replacement, $p(N|A)$ is the probability of drawing an N ball from cage A and $p(N|B)$ is the probability of drawing an N from cage B.
- In the California design, $K = 6$, $p(N|A) = 2/3$ and $p(N|B) = 1/2$.
- In the Wisconsin design, $K = 7$, $p(N|A) = .4$ and $p(N|B) = .6$.
- A neural net that does not use all 5 pieces of information cannot approximate human behavior in experiments where $K, p(N|A), p(N|B)$ are not fixed!

The value of adding more layers

- Suppose we assume that human subjects are incapable of quickly transforming the inputs $n \rightarrow T(n)$ 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*

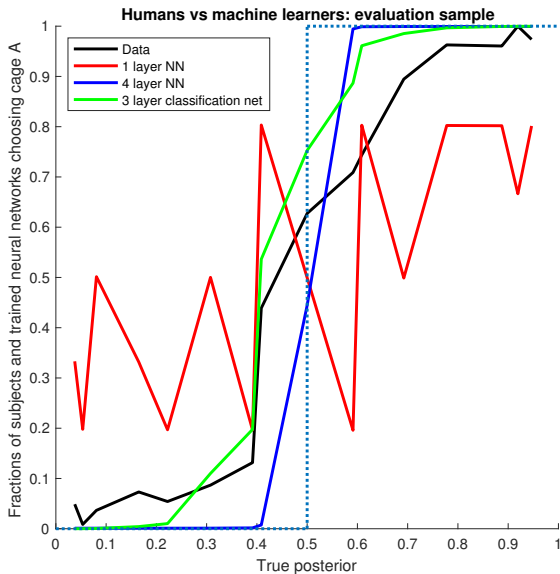
$$T(n) = \log(f(n|A)/f(n|B)) \quad \text{and} \quad T(\pi_A) = \log(\pi_A/(1 - \pi_A)).$$

- 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*.

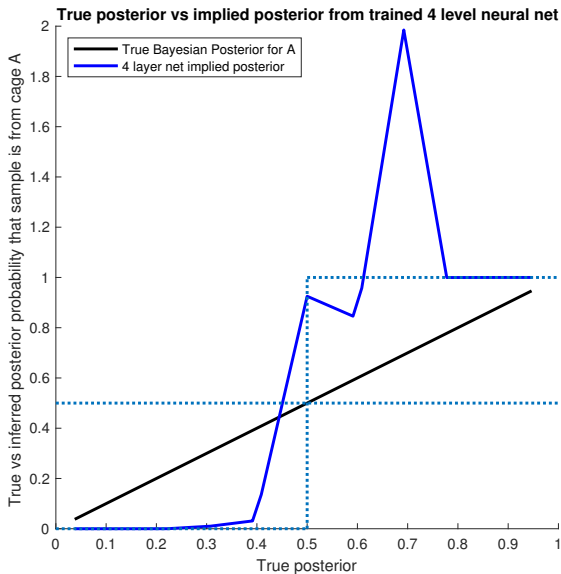
Do humans behave like poorly trained NNs?

- No.
- Human subjects had no explicit training before the experiment. Some subjects might have been exposed to Bayes Rule in a probability or stats class.
- The experiments intentionally provided no feedback to subjects to limit “learning by doing” – money payoffs were only provided at the end of the experiment.
- Human subjects had to rely on some sort of general, intuitive knowledge of probability theory to account for the different way n affects their assessment of the probability the sample was from cage A in the two designs.
- Neural nets need *sufficient complexity* and *extensive training* to make good choices, otherwise a poorly trained neural results in very poor choices in the evaluation sample.

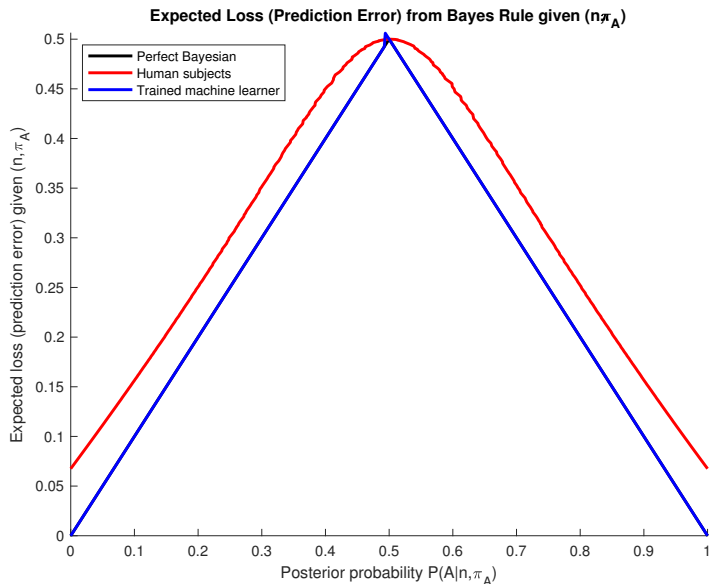
Predicted choices of well trained neural nets



Implied posterior beliefs: 3 layer classification net



The cost of not being fully rational: about 20%



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 process.”

Is ChatGPT Bayesian?

- Yes, though it can be noisy in how it applied Bayes Rule
- It is smart enough to understand, given a “word problem” that Bayes Rule can be applied to solve it, and thus it makes explicit reference to and use of Bayes Rule to make its choices.
- But in computing the actual numerical value of the posterior probability, it makes elementary algebraic errors so it sometimes gets the wrong value.
- Thus, chatGPT has the “right idea” but does somewhat “fuzzy math”.
- It also appears to be smart enough to understand the BDM mechanism.
- Humans, in contrast, find BDM quite confusing. “The downside is that BDM procedures may be difficult for subjects to comprehend.” “subjects often tense up trying to figure out what to do.” (Holt, 2022).

ChatGPT trial 1: direct elicitation of “beliefs”



Anonymous

I want you to play the role of a subject in an experiment where there are two bingo cages: cage 1 has 4 red balls and 2 blue balls and cage 2 has 3 red balls and 3 blue balls. A 6 sided die is thrown to determine which cage is used to draw a sample of 6 balls from the chosen cage where the draws are done with replacement. Suppose you are told that cage 1 is chosen if the outcome of the dice throw is 1,2,3 or 4 and cage 2 is chosen if the dice throw results in a 5 or 6. Suppose the die is tossed but you are not shown the outcome of the dice throw but then a bingo cage is chosen and 6 balls are drawn with replacement as described above. Suppose the drawing of the 6 balls results in an outcome of 4 red balls and 2 blue balls. Which of the two bingo cages was more likely to have been used to draw these 6 balls: cage 1 or cage 2? Can you estimate the probability this sample came from cage 1?



ChatGPT

To determine which cage was more likely to have been used based on the outcome of drawing 4 red balls and 2 blue balls, we'll follow a step-by-step process involving probability calculations.

Final Answer

Based on the observed outcome of drawing 4 red balls and 2 blue balls, it is more likely that the balls were drawn from cage 1, with an estimated probability of approximately 73.75%.

Key Concept

Bayes' Theorem and Probability with Replacement

- **Correct answer:** .4836

ChatGPT trial 2: using BDM to elicit “beliefs”



Anonymous

OK, I am going to draw a random number between 0 and 1, call this random uniform variable U . Let's call the number you report p . If you report p , and U is between 0 and p , then you will get \$1000 if the problem you just answered about the two bingo cages, where you observed 4 red balls and 2 blue balls drawn from either cage A or cage B (I have not revealed to you which one was actually used yet), you will get \$1000 if the cage used to draw this sample is actually cage A, but nothing if the sample was drawn from cage B. However if the random number U is bigger than p , the number you report between 0 and 1, then there is a separate lottery played where you will get 1000 with probability U and 0 otherwise. Remember that U itself is a random number, but once its value is realized, it will have some specific value such as $U=.75$. Then once the value of U is revealed, if it is .75 and you report $p=.5$, then you will get \$1000 with probability .75 and 0 with probability .25. Thus, we are asking you to report a single number p to maximize your chance of winning the \$1000 in this "compound lottery" (we call it a compound lottery since it involves several random drawings: the drawing of 6 balls from one of the bingo cages, the drawing of the random number U between 0 and 1, and then given the realized value of U , you get 1000 with probability U if U is between p the number you report and 1. What p should you report to maximize your chance of winning the \$1000 prize?

ChatGPT's answer to trial 2



ChatGPT

To maximize your probability of winning the \$1000 prize in this compound lottery, you should report p equal to the probability that Cage A was the source of the draw when observing 4 red balls and 2 blue balls. This probability was calculated earlier using Bayes' theorem and was found to be approximately 0.484.

So, the specific value of p you should report to maximize your chance of winning in this compound lottery, based on the information given and calculations performed, is $p = 0.484$. This value maximizes your chances because it aligns with the probability that Cage A was used, under the condition of observing 4 red balls and 2 blue balls. This strategy ensures that your chosen p optimally balances the likelihood of winning based on both the conditional probability related to Cage A and the unconditional probability related to U when $U > p$.

- This time chatGPT (4.0) gets it right despite the added complexity of the BDM mechanism and having to remember the answer to the previous question describing the bingo cage experiment and the result of the drawing (4 N balls and 2 G balls with a prior of $\pi_A = 2/3$).

Summary of chatGPT responses

- Direct elicitation of posterior probability (no BDM)

Truth	T1	T2	T3	T4
.4836	.5423	.737	.737	.798

- Using BDM for incentive-compatible elicitation of posterior probability

Truth	T1	T2	T3	T4	T5	T6	T7
.4836	.351	.4058	.483	.484	.484	.571	1.0 (chatGPT 3)

- Thus, despite making errors, chatGPT does not behave in a human-like manner
- Instead, it behaves like a math-challenged version of Star Trek's Dr. Spock
- Its ability to “see through” and “interpret” word problems seems amazing, particularly its ability to recognize that Bayes Rule is relevant for making these judgements, and its ability to see that under the BDM mechanism its optimal report of p is just its best estimate of the posterior probability of cage A.