Lecture 1: Introduction to Dynamic Programming and Structural Estimation

Econ 619, Structural Econometrics

John Rust, Georgetown University

March 10, 2021

Road Map

Intro to Dynamic Programming and Structural Estimation

1. What is dynamic structural estimation?
2. Static Structural Estimation: discrete choice models
3. Review of Dynamic Programming
4. Dynamic Structural Estimation: dynamic discrete choice models
5. The limits of inference with and without theory

What is structural estimation?

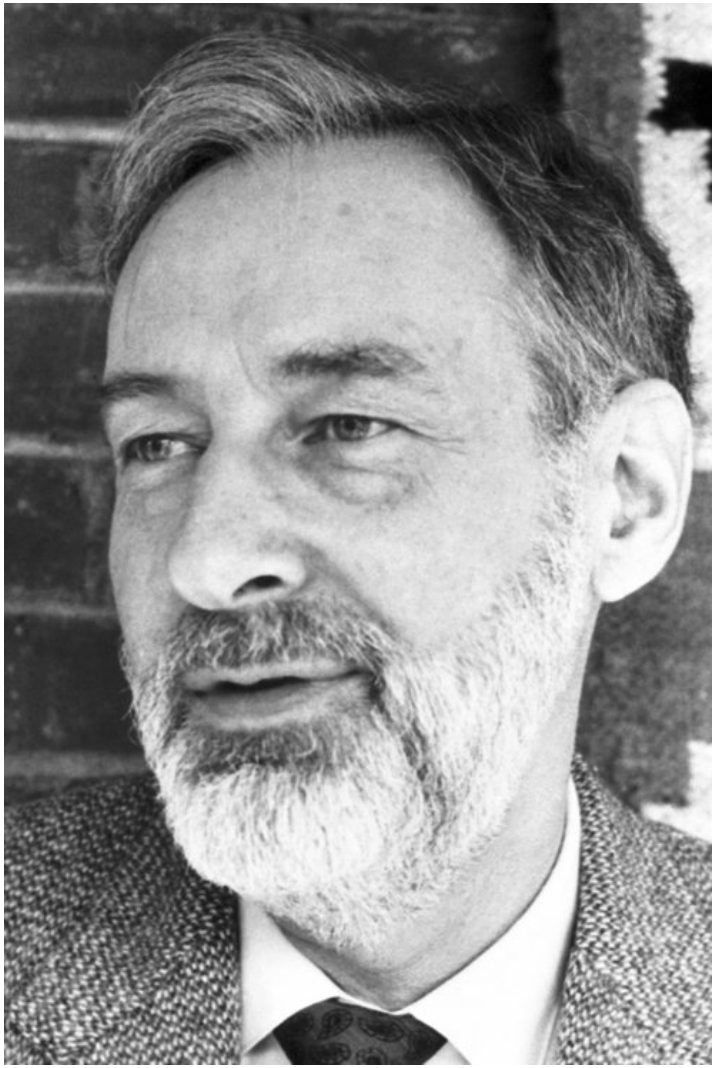
- **Econometrics:** a branch of statistics focused on economic measurement, prediction, and the development and testing of economic theories
- The term "structural" arose in the 1940s, and is credited to Trygve Haavelmo and Tjalling Koopmans, Jacob Marschak and other "founding fathers" at the Cowles Commission. It appears in the 1949 *Econometrica* paper by Koopmans, "Identification Problems in Economic Model Construction"

Statistical inference, from observations to economic behavior parameters, can be made in two steps: inference from the observations to the parameters of the assumed joint distribution of the observations, and inference from that distribution to the parameters of the structural equations describing economic behavior. The latter problem of inference, described by the term "identification problem," is discussed in this article in an expository manner, drawing on other more original work for concepts and theorems, and using a number of examples drawn partly from econometric literature.

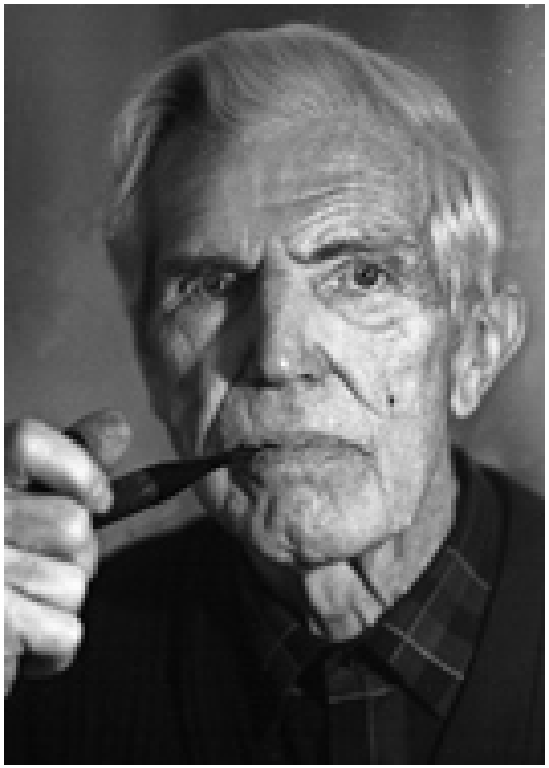
Origins of structural econometrics

- Intellectual origins: Cowles Commission and figures such as Jacob Marschak, Trygve Haavelmo and Tjalling Koopmans in the late 1940s
- Koopmans' essay, *Measurement without theory* criticized the "decision not to use theories of man's economic behavior, even hypothetically" because the absence of theory "limits the value to economic science and to the maker of policies" and "greatly restricts the benefit that might be secured from the use of modern methods of statistical inference"
- But disagreement on this is long-standing. Koopmans moved the Cowles Commission from Chicago to Yale due to "rising hostile opposition . . . by the department of economics at the University of Chicago in the 1950s"
- Why the hostility to efforts to bridge economic theory and empirical work?

Tjalling Koopmans



Trygve Haavelmo



Jacob Marschak

But even before them, there was Kant . . .

Kant paraphrased

Theory without empirics is empty, empirics without theory is blind

or

theory without practice is empty; practice without theory is blind

or

Karl Marx

practice without theory is blind; theory without practice is sterile

The Limits to inference without theory

- title of Ken Wolpin's 2010 monograph that grew out of lectures at the *Cowles Foundation* in honor of *Tjalling Koopmans* who wrote a 1947 book review of a book *Measuring Business Cycles* by Burns and Mitchell.
- Koopmans titled his review **Measurement without Theory**

From my review of Wolpin's monograph in the 2015 *Journal of Economic Literature*

His main message is clear from the title: the denial of the value of economic theory as a source of hypotheses and as a guide to empirical work remains pervasive in the profession, results in unnecessary limits on what we can learn, and cripples our ability to do counterfactual policy forecasting and analysis.

The Limits to inference without theory

- At the risk of oversimplifying, empirical work that takes theory "seriously" is referred to as *structural econometrics* whereas empirical work that avoids a tight integration of theory and empirical work is referred to as *reduced form econometrics.*
- The raison d'être of the Cowles Foundation is to promote a tighter integration between theory and measurement in economics.
- Cowles motto: *theory and measurement*
- However I believe it is a mistake to identify structural econometrics with *rational economic theory* (e.g. expected discounted utility maximization, optimization and equilibrium, etc) or the

Minnesota motto

Optimalité, Equilibraté, Calibraté

- Instead I am very comfortable embracing sub-rational, disequilibrium *behavioral* and *bounded rationality* theories of individual and firm behavior as part of a new branch, which we could call structural behavioral economics

Key to structural econometrics: Models

- What is a model? Wolpin's monograph *The Limits of Inference without Theory* did not define it. However models underlie all of structural econometrics.
- Implicit in Wolpin's title is the belief that methods of inference that fail to make use of a model or a theory will be somehow limited in what can be learned/inferred from any given data, compared to methods of inference that use models and theories.
- Sargent's definition of *model* "A model is a probability distribution over a sequence (of vectors) usually indexed by some fixed parameters"
- Marschak (1953) definition of *structure* "(1) a set of relations describing human behavior and institutions as well as technological laws and involving in general, nonobservable random disturbances and nonobservable random errors in measurement; (2) the joint probability distribution of these random quantities."

Structural behavioral economics

- Note there is no mention of rationality, optimization or even equilibrium in Marschak's definition. Does structural econometrics require the assumption of rationality, maximization and equilibrium?
- **NO!** Structural estimation/inference applies to all types of models including "behavioral" models that relax assumptions of rationality, optimization and equilibrium.
- Stefano DellaVigna *Handbook of Behavioral Economics* "Structural Behavioral Economics"
- "Behavioral economics has benefited from a close relationship between behavioral theory and empirics, which structural estimation can build on. Behavioral economics has also made important use of calibration of magnitudes of effects, and structural estimates take calibrations one step further. Experimental evidence has also played a key role in behavioral economics, and model-based designs can provide useful input already at the design stage of the experiment."

Structure and the Lucas Critique

- There is an older meaning of the terms *structure* and *structural model* originating from work by Koopmans, Haavelmo, and Marschak at Cowles dating back to the 1940s: under this view the goal of inference is to specify and identity *deep parameters* that are *policy invariant*
- "A structure is defined (by me, following Hurwicz 1962 and Koopmans and Bausch 1959) as something which remains fixed when we undertake a policy change, and the structure is identified if we can estimate it from the given data" Sims, 1980, "Macreconomics and Reality"
- *Lucas critique* — "given that the structure of an econometric model consists of optimal decision rules of economic agents, and that optimal decision rules vary systematically with changes in the structure of series relevant to the decision maker, it follows that any change in policy will systematically alter the structure of econometric models"

Reduced-form vs structure

- We use the term *reduced-form model* to denote an econometric model that may not necessarily be derived from, or tightly linked to a particular economic model or theory
- Whereas a *structural model* is an econometric model explicitly derived from an economic model or theory and we want to estimate/infer the parameters to test/evaluate the theory
- Example: *linear simultaneous equations model.* This is a simple theory of economic equilibrium with linear supply/demand equations that constitute the model.

$$\begin{aligned} &\text{structural model} \quad Y\Gamma = XB + U \quad \theta = (\Gamma, B, \Sigma) \\ &\text{reduced-form model} \quad Y = X\Pi + V \quad \Pi = B\Gamma^{-1} \end{aligned}$$

- θ are the structural parameters, (Π, Ω) are the reduced-form parameters. $\Pi = B\Gamma^{-1}$ and $\text{cov}(V) = \Omega = \Gamma'^{-1}\Sigma\Gamma^{-1}$.

The identification problem

- The reduced-form equation $Y = X\Pi + V$ holds regardless whether the underlying structural model is correct or not. The Π and Ω parameters can be estimated by OLS.
- *Identification problem:* when it is possible to invert the mapping from the reduced-form parameters to the structural parameters?

$$\theta = (\Gamma, B, \Sigma) \longleftrightarrow (\Pi, \Omega)$$

In general, there are more structural parameters than reduced-form parameters and we need additional restrictions such as *exclusion restrictions* and *cross-equation restrictions* to identify θ.

Hausman and Taylor, 1983 *Econometrica*

the early work at the Cowles Foundation determined necessary and sufficient conditions for identification and related these to maximum likelihood estimation and clarified the relationship between identifiability and instrumental variables estimation: i.e., that the restrictions required for identification give rise to the instrumental variables required for estimation

General view of identification

- Under what considitions can we do this inversion, and using only data we observe uniquely identify the model/theory that we hypothesized as the "data generating mechanism"?
- Is it always possible to find structural parameters to "rationalize" any data we observe? If so, does theory have testable content?
- Some theories are sufficiently general/flexible that without sufficient restrictions on the *primitives* of the the theory (e.g. preferences, beliefs of agents, etc) then it is possible to *rationalize any observed behavior.* When this is the case, why do structural estimation?

Heckman and Navarro, 2005

His paper (i.e. my 1994 Handbook of Econometrics chapter) has fostered the widespread belief that dynamic discrete choice models are identified only by using arbitrary functional form and exclusion restrictions. The entire dynamic discrete choice project thus appears to be without empirical content and the evidence from it at the whim of investigator choice about function forms of estimating equations and application of ad hoc exclusion restrictions.

Why do structural estimation?

- To do *counterfactual analyses and predictions* often known as *policy analysis*
- The general idea of the Cowles Commission approach to inference is that *structure is policy-invariant* which do not change if *policies* change (e.g. taxes, regulations, other government/firm policies).
- Let θ represent the structure (e.g. preference, production function parameters, etc). and let π represent policy variables under the *status quo*
- Suppose there are "endogenous variables" y (y might represent firm outputs or prices or consumer choices/demand) and observable "exogenous variables" x (i.e. x might represent weather) as well as ϵ unobserved exogenous variabes. The economic model predicts a reduced form relationship that could be written as

$$y = f(x, \epsilon, \theta, \pi) \tag{1}$$

Why do structural estimation?

- We can think of reduced-form estimation as avoiding the identification problem, and the need to specify a distribution of unobservables $g(\epsilon)$ and just use flexible, non-parametric methods to estimate $P(y|x)$, the conditional probability of y given x.

$$P(y \in B|x) = \int I\{f(x, \epsilon, \theta, \pi) \in B\} g(\epsilon) d\epsilon \tag{2}$$

- **Problem** While $P(y|x)$ may provide a good forecast of behavior/endogenous outcomes under the *status quo* policy π it will not provide a good prediction under a counterfactual policy π'. Further it will not tell us a great deal about *welfare* and *distributional consequences* (i.e. who is hurt who is helped by different policies).
- Structural estimation and identification attempts to invert and use the reduced form $P(y|x)$ to map back to the *structure* $[\theta, g, f]$ separately from the policy variables π.

Observational equivalence and identification

- The reduced form relationship $P(y|x)$ is identified because we can use non-parametric estimation to estimate it given sufficient data without many additional assumptions.
- However not so for the structure: $[\theta, g]$ we say that two structures $[\theta, g]$ and $[\theta', g']$ are **observationally equivalent** if they map into the same reduced form

$$\begin{aligned} P(y \in B|x) &= \int I\{f(x, \epsilon, \theta, \pi) \in B\} g(\epsilon) d\epsilon \\ &= \int I\{f(x, \epsilon, \theta', \pi) \in B\} g'(\epsilon) d\epsilon \end{aligned}$$

- We say that the structural model is *identified* if a) the data are generated by a "true structure" $[\theta^*, g^*]$ and b) no other structure $[\theta, g]$ is observationally equivalent to the true structure $[\theta^*, g^*]$.
- Equivalently the structure $[\theta^*, g^*]$ is identified if there is a 1 to 1 mapping between the reduced form $P(y|x)$ and $[\theta^*, g^*]$

Policy forecasting using structural models

- The economic model can enable us to assess welfare effects and distributional effects of a policy change from π to π' and results in a counterfactual prediction

$$P(y \in B|x, \theta, g, \pi') = \int I\{f(x, \epsilon, \theta, \pi') \in B\} g(\epsilon) d\epsilon \qquad (3)$$

- This is the big advantage of structural estimation: the ability to do counterfactual forecasts and simulations!
- Contrast this with reduced-form methods. If we estimate $P(y|x)$ under the *status quo* and do not attempt to disentangle policy parameters π from the structural parameters $[\theta, g]$ we will generally be unable to predict how $P(y|x)$ changes when the policy shifts from π to π'.
- In general, the reduced form will shift when the policy shifts, so

$$P(y \in B|x, \theta, g, \pi) \neq P(y \in B|x, \theta, g, \pi') \qquad (4)$$

and thus the reduced-form relationship we estimate under the *status quo* is not likely to hold, and provide accurate forecasts of behavior/outcomes under an alternative policy π'

Does policy forecasting require identification?

- Not always! But if we mis-identify the structure $[\theta, g]$ that is observationally equivalent to the true structure $[\theta^*, g^*]$ even though *in-sample* (i.e. under the *status quo*) we have $P(y|x, \theta, g, \pi) = P(y|x, \theta^*, g^*, \pi)$ for other policies $\pi' \neq \pi$ we could have $P(y|x, \theta, g, \pi') \neq P(y|x, \theta^*, g^*, \pi')$ so using the misidentified structure could result in incorrect policy forecasts
- When not all elements of the structure are identified, we say the model is *partially identified.* Sometimes we cannot *point identify* all parameters θ but can provide *bounds* on these parameters so we say the parameters are *set identified.*

From Heckman's slides

Marschak noted that for many specific questions of policy analysis, it is not necessary to identify fully specified economic models that are invariant to classes of policy modifications.

- See work by Kalouptsidi, Kitamura, Lima and Souza-Rodrigues, "Partial Identification and Inference for Dynamic Models and Counterfactuals" at the DSE2019 conference.

The Limits to Inference *With* Theory

- Title of my 2015 book review of Wolpin in the *Journal of Economic Literature*
- Asks the question: is structural inference in the strict Cowles Foundation sense really possible?
- Structural econometricians like to assume that preference parameters and technology parameters constitute the *deep structural parameters* and other parameters, particular those governing *economic policy* (e.g. tax rates, government regulations, and the sructure of economic institutions, such as the rules used in auctions, etc) are more like variables that can be changed.
- Using DP and game theory, we can use structural models to generate *counterfactual predictions* of how agents' behavior will change when these policy parameters are changed.
- But what if there really are no fully policy, technology, or socially/culturally independent parameters or objects?

The Limits to Inference *With* Theory

- Joseph Stiglitz, in his 2001 Nobel Prize lecture stated "There were other deficiencies in the theory, some of which were closely connected. The standard theory assumed that technology and preferences were fixed. But changes in technology, R&D, are at the heart of capitalism . . . I similarly became increasingly convinced of the inappropriateness of the assumption of fixed preferences."
- In fact, few things seem truly "invariant" these days, other than Kurzweil's (1999) *law of accelerating change* — the rate of technological progress is itself accelerating at an exponential rate. We are approaching *the singularity* when the rate of change will become nearly infinite.
- Far from there being any structural, invariant parameters or objects, accelerating change is making the future inherently more unpredictable. Rapidly evolving technology and knowledge alters our behavior and institutions, and the structure of individual preferences and decision making, and thus the economy as a whole.

Other Limits to Inference *With* Theory

- The curse of dimensionality: limits the degree of realism of models we can solve, simulate and estimate
- The multiplicity of equilibria: results in indeterminacy that limits the predictive empirical content of many of our models
- The identification problem: multiple *observationally equivalent structures* may predict the same reduced form behavior. However these different structures may result in different predicted counterfactual responses to policy changes
- Though these are very challenging problems, I agree with Wolpin and the Cowles Foundation that economists have far more to gain by trying to incorporate economic theory into empirical work and test and improve our theories than by rejecting theory and presuming that all interesting economic issues can be answered by well-designed controlled, randomized experiments and assuming that difficult questions of causality and evaluation of alternative hypothetical policies can be resolved by simply allowing the "data to speak for itself"

Though theory has limits, we can still learn a lot

- Gödel's Theorem shows that there are strict limits to knowledge using *deductive inference* — there are true mathematical theorems for which proofs do not exist.
- Given this, it should not be a surprise that there are also limits to knowledge using *inductive inference.*
- Yet, consider how much the human race has learned despite these fundamental limits to knowledge: in 2013 humans discovered the *God particle* (Higgs Boson) more than 5 decades after it was predicted to exist by fairly abstract physical theories
- In neuroscience, there is growing evidence that the human brain has an amazing innate, subconscious ability to *model and simulate reality.*
- Many neuroscientists believe that one of the keys to human intelligence is precisely our incredibly powerful ability to generate and modify internal mental models of the world.

Mental models: a key to human intelligence

- "People can infer causal relationships from samples too small for any statistical test to produce significant results . . . and solve problems like inferring hidden causal structure . . . that still pose a major challenge for statisticians and computer scientists." Griffiths and Tenenbaum (2009) "Theory-Based Causal Induction" *Psychological Review*
- Yet, our internal mental models can be wrong, or poor approximations to reality. But what do we do when this happens? If the consequences are serious enough, *we undertake experiments to gather more data and revise our mental models.*
- For example decades of experiments with infants suggest that humans are born with substantial amount of *a priori* theories and assumptions, such as basic understanding of laws of gravity, matter, and ability to do three dimensional visual spatial processing and reasoning, etc. Psychologists refer to this prior knowledge as *core knowledge*

Our brains are innate master structural econometricians!

- Numerous psychological experiments have shown how humans use experimentation to helps them develop better mental models in cases where their core knowledge conflicts with observations.
- A recent study by Stahl and Feigenson 2015 reported experiments that "tested learning after violations of expectations drawn from core knowledge of object behavior — knowledge that is available from early in life, is universal across human cultures, and is present in other species. . . . our experiments reveal that when infants see an object defy their expectations, they learn about that object better, explore that object more, and test relevant hypotheses for that object's behavior."
- Griffiths and Tenebaum argue that "human causal induction — the inference to causal structure from data — is the result of a statistical inference comparing hypotheses generated by a causal theory. This approach explains how people are able to infer causal relationships from small samples and identify complex causal structures."

Contrast this with the trend in econometrics

- The focus is how to do inference under the *weakest possible assumptions.* In fact, making assumptions is nearly demonized by Manski, who states that "policy predictions often are fragile. Conclusions may rest on critical unsupported assumptions or on leaps of logic. Then the certitude of policy analysis is not credible." 2013 *Public Policy in an Uncertain World*
- Three decades prior to Manski, Edward Leamer disparaged the state of applied econometrics for many of the same reasons in his famous (1983) "Let's Take the Con Out of Econometrics" paper. He questioned the validity of inference due to the practice of *specification searching*
- This trend towards the growing agnosticism is reflected in the newer literature on *partial identification* which shows that when we weaken assumptions sufficiently we can only *set identify* parameters or other objects of interest.
- While we know the saying "garbage in, garbage out", when we take partial identification to its logical extreme, we get the principle of *nothing in, nothing out*

Statistics: an inadequate theory of learning

- While the critiques by Leamer and Manski were motivated by legitimate concerns, they have had the unintended effect of demonizing activities that really are critical parts of the scientific process.
- In particular, specification searching can be viewed as an informal model selection procedure by which researchers discard models that do not fit the data well, as they search for others models that fit the data better. It seems crazy to demonize this activity because existing econometric theory has great difficulty formalizing how this search process affects our inferences.
- Similarly, it seems crazy to criticize researchers for the act of inventing models and making reasonable assumptions and other "leaps of logic" because some of these assumptions could be wrong
- What Leamer, Manski and their disciples ignore is that while yes, some models and assumptions are bad and can cause us to misunderstand casual relationships and make erroneous policy decisions, science advances through the process of rejecting bad models and searching for new, better ones.

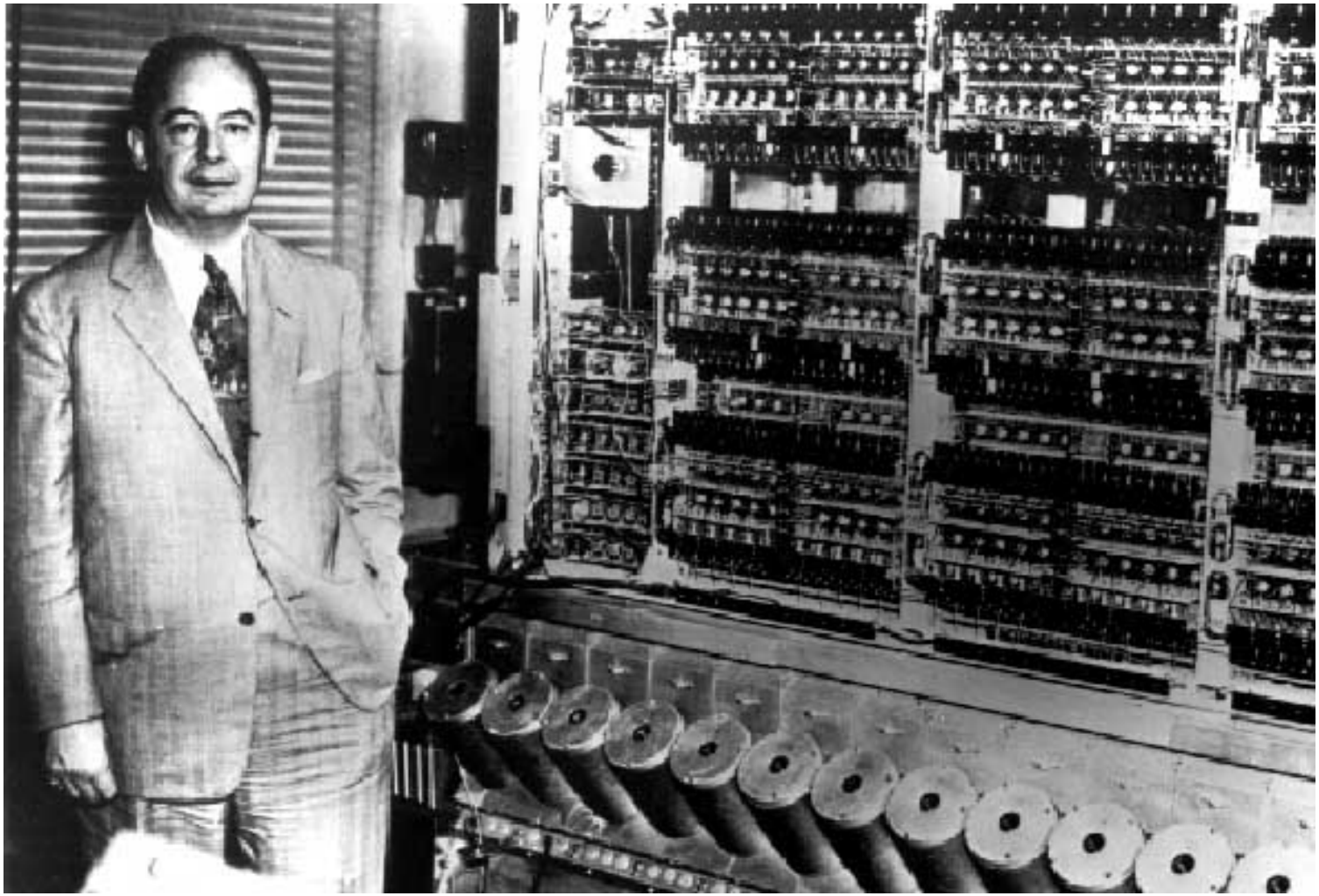
Statistics: an inadequate theory of learning

- To develop the newer, better models researchers need to 1) make assumptions, and 2) engage in specification searches. Discouraging these activities as examples of bad empirical practice actually has opposite effect: it greatly impedes our ability to make scientific progress.
- Econometric theory conveys an impression that statistics and econometrics is a completely "solved problem" and that all of the relevant pieces of a full mathematical model of inductive inference and learning are in place.
- Statistics and econometrics are very far from constituting an adequate theory or guide to empirical scientific discovery and are an inadequate and incomplete theory of how individual scientists and the scientific community at large should optimally learn from data.
- "Since we do not fully understand the process of discovery or the social nature of the knowledge achieved from this process (agreement in the scientific community) and the role of persuasion in forming consensus, it is not surprising that mathematically precise models of discovery are not available." Hackman 1992

We are neophytes in *formal* modeling and inference

- Taking modeling from the internal, subconcious domain to the conscious, formal, and symbolic domain is only relatively recent in evolutionary history. It may have begun with the advent of spoken language, then writing, and development of symbolic reasoning systems (e.g., mathematics) and modern science.
- Yet, the result of this has been fundamentally transformative to human evolution, in effect vastly speeding up the rate at which natural evolution occurs.
- The advent "artificial brain" — the modern digital computer or "von Neumann machine" is itself an extremely recent paradigm shift in evolution — having occured only 60 years ago.
- So perhaps we cannot be too hard on ourselves for being relatively clumsy at formal modeling and being relatively primitive in our attempts to build our first artificial brains and slow in finding the secrets of *artificial intelligence*

John von Neumann 1903–1957



But we are improving *very very quickly*

- The rate of change in our ability to do computations on artificial brains is breathtakingly rapid. For example, *Moore's law* implies that computing power is increasing at the amazing rate of 46 per year!
- Thus, I am very confident that a combination of better hardware and software will revolutionize our ability to formally model and predict complex phenomena, including *homo economicus*
- I would be willing to place a large bet that it won't be long before some of the key "secrets to intelligence" will be discovered and "reverse-engineered" and the people who do this will be the revolutionaries who usher us into the brave new world.
- In any event, *end of sermon!* If you want to find out more about my heretical views on the current state of econometric and economic theory, read my essay, *Most Useless Econometrics? Assessing the Causal Effect of Econometric Theory* in *Foundations and Trends in Accounting* in a volume arising from a conference on *Causality in the Social Sciences* held at Stanford Business School in December 2014.

Overview of where we are going

- Structural estimation of static models: the static discrete choice model
- Add dynamics. Fundamental tool: Dynamic Programming
- Positive application of DP: to interpret data in *structural estimation* (inverse optimal control)
- Normative application of DP: for *intelligent design* (finding better decision rules)
- In both cases we use *models* as laboratory for analyzing *causality* and *counterfactual predictions*
- DP is powerful due to its flexibility and breadth: it provides a framework to study decision making over time, under uncertainty, and can accomodate *learning,* strategic interactions between agents (game theory) and market interactions (equilibrium theory).

Static Discrete Choice Models

- We start by reviewing *static discrete choice theory*
- Problem: a decision maker (DM) in state x chooses an alternative d from a finite set $D(x)$ of possible alternatives to maximize $u(x, d)$
- Economic approach: subject uses a *decision rule* $d^*(x) = argmax_{d \in D(x)} u(x, d)$
- If we know the person's state x and utility function $u(x, d)$ and choice set $D(x)$, then we can *perfectly predict the choice* $d^*(x)$ *they will make*
- *Probabilistic choice theory* The choice d^* is not perfectly predictable because it depends on *private information* ϵ that the DM observes that we (as the econometrician) do not observe
- Decision rule is $d^*(x, \epsilon) = argmax_{d \in D(x)}[u(x, d) + \epsilon(d)]$ which is a *random variable* that is not perfectly predictable. RUM = *Random Utility Maximization*
- Define the *conditional choice probability* $P(d|x) = \text{Prob}\{d^*(x, \epsilon) = d|x\} = \int_\epsilon I\{d^*(x, \epsilon) = d\}\, q(\epsilon|x)$

Probabilistic choice theory

- Early work by psychologists such as Thurstone *The measurement of values* and Duncan Luce *Individual choice behavior* both published in 1959
- $P(d|D,x)$ conditional probability of choosing alternative $d \in D(x)$ for a subject with observed characteristics x (x might also capture *attributes* of the alternatives in $D(x)$)

Independence from Irrelevant Alterantives (IIA) Axiom

If $B(x) \subset D(x)$ and $d \in B(x)$ then

$$P(d|D,x) = P(d|B,x)P(B|D,x)$$

where

$$P(B|D,x) = \sum_{d \in B(x)} P(d|D,x)$$

The multinomial logit model

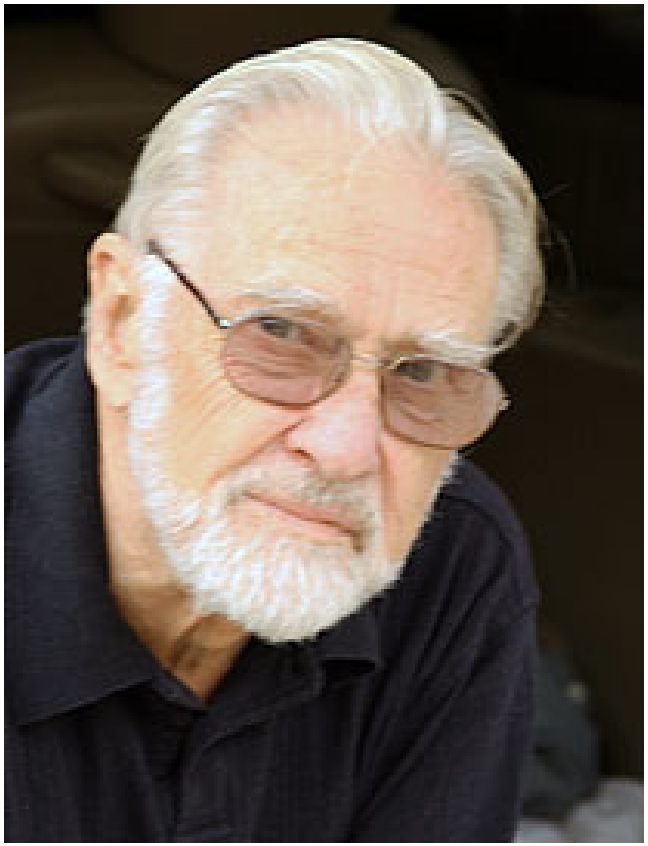
Luce's Theorem

If IIA holds, then there exist non-negative weights $v(x,d)$, $d \in D(x)$ such that

$$P(d|D,x) = \frac{exp\{v(x,d)\}}{\sum_{d' \in D(x)} \exp\{v(x,d')\}} \tag{5}$$

- These choice probabilities are known as the multinomial logit model
- Luce: "I have worked in the highly probabilistic area of psychophysical theory; but the empirical materials have lead me away from axiomatic structures, such as the choice axiom, to more structural, neural models which are not easily axiomatized at the present time. After some attempts to apply choice models to psychophysical phenomena I was lead to conclude that it is not a very promising approach to these data"

Duncan Luce 1925–2012



The IIA property of the MNL model

- The MNL model is based on the assumption of *independence of the error terms* in the RUM: i.e. $\epsilon(d)$ and $\epsilon(d')$ are independently distributed if $d \neq d'$.
- Luce's *Independence from Irrelevant Alternatives* (IIA) Axiom: *The odds of selecting one alternative d relative to another alternative d′ is independent of the size/composition of the choice set D(x) or the utilities of other choices not equal to d or d′*
- The MNL model satisfies the IIA Axiom:

$$\frac{P(d|x)}{P(d'|x)} = \frac{\exp\{u(x,d)\}}{\exp\{u(x,d')\}}$$

  independent of the size of $D(x)$ or the values of $u(x,d'')$ for $d'' \neq d$ and $d'' \neq d'$.
- Luce (1959) proved that the MNL is the *only choice probability* that satisfies the IIA axiom. Marshak (1960) provided a RUM interpretation of the MNL model using the Type 1 extreme value (GEV/Gumbel) family of probability distributions.

Testable implications of the IIA property

- Suppose we can do experiments and alter subject choice sets. Then via *laboratory or field experimentation* it may be possible to reject some specifications of the RUM.
- **Red bus/blue bus paradox** (Debreu) Suppose a person can commute to work by "blue bus" b or car c and $u(x,b) = u(x,c)$. Then the MNL model predicts a 50/50 probability for either model choice: $P(b|x) = P(c|x) = .5$.
- Now suppose we do an experiment introducing a third artificial alternative, "red bus" r and assume $u(x,r) = u(x,b) = u(x,c)$. Then the MNL predicts the choice probabilities following this new artificial alternative is $P(r|x) = P(b|x) = P(c|x) = 1/3$. But a more reasonable prediction is that the choice probabilities should be

$$P(r|x) = P(b|x) = .25 \quad P(c|x) = .5$$

- The independence of the random errors $\{\epsilon(r), \epsilon(b), \epsilon(c)\}$ is called into question here. The observed attributes of two bus alternatives are essentially identical (except color) and thus we would expect that $\epsilon(b)$ and $\epsilon(r)$ are either identical or highly correlated.

Avoiding IIA: Random coefficients models

- Suppose random utilities are given by

$$u(x, d, \tilde{\beta}) = x_d \tilde{\beta}$$

where x_d is a $K \times 1$ vector of attributes of alternative d and $\tilde{\beta} \sim N(\mu, \Sigma)$ are *random coefficients* representing different weights put on the attributes by different types of consumers.

- Let $\eta = \tilde{\beta} - \mu$ and we have $\eta \sim N(0, \Sigma)$ and we can write the model as an additively separable RUM as

$$u(x, d) + \epsilon(d) = x_d \mu + x_d \eta$$

where $\epsilon(d) = x_d \eta \sim N(0, x_d \Sigma x_d')$.

- Notice now that for two alternatives d and d' whose observed attributes become close, so that $x_d \to x_{d'}$, then their corresponding errors $\epsilon(d)$ and $\epsilon(d')$ become perfectly correlated as well

$$\text{var}(\epsilon(d) - \epsilon(d')) = (x_d - x_{d'})\Sigma(x_d - x_{d'})' \to 0 \quad \text{as } x_d \to x_{d'}$$

Social Surplus Function (EMAX or Smoothed Max)

- $E\left\{\max_{d\in D(x)}[u(x,d)+\epsilon(d)]\right\} = \int_\epsilon \max_{d\in D(x)}[u(x,d)+\epsilon(d)]q(\epsilon|x)$
- *Williams-Daly-Zachary Theorem*

$$\frac{\partial}{\partial u(x,d)} E\left\{\max_{d\in D(x)}[u(x,d)+\epsilon(d)]\right\} = P(d|x). \tag{6}$$

- Suppose $q(\epsilon|x)$ is a *multivariate Type 1 Extreme value distribution* with $\#D(x)$ components

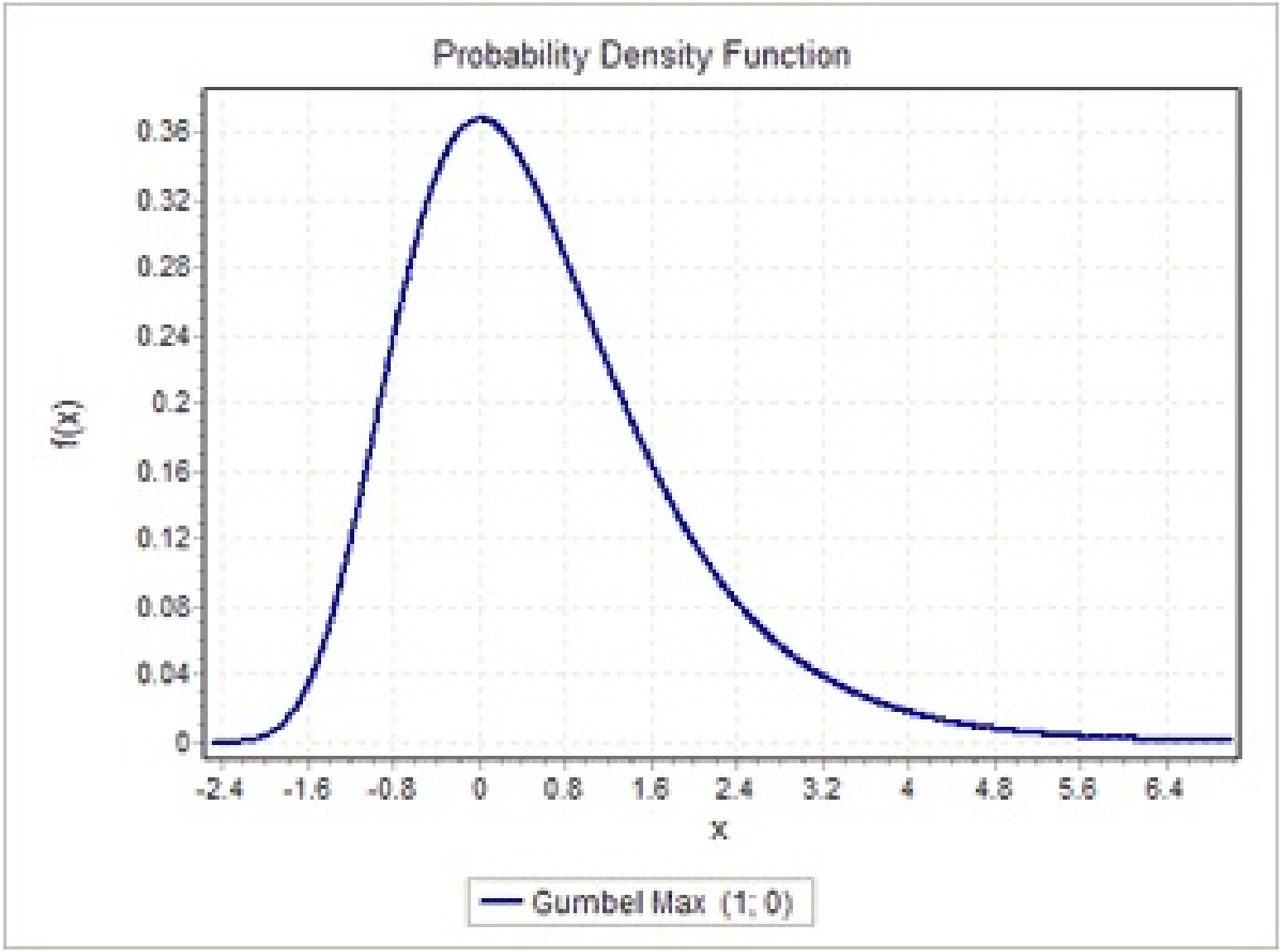
$$q(\epsilon|x) = \prod_{d\in D(x)} \exp\left\{-\exp\left\{-\frac{(\epsilon(d)-\mu(d))}{\sigma}\right\}\right\} \tag{7}$$

- Note: $\epsilon(d)$ and $\epsilon(d')$ are *independently distributed* for $d \neq d'$ and

$$\begin{aligned} E\{\epsilon(d)\} &= \mu(d)+\sigma\gamma \\ \text{var}\{\epsilon(d)\} &= \sigma^2\frac{\pi^2}{6} \end{aligned}$$

where $\gamma \simeq 0.57721$ is *Euler's constant*

Standardized Type 1 Extreme value probability density



Max-stability property of extreme value distributions

- Recall what a *stable family* of distributions is: *if $\tilde{X}$ and $\tilde{Y}$ are stable, then $\tilde{X} + \tilde{Y}$ is stable.* Example: Gaussian random variables are stable family since *sums of normal random variables are normal*
- The Type 1 extreme value family $\{\epsilon(d)|d \in D(x)\}$ is *max-stable* — the *maximum of Type 1 extreme value distributions also has a Type 1 extreme value distribution*
- Recall the distribution of a max of independent random variables:

$$\text{Prob}\left\{\max_{d\in D(x)}[\epsilon(d)] \leq x\right\} = \prod_{d\in D(x)} \text{Prob}\left\{\epsilon(d) \leq x\right\} \tag{8}$$

- For the Type 1 Extreme value family we have

$$\begin{aligned}\prod_{d\in D(x)} \text{Prob}\left\{\epsilon(d) \leq x\right\} &= \prod_{d\in D(x)} \exp\left\{-\exp\left\{-\frac{(\epsilon(d)-\mu(d))}{\sigma}\right\}\right\} \\ &= \exp\left\{-\exp\left\{-\frac{(x-I)}{\sigma}\right\}\right\}\end{aligned}$$

Max-stability of extreme value family, continued

- where I is the *inclusive value* or *log-sum* formula

$$I = \sigma \log \left(\sum_{d \in D(x)} \exp \left\{ \frac{\mu(d)}{\sigma} \right\} \right). \tag{9}$$

- Thus the Type 1 extreme value family is max-stable: *the maximum of Type 1 extreme value random variables is a Type 1 extreme value random variable.*
- Now apply the Williams-Daly-Zachary Theorem to get

$$\begin{aligned} P(d|x) &= \frac{\partial}{\partial u(x,d)} E \left\{ \max_{d \in D(x)} [u(x,d) + \epsilon(d)] \right\} \\ &= \frac{\partial}{\partial u(x,d)} \sigma \log \left(\sum_{d \in D(x)} \exp\{u(x,d)/\sigma\} \right) \\ &= \frac{\exp\{u(x,d)/\sigma\}}{\sum_{d' \in D(x)} \exp\{u(x,d')/\sigma\}} \end{aligned}$$

Econometric estimation of MNL models

- The MNL model has been used extensively in empirical work. If θ are unknown parameters of $u(x, d, \theta)$ and we observe data on individual states and choices, $\{(x_i, d_i)|i = 1, \ldots, N\}$, then McFadden showed how to estimate θ by *maximum likelihood*

$$\hat{\theta}_N = \underset{\theta}{\mathit{argmax}} \log(L_N(\theta)) = \log\left(\prod_{i=1}^{N} \frac{\exp\{u(x_i, d_i, \theta)/\sigma\}}{\sum_{d' \in D(x_i)} \exp\{u(x_i, d', \theta)/\sigma\}}\right)$$

- McFadden showed that $\log(L_N(\theta))$ is *globally concave in* θ when $u(x, d, \theta)$ is *linear in parameters* i.e. where $u(x, d, \theta) = w(x, d)'\theta$ where $w(x, d)$ and θ are $J \times 1$ vectors.
- However these days there is no reason to restrict $u(x, d, \theta)$ to be linear in parameters: modern "hill climbing" algorithms can also estimate $\hat{\theta}_N$ for more general specifications where $u(x, d, \theta)$ is not linear in parameters. The main complication is that $\log(L_N(\theta))$ is no longer necessarily concave in θ, creating the possibility of *multiple local optima.*

Brief review of Maximum Likelihood estimation

- Consider a general model for *IID* data for simplicity. The *model* is a condition density $f(x|\theta)$ that depends on unknown parameter θ^* to be estimated.
- Assume model is *correctly specified* that is the observed data $\{x_i\}_{i=1}^N$ are *IID* draws from $f(x|\theta^*)$, i.e. $x_i \sim f(x|\theta^*)$.
- Under standard smoothness and interiority regularity conditions ($f(x|\theta)$ is twice differentiable, has bounded expectation, parameter space compact subset of Euclidean space and θ^* is in the interior of this set, blah blah bhah) then we have the *asymptotic normality and efficiency of maximum likelihood*

$$\sqrt{N}[\hat{\theta}_N - \theta^*] \Longrightarrow N(0, I^{-1}(\theta^*)) \tag{10}$$

where $I(\theta^*)$ is the *Information Matrix* given by

$$I(\theta^*) = E\left\{ \frac{\partial}{\partial \theta} \log(f(\tilde{x}|\theta^*)) \frac{\partial}{\partial \theta'} \log(f(\tilde{x}|\theta^*)) \right\}. \tag{11}$$

The Information Equality

- Let $H(\theta^*)$ be the expected hessian, i.e expectation of $\log(f(\tilde{x}|\theta))$, i.e.

$$H(\theta^*) = E\left\{\frac{\partial^2}{\partial\theta\partial\theta'}\log(f(\tilde{x}|\theta^*))\right\}. \tag{12}$$

- The *Information Equality* is that for any true θ^* we have

$$I(\theta^*) = -H(\theta^*). \tag{13}$$

- Differentiate the key *orthogonality condition* of maximum likelihood, namely that the "expectation of the score function is zero" identity to get

$$\begin{aligned} 0 &= \nabla E\{\nabla \log(f(\tilde{x}|\theta)|\theta\} = \frac{\partial}{\partial\theta}\int_x \frac{\partial}{\partial\theta'}\log(f(x|\theta)f(x|0)dx \\ &= \int_x \frac{\partial^2}{\partial\theta\partial\theta'}\log(f(x|\theta))f(x|\theta)dx + \int_x \frac{\partial}{\partial\theta}\log(f(x|\theta))\frac{\partial}{\partial\theta'}f(x|\theta)dx \\ &= H(\theta) + I(\theta). \end{aligned}$$

Computation of MLEs

- We use *quasi Newton algorithms* for smooth problems. These are all based on *Newton's Method* for solving the system of first order conditions for a maximum of the log-likelihood function

$$\nabla \log L_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial \theta} \log(f(x_i|\theta)) = 0 \quad \text{at } \theta = \hat{\theta}_N. \tag{14}$$

- Using Newton's method to solve these first order conditions we get iterations of the form

$$\theta_{t+1} = \theta_t - \left[\nabla^2 L_N(\theta_t)\right]^{-1} \nabla L_N(\theta_t), \tag{15}$$

where $H_N(\theta) = \nabla^2 L_N(\theta)$ is the *Hessian matrix* of log-likeihood function $L_N(\theta)$ given by

$$H_N(\theta) = \nabla^2 L_N(\theta) = \frac{\partial^2}{\partial \theta \partial \theta'} L_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2}{\partial \theta \partial \theta'} \log(f(x_i|\theta)). \tag{16}$$

BHHH

- BHHH is the acronym for a 1974 paper by Berndt, Hall, Hall and Hausman the key idea of which is to use the Information Equality and replace the *empirical hessian* $H_N(\hat{\theta}_N)$ by the (negative of) the *empirical information matrix* $I_N(\hat{\theta}_N)$ given by

$$I_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \theta} \log(f(x_i|\theta)) \frac{\partial}{\partial \theta'} \log(f(x_i|\theta)). \tag{17}$$

- Because $I_N(\theta)$ is always *positive semidefinite* the BHHH iteration *always results in an improvement in the likelihood* at least for a sufficiently small scalar *step size* $\lambda_t > 0$ chosen to approximately maximize the likelihood along the line of search, i.e.

$$\theta_{t+1} = \theta_t - \lambda_t \left[-I_N(\theta_t)\right]^{-1} \nabla L_N(\theta_t), \tag{18}$$

  i.e. we have $L_N(\theta_{t+1}) \geq L_N(\theta_t)$ using only first derivatives.

Bayesian estimation of the MNL model

- Recall that in Bayesian inference we start with a *prior* $\pi(\theta)$ representing our subjective initial beliefs about plausible or "most likely" parameter values prior to observing any data.
- Then we observe $\{x_i, d_i\}_{i=1}^N$ *IID* observations on the states/choices of N different decision makers. Bayesian learning is the process of learning from the data, going from *prior beliefs* $\pi(\theta)$ about θ to the *posterior beliefs* $\pi(\theta|\{x_i, d_i\}_{i=1}^N)$ given by

$$\pi(\theta|\{x_i, d_i\}_{i=1}^N) = \frac{\prod_{i=1}^N P(d_i|x_i, \theta)\pi(\theta)}{\int_{\theta'} \prod_{i=1}^N P(d_i|x_i, \theta')\pi(\theta')d\theta'}. \tag{19}$$

  where $P(d|x, \theta)$ is just the MNL formula derived above.
- Bayesian estimation is elegantly simple and *does not need to rely on asymptotic approximations such as the Central Limit Theorem* to get an *exact finite sample distribution for the Bayesian estimator.* The exact finite sample distribution is just the posterior, $\pi(\theta|\{x_i, d_i\}_{i=1}^N)$.

Mother logit, CCPs and Hotz-Miller Inversion

- *Conditional choice probability* (CCP) $P(d|x)$, is the probability an individual with characteristics x chooses alternative $d \in D(x)$.
- The CCP may or may not be consistent with a "rational" underlying model of static discrete choice, i.e. RUM.
- Is it possible to find some utility function to "rationalize" any CCP $P(d|x)$? *Yes!* define $u(x,d) = \log(P(d|x))$. Then for the MNL model we have

$$\begin{aligned} P(d|x) &= \frac{\exp\{u(x,d)\}}{\sum_{d' \in D(x)} \exp\{u(x,d')\}} = \frac{\exp\{\log(P(d|x))\}}{\sum_{d' \in D(x)} \exp\{\log(P(d'|x))\}} \\ &= \frac{P(d|x)}{\sum_{d' \in D(x)} P(d'|x)} = P(d|x). \end{aligned}$$

- **Hotz-Miller Inversion Theorem** *There is a 1 to 1 mapping between CCPs and normalized utility functions*

$$\{P(d|x) | d \in D(x)\} \longleftrightarrow \{u(x,d) - u(x,0) | d \in D(x)\}$$

where $0 \in D(x)$ is some fixed element in the choice set.

Implications for identification

- Example of Hotz-Miller inversion for MNL model

$$\log(P(d|x)/P(0|x)) = u(x,d) - u(x,0) \tag{20}$$

We often normalize the utility of one of the alternatives to an aribtrary value, e.g. $u(x,0) = 0$ since any monotonic transformation of utilities are observationally equivalent, i.e for any constant k and scalar $\lambda > 0$ we have

$$u(x,d) + \epsilon(d) \geq u(x,d') + \epsilon(d') \Longleftrightarrow$$
$$\lambda[u(x,d) + \epsilon(d) + a] \geq \lambda[u(x,d') + \epsilon(d') + a]$$

- Therefore we typically make *location and scale normalizations* such as $a = 0$ and $\lambda = 1$ unless $u(x,d)$ is a payoff we can potentially observe, e.g. profits.
- But even with the normalizations, the idea behind the Mother Logit and the Hotz-Miller inversion is the following **Non-parametric non-identification of the RUM** *If we are unwilling to put any restrictions on preferences beyond location and scale normalizations, then we can always "rationalize" any conditional choice probabilities.*

Recap on identification

- The CCP $\{P(d|x)|d \in D(x), x \in X\}$ is the *reduced-form object* — it can be estimated regardless of whether the underlying theory of RUM holds, and so with sufficient data we can estimate $P(d|x)$ nonparametrically under very weak assumptions and hence treat it as "known"
- The utility function $\{u(x,d)|d \in D(x), x \in X\}$ and the distribution of RUM components $\{q(\epsilon|x)|x \in X\}$ are the *structural objects.*
- The RUM is identified if there is a 1 to 1 mapping between the structure and reduced form

$$\{P(d|x)|d \in D(x), x \in X\} \longleftrightarrow \{[u(x,d), q(\epsilon|x)]|d \in D(x), x \in X\}$$

- Unfortunately the Mother Logit result, or more generally the Hotz-Miller Inversion Theorem tell us that the discrete choice model is *non-parametrically unidentified* — i.e. without any further restrictions, for any distribution $q(\epsilon|x)$ of random utility components, we can find a utility function $\{u(x,d)|d \in D(x), x \in X\}$ that rationalizes the CCPs.

Aside: proof of the Hotz-Miller Inversion Theorem

- We have shown above, for the logit model that we can invert the MNL choice probabilities (via taking logs of the choice probability ratios) to get the *utility differences* $\Delta u(x,d) \equiv u(x,d) - u(x,0)$ where we can interpret choice $d = 0$ as the "outside good".
- Note that the if $|D(x)|$ is the number of elements in the choice set, then there are only $|D(x)| - 1$ "free choice probabilities" since the choice probabilities sum to 1. These $|D(x)| - 1$ "free choice probabilities" can be inverted into $|D(x)| - 1$ "utility differences" $\Delta u(x,d)$, $d \in D(x)$ but excluding the outside good $d = 0$ whose utility difference is 0, $\Delta u(x,0) = u(x,0) - u(x,0) = 0$.
- Is this inversion specific to the logit specification, or will it hold for a multinomial probit model or some other distribution for the error terms $\{\epsilon(d) | d \in D(x)\}$ in the discrete choice model?
- The *Hotz-Miller Inversion Theorem* says that in pretty great generality, the answer is yes: *we can invert conditional choice probabilities* $\{P(d|x),\ d \in D(x),\ d \neq 0\}$ *to obtain* $|D(x)| - 1$ *utility function differences* $\{\Delta u(x,d),\ d \in D(x),\ d \neq 0\}$ *for essentially any discrete choice model.*

The Inverse Function Theorem

- Their result is an application of a basic result in functional analysis called the *Inverse Function Theorem*
- **Inverse Function Theorem** *Let $F : R^n \to R^n$ be a differentiable mapping and let p be an interior point of the domain of F. If $F(p) = q$ and the $n \times n$ Jacobian matrix $\nabla F(p)$ is non-singular at p, then there is an open neighborhood of q such that the inverse mapping $F^{-1}(q) = p$ exists and is continuously differentiable, and the Jacobian matrix of the inverse mapping at q exists and is given by $\nabla F^{-1}(q) = [\nabla F(p)]^{-1}$.*
- So to apply this to prove the Hotz-Miller theorem, we identify the point p with the utility differences $\Delta u(x, d)$, so then $n = |D(x)| - 1$. We can write $F(p) = \vec{P}(\{\Delta u(x, d)\})$, i.e. the conditional choice probabilities map the utility differences into the vector of choice probabilities (excluding the probability of the outside good), so $\vec{P}$ denotes a $|D(x)| - 1 \times 1$ vector with elements $[P(1|x), P(2|x), \ldots, P(|D(x)| - 1|x)]'$ where we have indexed the choices by integers, $\{1, 2, \ldots, |D(x)| - 1\}$ but excluded the "outside good" choice, $d = 0$.

Now use IFT to prove Hotz-Miller Inversion Theorem

- To complete the proof using the Inverse Function Theorem, we need to show:
- 1) the conditional choice probability mapping $\vec{P}(\{\Delta u(x,d)\})$ is a continuously differentiable function of p, the utility differences $\{\Delta u(x,d)\}$ arranged as a $|D(x)|-1 \times 1$ vector,
- 2) the Jacobian of this mapping is an invertible $(|D(x)|-1) \times (|D(x)|-1)$ matrix.
- Our strategy for doing this is to use the *social surplus function* $S(\{\Delta u(x,d)\})$ expressed as a function of the utility differences:

$$S(\{\Delta u(x,d)\}) \equiv E\left\{\max\left[\epsilon(0), \max_{d\in\{1,\ldots,|D(x)|-1\}}[\Delta u(x,d)+\epsilon(d)]\right]\right\}$$
$$= \int_{\epsilon(0)} \cdots \int_{\epsilon(|D(x)-1)} \max\left[\epsilon(0), \max_{d\in\{1,\ldots,|D(x)|-1\}}[\Delta u(x,d)+\epsilon(d)]\right]$$
$$\times dq(\epsilon(0),\epsilon(1),\ldots,\epsilon(|D(x)|-1)|x). \quad (21)$$

Use key properties from convex duality theory

- The Social Surplus function is akin to a *support function* in convex duality theory. Notice

$$S(\{\Delta u(x,d)\}) + u(x,0) = S(\{u(x,d)|\ d \in D(x)\}), \tag{22}$$

where $S(\{u(x,d)|\ d \in D(x)\})$ is the usual social surplus or EMAX function given by

$$S(\{u(x,d)|\ d \in D(x)\}) = E\left\{\max_{d \in \{0,1,\ldots,|D(x)|-1\}} [u(x,d) + \epsilon(d)]\right\} \tag{23}$$

i.e. the Social Surplus function as a function of all the $|D(x)|$ utilities $\{u(x,d)\}$ instead of the utility differences $\{\Delta u(x,d)\}$.

- The strategy is to show that $S(\{\Delta u(x,d)\})$ is a strictly convex function of the utility differences and then use the *Williams Daly Zachary Theorem.*

Showing that EMAX is strictly convex

- Let n be the number of choices, $x \in R^n$ is the vector of utility differences (setting choice $d = 0$ as the normalizing choice), and fix $\epsilon \equiv (\epsilon(0), \epsilon(1), \ldots, \epsilon(n))$. Then define the function $M(x)$ as follows

$$M(x, \epsilon) = \max \left[\sigma\epsilon(0), \max_{d \in \{1,\ldots,n\}} [x(d) + \sigma\epsilon(d)] \right] \tag{24}$$

and note that when we let ϵ have a conditional distribution $q(\epsilon|x)$, then the Social Surplus function $S(x)$ is the expectation of $M(x, \epsilon)$ with respect to $q(\epsilon|x)$

$$S(x) = E\{M(x, \epsilon)\} = \int_\epsilon M(x, \epsilon) dq(\epsilon|x). \tag{25}$$

So we first show that, holding ϵ fixed, $M(x, \epsilon)$ is a convex function of x, i.e. for any $x, y \in R^n$ and $\theta \in [0, 1]$ we have

$$\theta M(x, \epsilon) + (1 - \theta) M(y, \epsilon) \geq M(\theta x + (1 - \theta) y, \epsilon). \tag{26}$$

Showing that EMAX is convex

- It is not hard to show that $M(\theta x + (1-\theta)y, \epsilon) = \max\left[\sigma\epsilon(0), \theta x(d) + (1-\theta)y(d) + \sigma\epsilon(d)\right]$ for some $d \in \{1, \ldots, n\}$. But we also have

$$\begin{aligned} M(x, \epsilon) &\geq \max[\sigma\epsilon(0), x(d) + \sigma\epsilon(d)] \\ M(y, \epsilon) &\geq \max[\sigma\epsilon(0), y(d) + \sigma\epsilon(d)], \end{aligned} \tag{27}$$

  but this implies convexity, since by using the inequalities above we have

$$\theta M(x, \epsilon) + (1-\theta)M(y, \epsilon) \geq M(\theta x + (1-\theta)y, \epsilon). \tag{28}$$

- Integrating with respect to $q(\epsilon|x)$ on both sides of inequality (28), we obtain

$$\theta S(x) + (1-\theta)S(y) \geq S(\theta x + (1-\theta)y). \tag{29}$$

- But we want to show *strict convexity* of $S(x) = E\{M(x, \epsilon)\}$.

Showing that EMAX is *strictly convex*

- This is harder to show, but roughly, if the distribution of unobserved ϵ shocks is 1) continuously distributed, and 2) has unbounded support, then the corresponding social surplus, smoothed max, or EMAX function $S(x)$ will be strictly convex in x, i.e. for any $x \neq y$ and $\theta \in (0,1)$ we have

$$\theta S(x) + (1-\theta)S(y) > S(\theta x + (1-\theta)y). \tag{30}$$

  In terms of our original notation, $S(\{\Delta u(x,d) | d \in \{1,\ldots,|D(x)|\})$ will be strictly convex. Since strictly convex functions are "almost everywhere twice continuously differentiable" with invertible hessian (except at a countable set of points), we now have the key ingredients to prove the Hotz-Miller Inversion Theorem.

Finish the proof of the Hotz-Miller Inversion Theorem

- To see that the Hotz-Miller Inversion Theorem is a special case of the Inverse Function Theorem, first apply the Williams-Daly-Zachary Theorem to get

$$\vec{P}(\{\Delta u(x,d)\}) = \nabla S\left(\{\Delta u(x,d)\}\right). \tag{31}$$

- Next, differentiate both sides, i.e. take Jacobians on both sides of equation (31) to get

$$\nabla \vec{P}(\{\Delta u(x,d)\}) = \nabla^2 S((\{\Delta u(x,d)\}). \tag{32}$$

- Notice the right hand side of equation (32) is the second derivative matrix of Hessian of the Social Surplus/EMAX function, which is invertible for almost all values of $\{\Delta u(x,d)\}$ due to strict convexity. Hence the Jacobian of the choice probability is invertible, so by the Inverse Function Theorem it follows that we can (locally) invert the CCPs to obtain $\{\Delta u(x,d)|d \in D(x)\}$.

Tying up some loose ends

- In summary, there is (locally, i.e. at any fixed x for almost any utility values $\{u(x,d)|d \in D(x)\}$) a 1 to 1 mapping between the vector of $|D(x)|-1$ choice probabilities that excludes the "normalizing choice" $d=0$ that we used to define $\Delta u(x,d) \equiv u(x,d) - u(x,0)$,

$$\left\{\vec{P}(d|x,\{\Delta u(x,d)\})|d \in \{1,\ldots,|D(x)|-1\}\right\}, \tag{33}$$

  and the corresponding vector of normalized utilities, with $|D(x)|-1$ components

$$\{\Delta u(x,d)|d \in \{1,\ldots,|D(x)|-1\}\}. \tag{34}$$

- What we still have not proved, though, is the *strict convexity property* of the Social Surplus function S when the unobserved utility shocks $\{\epsilon(d)|d \in D(x)\}$ have a continuous density with full support on $R^{|D(x)|}$. Showing this is harder, but the next slides show it does hold.
- Can you come up with an analytical proof?

Expectations of convex functions can be strictly convex

- Recall the function

$$M(x,\epsilon) = \max\left[\sigma\epsilon(0), \max_{d\in\{1,\ldots,n\}}[x(d)+\sigma\epsilon(d)]\right] \tag{35}$$

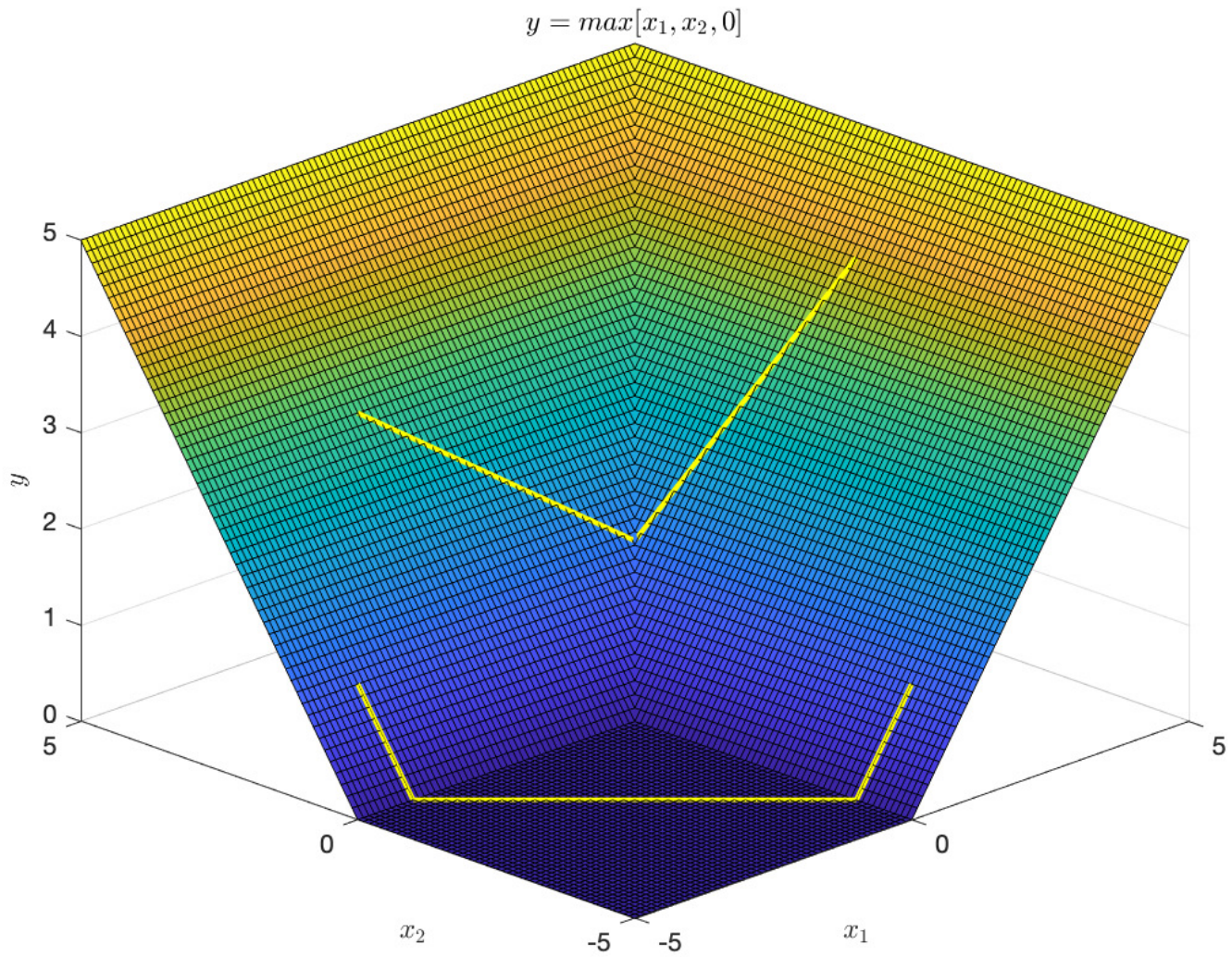
that was the key to establishing our result. For any *fixed* ϵ vector, this function is only convex in x and not strictly convex in x as we showed in inequality (26) above. The next slide plots this function for both $\epsilon = 0$ and a randomly drawn ϵ and we can verify that in either case $M(x,\epsilon)$ is convex but not strictly convex.

- But when we integrate with respect to the continuous distribution $q(\epsilon|x)$ to get $S(x)$, the resulting smoothed max function is indeed strictly convex. For example if ϵ has a Type 1 Extreme value distribution, recall our analytical expression for $S(x)$. You can use it to show $S(x)$ is strictly convex

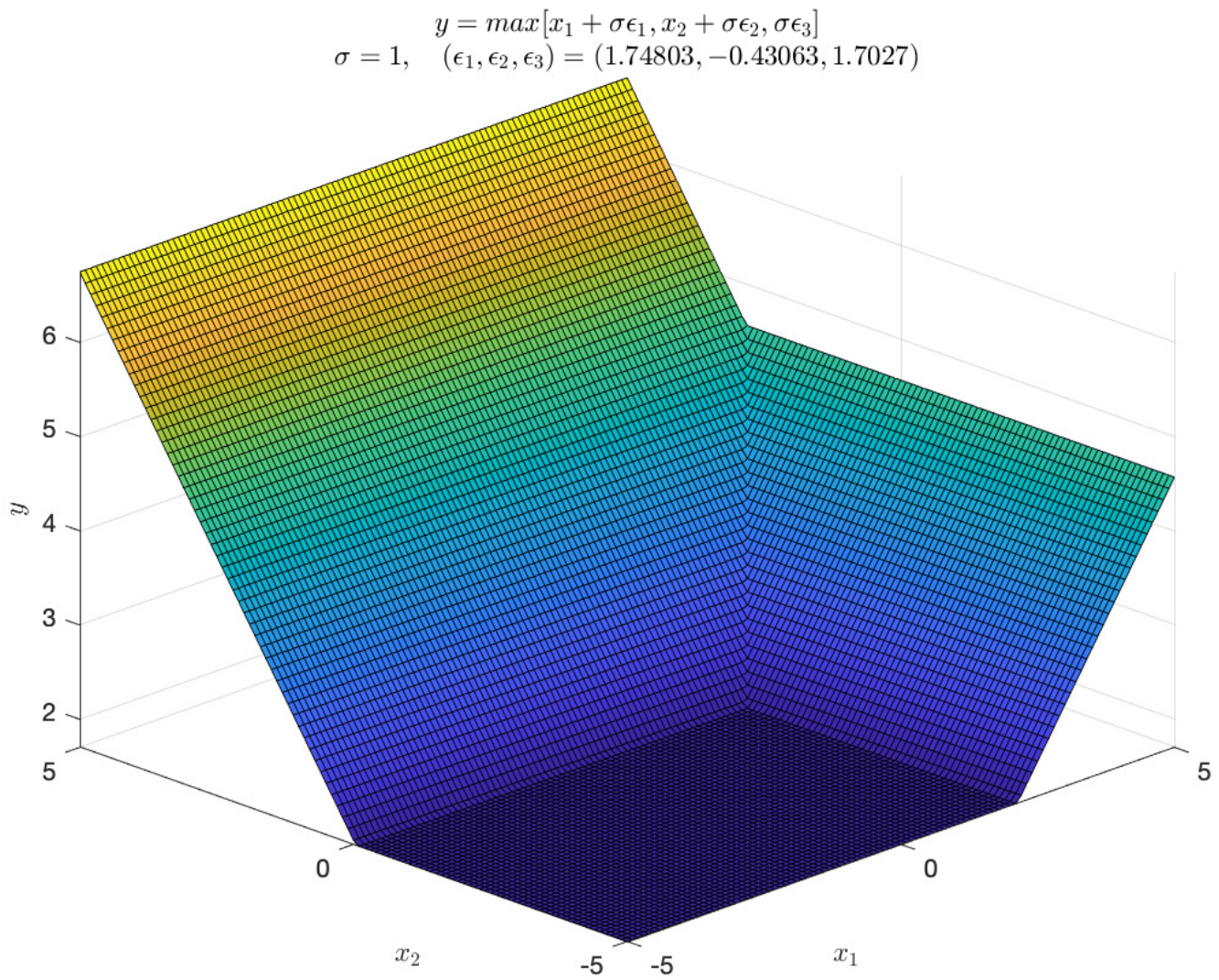
$$S(x) = S(x_1,\ldots,x_n) = \sigma\log\left(1+\sum_{d=1}^{n}\exp\{x_d/\sigma\}\right). \tag{36}$$

But what other distributions? Such as if ϵ is multivariate normal?

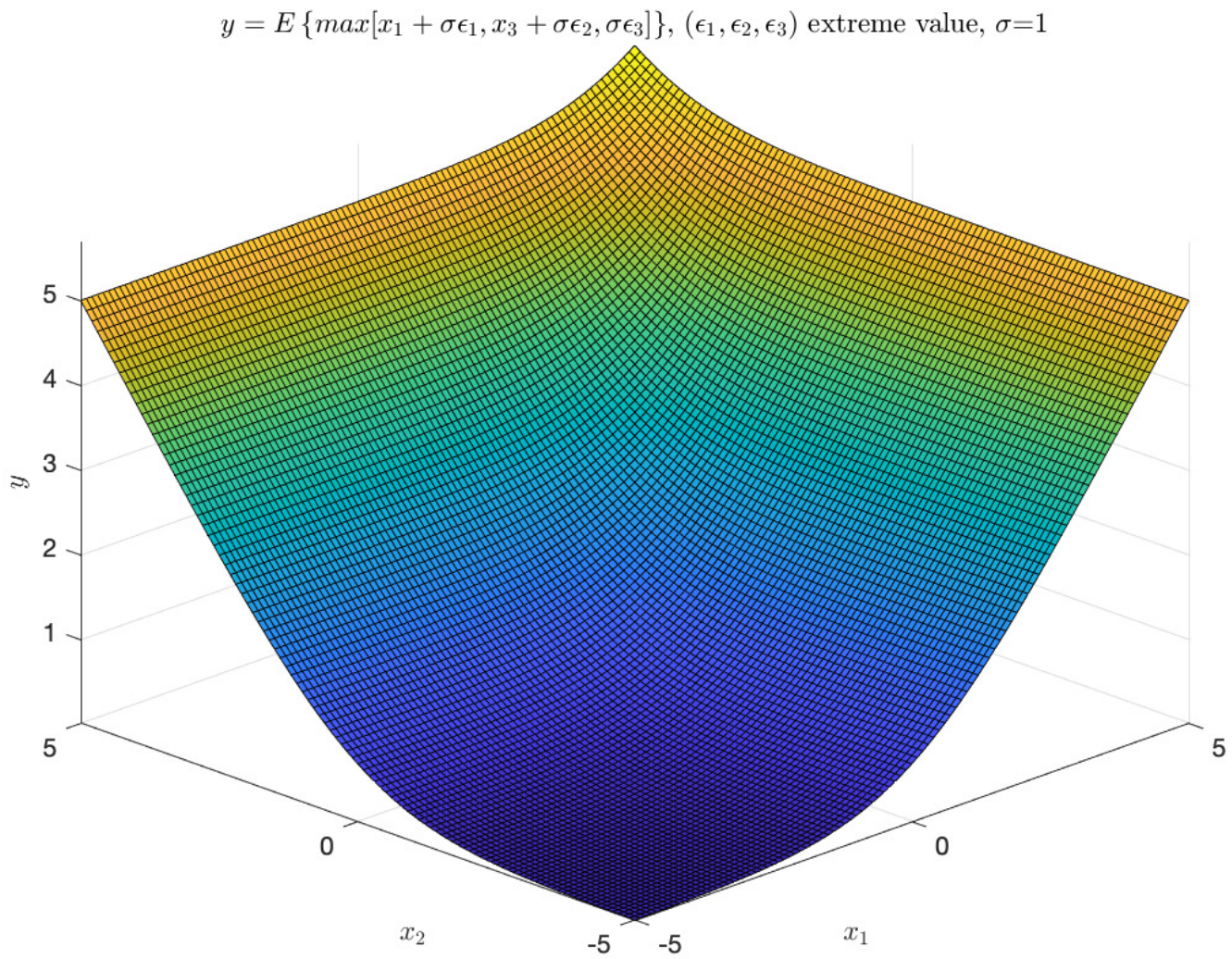
A plot of $M(x, \epsilon)$ when $\epsilon = (0, 0, 0)$



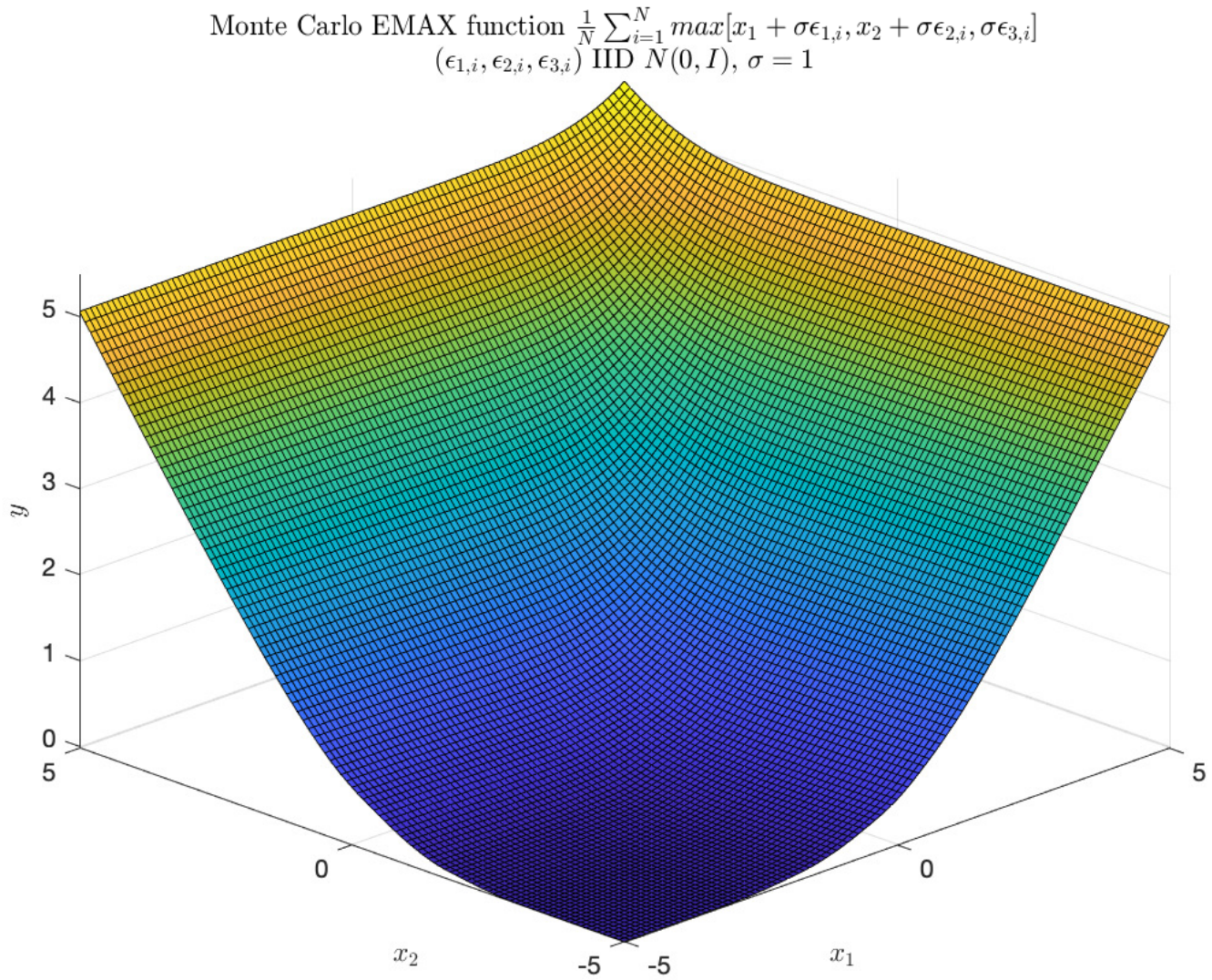
A plot of $M(x, \epsilon)$ for a randomly drawn ϵ



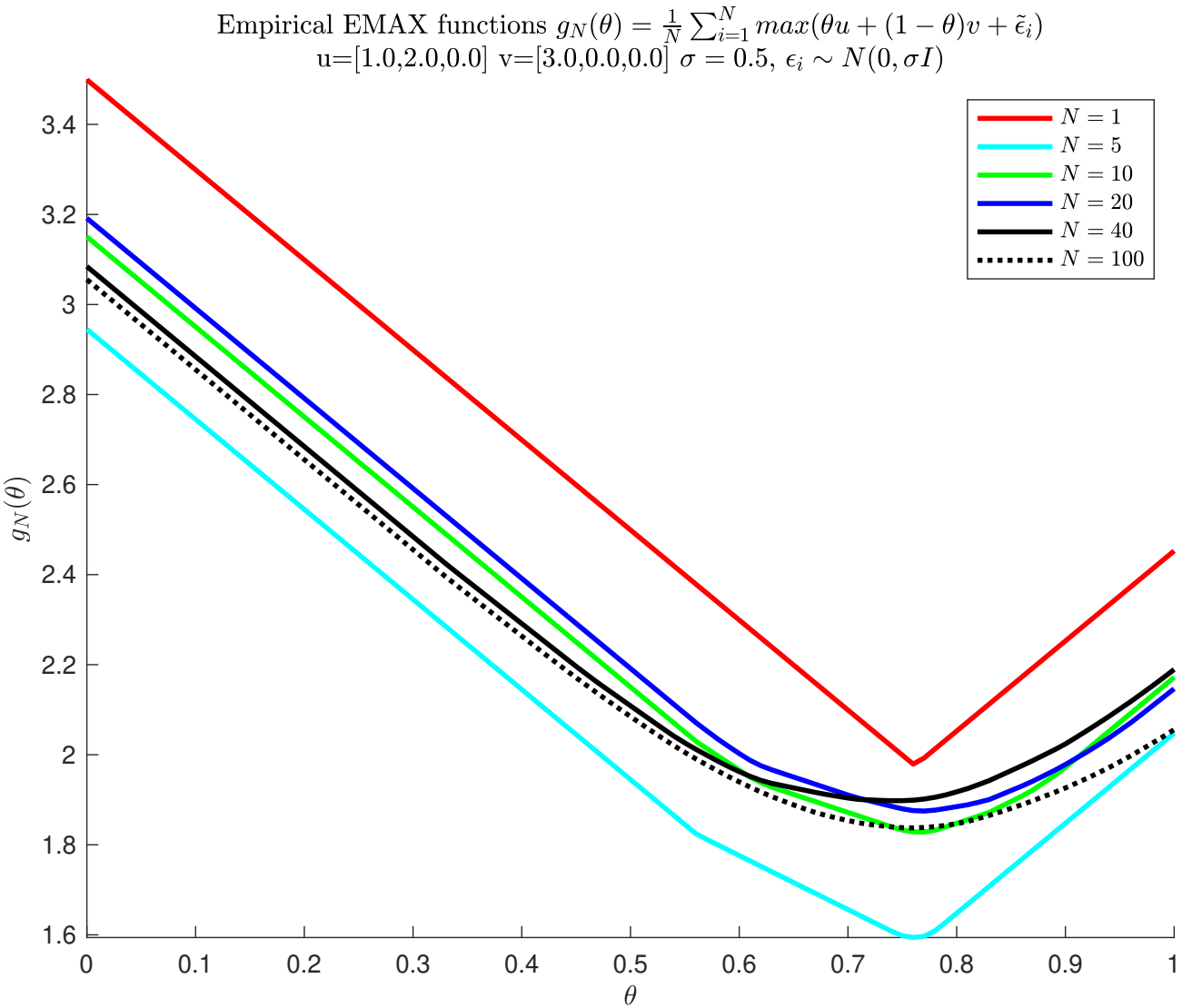
A plot of $S(x) = E\{M(x, \epsilon)\}$ for extreme value ϵ



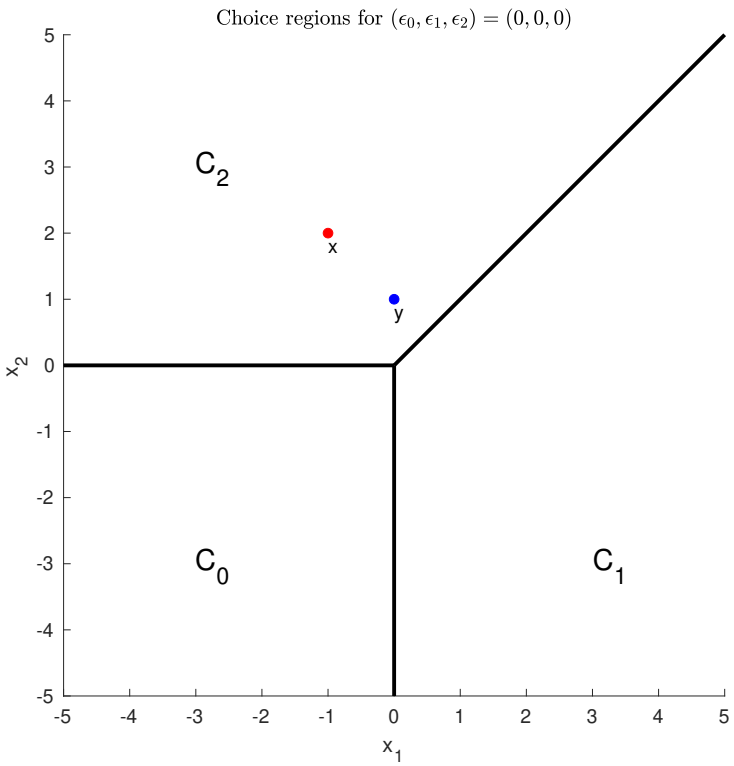
Monte Carlo EMAX, $N = 100$ *IID* $N(0, \sigma^2)$ ϵ draws

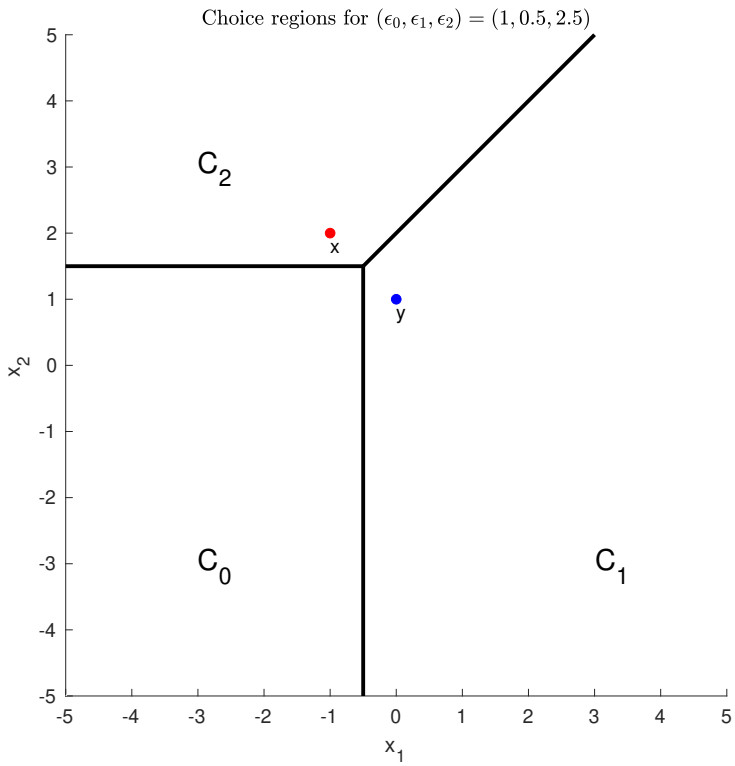


How increasing N affects convexity and smoothness



How choice shocks can "separate" vectors x and y





Insights from the graphs

- The graph of $\max[x_1, x_2, 0]$ decomposes into 3 planes and each plane corresponds to a different *choice region* (C_0, C_1, C_2) in (x_1, x_2) space. In region C_1, $\max(x_1, x_2, 0) = x_1$ so choice 1 is optimal. In region C_2 alternative 2 is optimal and $\max[x_1, x_2, 0] = x_2$. In region C_0 the outside option is optimal so $\max[x_1, x_2, 0] = 0$. We plot these choice regions above.
- Since $M(x, \epsilon) = \max[x_1, x_2, 0]$ when $\epsilon = (\epsilon_0, \epsilon_1, \epsilon_2) = (0, 0, 0)$ we see that strict convexity holds if x and y are in different choice regions, since then $\theta M(x, 0) + (1 - \theta)M(y, 0) > M(\theta x + (1 - \theta)y, 0)$ for all $\theta \in (0, 1)$.
- But if x and y are in the same choice region, since $M(x, 0)$ is linear in each subregion, then strict convexity does not hold: $\theta M(x, 0) + (1 - \theta)M(y, 0) = M(\theta x + (1 - \theta)y, 0)$.
- The right hand panel of the graph shows that given any two points x and y, $x \neq y$ there are infinitely many"perturbations" ϵ that *separate* x and y, i.e. these two points lie in different choice regions of the perturbed max function $M(x, \epsilon)$.

Using the insights to prove strict convexity of $S(x)$

- The proof involves two steps: 1) show that if ϵ has a continuous density with unbounded support, then for any $x \neq y$ there is always a positive probability of ϵ shocks that separate x and y, i.e. that cause them to lie in different choice regions of the perturbed max function $M(x, \epsilon)$, 2) this implies that $S(x) = E\{M(x, \epsilon)\}$ is a weighted average of linear functions $M(x, \epsilon)$ (for values of ϵ where x and y are in the same choice region) and strictly convex functions $M(x, \epsilon)$ (for values of ϵ where x and y are in different choice regions). Hence $S(x)$ is strictly convex in x.
- We will prove the first step for the general case of n alternatives. Suppose $x \neq y$ and both x and y are vectors in R^n. Let $\delta(x, 0)$ be the optimal choice, i.e. the index $\delta(x, 0) = j$ such that $x_j = \max(x_1, \ldots, x_n) = \max(x)$, and similarly let $\delta(y, 0)$ be the index such that $y_{\delta(y,0)} = \max(y)$. Assume that $\delta(x, 0) = \delta(y, 0)$ so that x and y are in the same choice region of $M(x, 0)$.

Showing that ϵ perturbations can separate any $x \neq y$

- We now show that there is an open set $O(x, y) \subset R^n$ of "perturbations" ϵ such that for each $\epsilon \in O(x, y)$ we have $\delta(x, \epsilon) \neq \delta(y, \epsilon)$, i.e. x and y are "separated" and lie in different choice regions of the perturbed max function $M(x, \epsilon)$.
- There are two cases to consider: 1) x is a "uniform shift" of y so that for some scalar λ we have $x = y + \lambda$, and 2) $x \neq y + \lambda$ for any scalar λ.
- If x is a uniform shift of y, then without loss of generality suppose that if $\lambda > 0$ and let $\epsilon = -(x + \lambda/2)$. Notice that $x + \epsilon < 0$, i.e. it is a vector whose components are all negative and equal to $-\lambda/2$ and hence $x + \epsilon$ lies in the region C_0 where the outside good is chosen. But $y + \epsilon = x + \lambda - (x + \lambda/2)$ is a vector whose elements all equal $\lambda/2$ and hence does not lie in region C_0. It follows this ϵ separates x and y. Clearly a continuum of other choices of ϵ separate x and y this way. For example any $\epsilon = -(x + \gamma)$ where γ is any positive scalar satisfying $\gamma < \lambda$ will also do.

Showing that ϵ perturbations can separate any $x \neq y$

- Now consider the case where x is not a shifted version of y. Without loss of generality, suppose there are no ties in the rankings of the components of x and y, so that $x_{\delta(x,0)} > \max_{d \neq \delta(x,0)} x_d$ and $y_{\delta(y,0)} > \max_{d \neq \delta(y,0)} y_d$. Let $\delta_2(x,0)$ be the index of the second highest component of x, so $x_{\delta(x,0)} > x_{\delta_2(x,0)}$ and similarly let $\delta_2(y,0)$ be the index of the second highest component of y.
- Suppose, again without loss of generality, that $x_{\delta(x,0)} - x_{\delta_2(x,0)} > y_{\delta(y,0)} - y_{\delta_2(y,0)}$ and let ϵ be any vector all whose components are zero except for the common maximal component of x and y, $\delta(x,0) = \delta(y,0)$, (which are the same by the assumption that x and y are in the same choice region of $M(x,0)$). Let this component, $\epsilon_{\delta(x,0)} = -\lambda$, were λ is any scalar satisfying the inequality $x_{\delta(x,0)} - x_{\delta_2(x,0)} > \lambda > y_{\delta(y,0)} - y_{\delta_2(y,0)}$. Then is is easy to see that $x + \epsilon$ continues to be maximized at alternative $\delta(x,0)$, i.e. $\delta(x,\epsilon) = \delta(x,0)$. However the maximizer of $y + \epsilon$ is $\delta_2(y,0)$ since $y_{\delta_2(y,0)} - \lambda > y_{\delta(y,0)}$. It follows that $\delta(x,\epsilon) \neq \delta(y,\epsilon) = \delta_2(y,0)$, and hence there is an open set of perturbations ϵ that separate x and y.

Empirical EMAX implies locally flat choice probabilities

- Though we have shown that even finite mixtures of perturbed max functions will be strictly convex for "many" pairs of point (x, y), only be in the limit will the Social Surplus function $S(x) = E\{\max(x, \tilde{\epsilon})\}$ be strictly convex *everywhere* i.e. for *all* pairs (x, y).
- Consider the *empirical EMAX (or Social Surplus) function* $S_N(x)$ resulting from drawing N *IID* perturbation vectors $(\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_N)$ from a distribution with a continuous density and full support over R^n (where small n is the number of choices, so each $\tilde{\epsilon}_i \in R^n$). Then $S_N(x)$ is the (random) function given by

$$S_N(x) = \frac{1}{N} \sum_{i=1}^{N} M(x, \tilde{\epsilon}_i). \tag{37}$$

- Note that for any two points x and y which are not separated by the N random draws $(\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_N)$ they will be in the same choice region and hence the second derivative of $S_N(x)$ at either of these points will be all *zero.*
- This implies that the choice probability will be *flat* in sufficiently small neighborhoods of points x where $S_N(x)$ is *linear in* x.

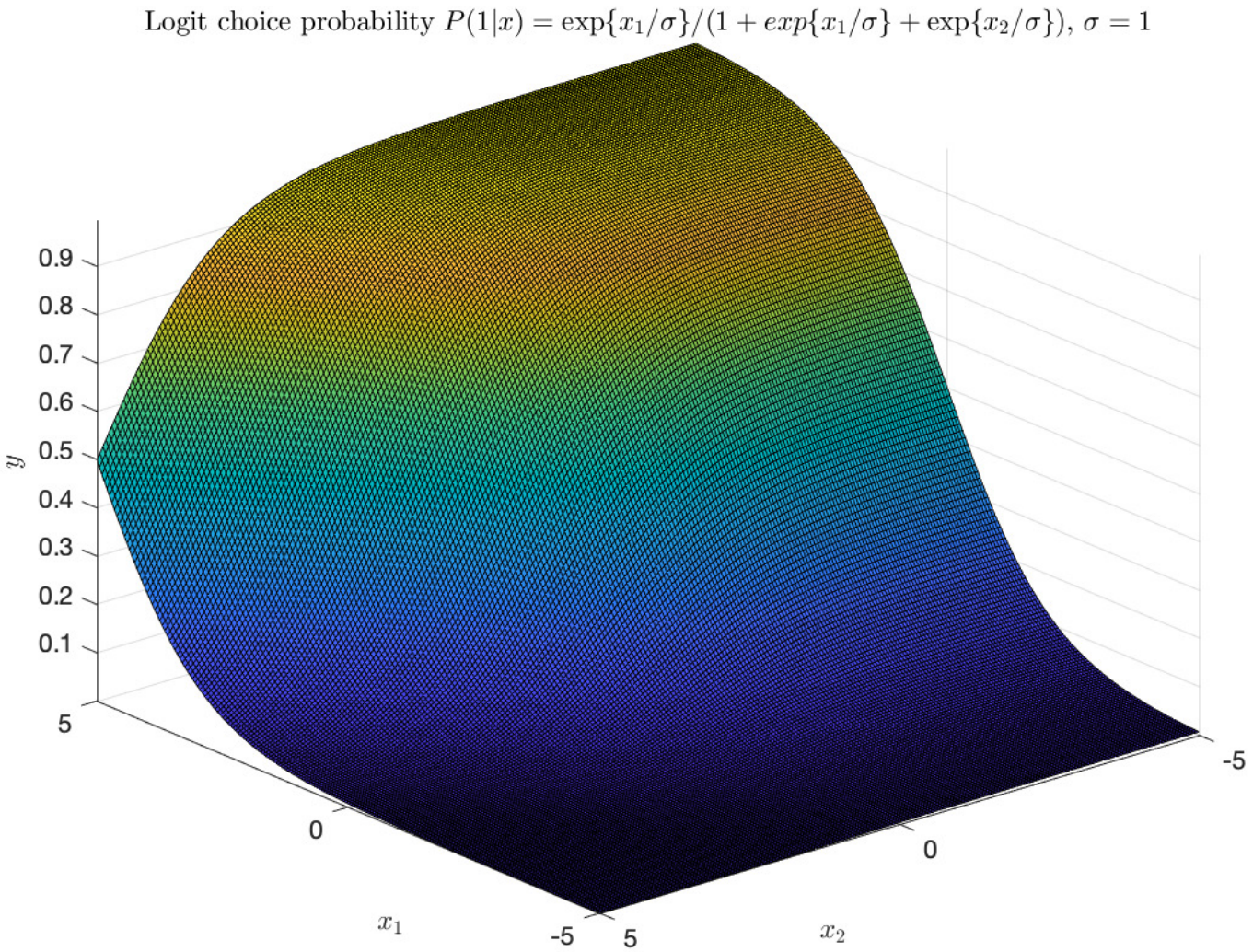
What if we differentiate the empirical EMAX function?

- Similar to how an empirical CDF $F_N(x)$ converges uniformly to the true limiting CDF $F(x)$, the empirical EMAX function $S_N(x)$ will converge uniformly to the limiting true EMAX function $S(x)$ using *empirical process theory* (i.e. a uniform (over x) Law of Large Number).
- Since $\max(x+\epsilon)$ is differentiable for almost all x, we can apply the Williams-Daly-Zachary Theorem to $S_N(x)$ to get

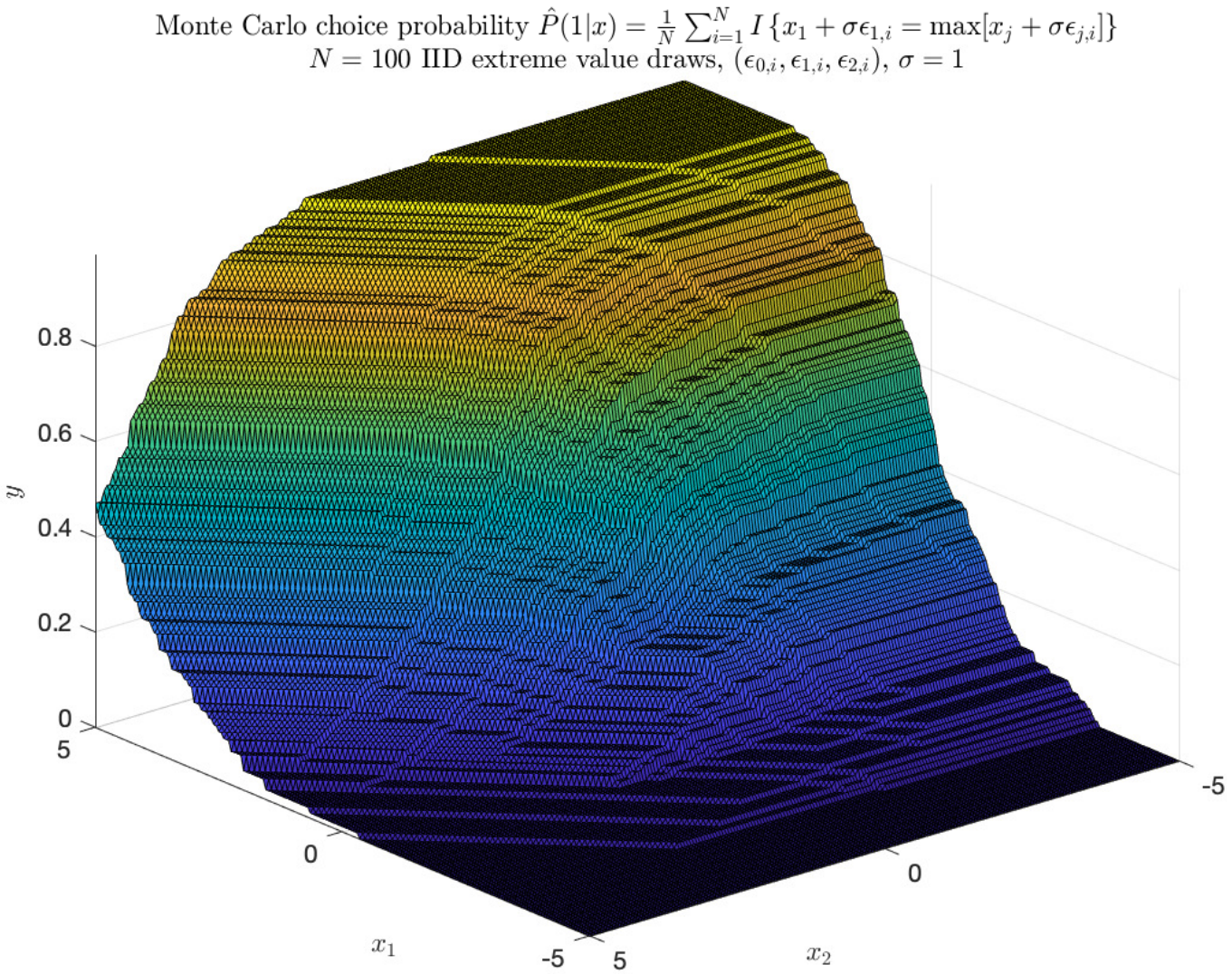
$$\begin{aligned}
\frac{\partial}{\partial x_j} S_N(x) &= \frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial x_j} \max_{d\in\{1,\ldots,n\}} [x_d + \tilde{\epsilon}_{i,d}] \\
&= \frac{1}{N}\sum_{i=1}^{N} I\left\{ x_j + \tilde{\epsilon}_{i,j} \geq \max_k [x_k + \tilde{\epsilon}_{i,k}] \right\} \\
&\equiv \hat{P}(j|x) \qquad (38)
\end{aligned}$$

- This is the *crude Monte Carlo frequency simulator* for choice probabilities, which is *locally flat in x.*

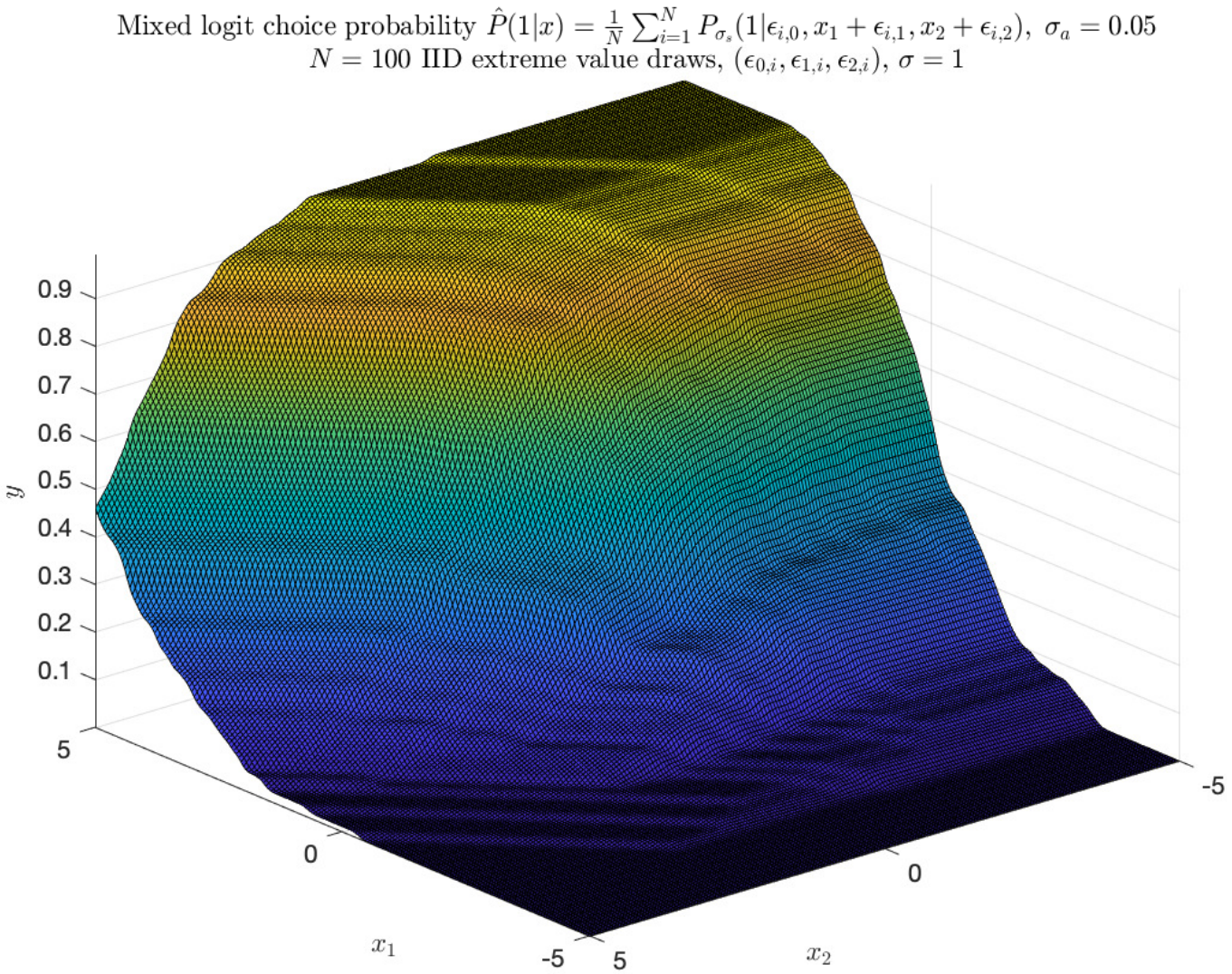
True logit probability $P(1|x)$



Crude frequency $\hat{P}(1|x)$ using $N = 100$ draws



Logit-smooothed $\hat{P}_{\sigma_s}(1|x)$, $\sigma_s = .05$, $N = 100$ draws



Additional thoughts/results on the EMAX function

- When the distribution of shocks $q(\epsilon|x)$ is continuous, the probability of ties for the utility maximizing choice is zero, so fro almost all ϵ values there is a unique utility maxizing choice, which we have denoted by the *decision rule* or function $\delta(x, \epsilon)$. As a result, we can write

$$\max_{d=1,\ldots,n}[x_d + \epsilon_d] = \sum_{d=1}^{n}[x_d + \epsilon_d]I\{\delta(x, \epsilon) = d\} \tag{39}$$

  and actually this holds for *all* (x, ϵ) even in the case of ties (where in such cases we choose any of the utility maximizing actions to be $\delta(x, \epsilon)$).
- Taking expectations of both sides with respect to ϵ we have

$$S(x) = E\left\{\max_{d=1,\ldots,n}[x_d + \epsilon_d]\right\} = \sum_{d=1}^{n} E\left\{[x_d + \epsilon_d]I\{\delta(x, \epsilon) = d\}\right\} \tag{40}$$

Conditional EMAX function $S_j(x)$

- Define the following *conditional EMAX function* $S_j(x)$ by

$$\begin{aligned} S_j(x) &= E\{x_j + \epsilon_j | \delta(x,\epsilon) = j\} \\ &= \frac{\int_\epsilon (x_j + \epsilon_j) I\{\delta(x,\epsilon) = j\} q(d\epsilon|x)}{P(j|x)} \end{aligned} \tag{41}$$

- $S_j(x)$ is the expected payoff the agent expects from choosing alternative j given that j is the utility maximizing choice, $j = \delta(x,\epsilon)$, but *before* the agent observes the ϵ shocks. Using the identity on the previous slide we have

$$S(x) = \sum_{j=1}^{n} S_j(x) P(j|x) \tag{42}$$

so the EMAX $S(x)$ is a choice-probability weighted average of the conditional EMAX functions $S_j(x)$.

Special case of conditional EMAX function

- Hotz-Miller (1993) showed that when $q(\epsilon|x)$ is Type 1 Extreme value with location parameter normalized to zero, we have

$$S_j(x) = E\left\{[x_j + \epsilon_j] | j = \delta(x,\epsilon)\right\} = S(x) = E\left\{\max_{d=1,\ldots,n}[x_d + \epsilon_d]\right\} \tag{43}$$

  but for other distributions, such as multivariate normal, generally $S_j(x) \neq S(x)$ though the choice probability weighted average of $\{S_j(x)\}$ does equal $S(x)$.
- Exercise: in the case where ϵ has a Type 1 extreme value distribution with location parameter normalized to 0 and scale parameter σ, show $S_j(x) - x_j = E\{\epsilon_j | \delta(x,\epsilon) = j\} = \sigma \log(P(j|x))$, which is equation (4.12) of Hotz and Miller 1993.

Simulating CCPs

- The normality assumption for random coefficients leads to an alternative model to MNL: the *multivariate probit model* (MNP)
- However unlike the MNL model the Emax/smoothed max function and choice probabilities no longer have the beautiful and easy to evaluate closed form expressions. Instead, we must do *numerical integration* to estimate the MNP by maximum likelihood. For $|D(x)| > 5$ this becomes highly burdensome due to the *curse of dimensionality of deterministic numerical integration.*
- *Monte carlo simulation* is an alternative way to avoid the curse of dimensionality by *simulating random utilities* to get an unbiased Monte carlo estimate of the CCP, $\hat{P}(d|x)$ given by

$$\hat{P}(d|x,\theta) = \frac{1}{N}\sum_{i=1}^{N} I\{u(x,d,\theta) + \tilde{\epsilon}_i(d) \geq u(x,d',\theta) + \tilde{\epsilon}_i(d') | d' \in D(x)\}$$

where $\{\tilde{\epsilon}_i\}_{i=1}^N$ are N *IID* draws from the conditional density of the random utility distribution $q(\epsilon|x)$.

Problems with Simulated Maximum Likelihood (SML)

- We could estimate θ by *simulated maximum likelihood* (SML)

$$\hat{\theta}_T = \underset{\theta}{argmax}\, \frac{1}{T} \sum_{t=1}^{T} \log(\hat{P}(d_t|x_t, \theta))$$

- Even though $E\{\hat{P}(d|x,\theta)\} = P(d|x,\theta)$ so the Monte Carlo simulator is unbiased, if we use $\hat{P}(d|x,\theta)$ instead of $P(d|x,\theta)$ we can get poor results due to *Jensen's inequality bias*

$$E\left\{\log(\hat{P}(d|x,\theta))\right\} < \log(P(d|x,\theta))$$

- Thus we generally require that $N \to \infty$ as $T \to \infty$ to remove the Jensen's Inequality bias and ensure consistency of $\hat{\theta}_T$.
- When N is sufficiently large, it is so time-consuming to compute $\hat{P}(d|x,\theta)$ that SML is not attractive: "Lerman and Manski (1981) suggest a Monte Carlo procedure for estimating $P(d|x,\theta)$ ... but find that it requires an impractical number of Monte Carlo draws to estimate small probabilities and their derivatives with acceptable precision." (McFadden, 1989)

Mixed Logit: a tractable solution to IIA problem

- Suppose we consider a combination of Type 1 extreme value additive RUM and random coefficients $\tilde{\beta}$ to get

$$u(x, d, \tilde{\beta}) + \tilde{\epsilon}(d) = x_d\tilde{\beta} + \tilde{\epsilon}(d) = x_d\mu + \tilde{\epsilon}(d) + x_d\tilde{\eta}$$

  where $\epsilon(d)$ is an *IID* extreme value random utility component and $\tilde{\beta} = \mu + \tilde{\eta}$ where $\tilde{\beta} \sim N(0, \Sigma)$ is $K \times 1$ random coefficients, and we assume that $\tilde{\eta}$ and $\tilde{\epsilon}$ are independently distributed.
- The composite error term $\tilde{\varepsilon}(d) = \tilde{\epsilon}(d) + x_d\tilde{\eta}$ will not satisfy IIA since $\tilde{\varepsilon}(d)$ and $\tilde{\varepsilon}(d')$ will be increasingly (though not perfectly) correlated as $x_d \to x_{d'}$.
- The implied choice probability is given by the *mixed logit* or *random coefficients logit* formula

$$P(d|x, \theta) = \int_\beta \left[\frac{\exp\{x_d\beta)\}}{\sum_{d' \in D(x)} \exp\{x_{d'}\beta)\}} \right] \phi(\beta|\mu, \Sigma)d\beta$$

  where $\theta = (\mu, \Sigma)$. In case where x_d has relatively low dimension, the mixed MNL model can be tractable even for large choice sets $D(x)$.

Mixed Logits: universal probability approximators

- Theorem 1 of McFadden and Train (2000) "Mixed MNL Models of Discrete Response" showed that if $P(d|z,x)$ is an arbitrary conditional choice probability resulting from random utility maximization over n choices,

$$P(d|z,s) = \text{Prob}\left\{ U_d(z_d,s,\epsilon_d,\nu(s)) = \max_{d'=1,\ldots,n}[U_{d'}(z_{d'},\epsilon_{d'},\nu(s))] \right\}, \tag{44}$$

where $z = (z_1,\ldots,z_n)$ is a vector of attributes of the n alternatives (so z_d are the attributes of alternative d) and s are characteristics of the decision maker, and (ϵ,ν) are the unobserved preference shocks with an arbitrary (but continuous) density $q(\epsilon,\nu|s,z)$, and for each choice d $U_d(z_d,s,\epsilon_d,\nu(s))$ is a continuous function of its arguments. Then there exists a mixed logit model that approximates $P(d|z,s)$ arbitrarily closely over its (compact) domain.

Continuous and discrete mixed logits

- That is, for any $\eta > 0$ there exists a K and a continuous function $x = (x(z_1, s), \ldots, x(z_n, s))$ where $x(z_j, s) \in R^K$ and a mixing distribution $G(\beta)$ over the K random coefficients $\beta \in R^K$ of a mixed MNL model such that $|P(d|x, G) - P(d|z, s)| < \eta$ for all $d \in \{1, \ldots, n\}$ and all (z, s) in a compact set, where the mixed logit model $P(d|x, G)$ is given by

$$P(d|x, G) = \int_\beta \left[\frac{\exp\{x_d \beta\}}{\sum_{d'=1}^n \exp\{x_{d'} \beta\}} \right] G(d\beta) \tag{45}$$

- In fact, one could use a finite mixture with r "types" to do this approximation, so let β_t be a coefficient vector of "type t" and let there be a total of r types and let ω_t be the probability of a type t agent in the population. Then we can also write an approximator as a finite mixture as follows

$$P(d|x, G) = \sum_{t=1}^r \omega_t \left[\frac{\exp\{x_d \beta_t\}}{\sum_{d'=1}^n \exp\{x_{d'} \beta_t\}} \right]. \tag{46}$$

Relationship to neural networks

- This can be seen as a special case of more general approximation results for *single layer feedfoward neural networks* (also called multi-layered perceptrons). Suppse we want to approximate a given continuous function $y = f(x)$ where $f : R^n \to R$ (i.e. $x = (x_1, \ldots, x_n)$). Let $\sigma : R \to R$ be a *squashing function* such as $\sigma(z) = 1/(1 + exp\{z\})$ (i.e. a logit). Asingle layer feedfoward neural network with r "hidden units", $f(x, r, \omega, \beta, \gamma)$, can approximate f arbitrarily well if r and the parameters $\theta = (\omega, \beta, \gamma)$ are chosen appropriately

$$f(x, \theta) = \sum_{i=1}^{r} \omega_i \sigma(x\beta_i + \gamma_i) \tag{47}$$

- We can regard $f(x, \theta)$ as a *nonlinear regression* and do nonlinear least squares to "fit" it, i.e. $\hat{\theta} = argmin_\theta \|f(x) - f(x, \theta)\|$ where $\|f\|$ is a *norm* such as $\|f\| = \sup_x |f(x)|$, the *sup-norm.*

Universal approximation theorem for neural networks

- Let $E(r)$ be the *worst case error* for using a single hidden layer neural network with r hidden units to approximate a class of smooth functions $\mathcal{F}$ such as the class of all m-times differentiable functions of n whose domain D is a compact subset of R^n. So we define $E(r)$ as follows

$$E(r) = \sup_{f \in \mathcal{F}} \inf_{\theta} \sup_{x \in D} |f(x) - f(x, \theta)|. \tag{48}$$

- **Universal Approximation Bound (Maiorov 1999)** *Let $E(r)$ denote the worst case error between an m-times differentiable function f and a single layer feedforward neural network with r hidden units. Then we have for each $n \geq 2$*

$$E(r) = \Theta\left(r^{-\frac{m}{n-1}}\right). \tag{49}$$

where Θ denotes a lower and upper bound on the rate of convergence, i.e. there exist universal constants C_1 and C_2 such that

$$C_1 r^{-\frac{m}{n-1}} \leq E(r) \leq C_2 r^{-\frac{m}{n-1}}. \tag{50}$$

Smooth unbiased simulation of choice probabilities

- For mixed logit model, take S *IID* draws $(\tilde{\beta}_1, \ldots, \tilde{\beta}_S)$ from the $N(\mu, \Sigma)$ distribution. Let $\Sigma^{1/2}$ be a Cholesky factor of Σ so we can write $\tilde{\beta}_j = \mu + \Sigma^{1/2}\tilde{\eta}_j$ where $\tilde{\eta}_j \sim N(0, I)$.

$$\hat{P}(d|x,\theta) = \frac{1}{S}\sum_{s=1}^{S} \frac{\exp\{x_d(\mu + \Sigma^{1/2}\tilde{\eta}_j)\}}{\sum_{d' \in D(x)} \exp\{x_{d'}(\mu + \Sigma^{1/2}\eta_j)\}}$$

  is a smooth, unbiased simulator of $P(d|x,\theta)$.
- For multivariate probit the *Geweke-Hajivassiliou-Keane* (GHK) simulator is a smooth, unbiased simulator of MNP choice probabilities and other orthant probabilities under the multivariate normal distribution. It is based on the Cholesky factorization of the multivariate normal covariance matrix and a recursive procedure for simulating univariate truncated normal distributions, combined wtih importance sampling to reduce the variance of the simulator.
- The comparison by Hajivassiliou, McFadden and Ruud (1994) concludes "the GHK simulator appears overall the most reliable method, especially for simulating orthant probabilities."

Method of Simulated Moments

- Notice that mixed logit still requires numerical integration but it can be feasible if the dimension K of the random coefficients is not to large. Simulation methods make mixed logit or MNP feasible for bigger problems, but consistency of SML requires $S \to \infty$ as $T \to \infty$ and so SML is not a panacea either.
- McFadden (1989) introduced the *method of simulated moments* (MSM) which can enable consistent (CAN) estimation of discrete choice models with *only 1 simulated utility draw per observation.* The basic idea of MSM is that *we can use the law of large numbers to average out simulation error in the same way we use the LLN to average out sampling error.*
- Let d_t be *choice vector* for person t who is in observed state x_t. That is, if there are $|D(x_t)|$ choices for this consumer, then d_t is a vector which equals 0 for alternatives that the person did not choose and 1 for the alternative the person chose. Let $P(x_t, \theta^*)$ be the stacked vector of choice probabilities of dimension $|D(x_t)|$.

Method of Simulated Moments

- For example if $D(x_t) = \{1, 2, 3\}$ then if the person chose alternative 2, then $d_t = (0, 1, 0)$ and $P(x_t, \theta^*)$ is given by

$$P(x_t, \theta^*) = [P(1|x_t, \theta^*), P(2|x_t, \theta^*), P(3|x_t, \theta^*)].$$

- If we assume the discrete choice model is correctly specified, then

$$E\{d_t|x_t\} = P(x_t, \theta^*)$$

so $d_t = P(x_t, \theta^*) + \varepsilon_t(\theta^*)$ where $E\{\varepsilon_i(\theta^*)|x_i\} = 0$.

- Consider GMM/Minimum distance estimation if we could calculate $P(x_t, \theta)$ exactly for any θ. Then we have the unconditional moment restriction $E\{x_t \otimes \varepsilon_t(\theta^*)\} = E\{x_t \otimes (d_t - P(x_t, \theta^*))\} = 0$ where the $K \times 1$ vector x_t is multiplied by each of the $D_t = |D(x_t)|$ components of $d_t - P(x_t, \theta^*)$ so it is a $K * D_t \times 1$ vector given by the Kronecker product $x_t \otimes (d_t - P(d_t, \theta^*)$. Then the GMM estimator $\hat{\theta}_T$ is

$$\hat{\theta}_T = \underset{\theta}{argmin} \frac{1}{T} \sum_{t=1}^{T} [x_t \otimes (d_t - P(x_t, \theta)]' [x_t \otimes (d_t - P(x_t, \theta))]$$

Method of Simulated Moments

- Notice that the Law of Large Numbers implies that sampling error averages out to zero, so with probability 1 we have

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} x_t \otimes (d_t - P(x_t, \theta^*)) = E\{x_t \otimes (d_t - P(x_t, \theta^* a))\} = 0.$$

- Now consider a simulated set of choices $\tilde{d}_t(\theta)$ at parameter θ. Suppose we draw a vector of random additive error terms $\tilde{\epsilon} \sim q(\epsilon|x_t)$ (e.g from the Type 1 extreme value distribution), one for each choice. Then $d_t(\theta)$ will have a 1 in the component corresponding to the *simulated chosen alternative* and zeros elsehwere. Note that we can use only *only 1 draw of the* ϵ *vector* to generate $\tilde{d}_t(x)$ from the simulated utilities $u(x_t, d, \theta) + \tilde{\epsilon}(d)$, $d \in D(x_t)$. Then we have

$$E\{\tilde{d}_t(\theta)\} = P(x_t, \theta)$$

so we also have the following regression equation $\tilde{d}_t(\theta) = P(x_t, \theta) + \tilde{\varepsilon}_t(\theta)$ where $E\{\varepsilon_t(\theta)|x_t\} = 0$ and the unconditional moment condition $E\{x_t \otimes \tilde{\varepsilon}_t(\theta)\} = 0$ holds for each θ.

Method of Simulated Moments

- Now define the MSM estimator $\hat{\theta}_T^{\text{msm}}$ by

$$\hat{\theta}_T^{\text{msm}} = \underset{\theta}{\text{argmin}} \frac{1}{T} \sum_{t=1}^{T} [x_t \otimes (\tilde{d}_t(\theta) - d_t)]' [x_t \otimes (\tilde{d}_t(\theta) - d_t)]$$

- Notice that this estimator is based on *only a single set of simulated utilities per observation.* We can of course use more than a single simulation per observation. If we draw S simulations per observation, then let $\hat{d}_t^S(\theta)$ be the average of the simulated choice vectors over the S independent simulations of utilities for observation t.
- It is critical to use *common random numbers* when minimizing the MSM objective function (51). That is, we draw the random additive utility shock vectors $\tilde{\epsilon}_t$ for each person in the sample at the start of the estimation process and keep the same set of shocks fixed as we search over θ to minimize (51).

Method of Simulated Moments

- For Type 1 extreme value utility shocks, we use the probability integral transform

$$F(\tilde{\epsilon}) = \exp\{-\exp\{-\tilde{\epsilon}\}\} = \tilde{u} \sim U[0,1]$$

so we have $\tilde{\epsilon} = -\log(-\log(\tilde{u}))$.

- Intuition for consistency of MSM estimator

$$\frac{1}{T}\sum_{t=1}^{T}[\tilde{d}_t(\theta) - d_t]$$

$$= \frac{1}{T}\sum_{t=1}^{T}[\tilde{d}_t(\theta) - P(x_t,\theta)] + [P(x_t,\theta) - P(x_t,\theta^*)] + [P(x_t,\theta^*) - d_t]$$

The probability limit of this sum is $E\{\tilde{d}_t(\theta) - P(x_t,\theta)\}$ $+E\{[P(x_t,\theta) - P(x_t,\theta^*)]\} + E\{[P(x_t,\theta^*) - d_t]\}$. The first and last term are zero, so the limit converges to $E\{[P(x_t,\theta) - P(x_t,\theta^*)]\}$. Thus, by setting $\theta = \theta^*$ can set the surviving term to zero, so asymptotically the MSM estimation criterion is minimized at θ^* as $T \to \infty$, so $\hat{\theta}_T^{\text{msm}} \to \theta^*$.

Daniel McFadden

Another drawback of the Multinomial logit model

Equal cross elasticities

An implication of MNL is that all cross elasticities are equal, i.e. if $v(x,d) = x_d\beta$ where x_d is a vector of attributes of alternative d (such as its price), then we have for all $d' \neq d$

$$\frac{\partial \log P(d|x)}{\partial \log x_{d'k}} = \beta_k x_{d'k} P(d'|x)$$

- The tractability of MNL makes it very popular in IO as a way of deriving market demand via *microaggregation* of individual discrete choices of individual consumers
- Then the limit, market shares for a finite number of goods equal a weighted average of logit choice probabilities for individual products in the market

Equality of cross-elasticities problematic in IO

- Berry, Levinson and Pakes (1995) (BLP) noted that the logit form can be problematic for IO and a typical specification has utility given by

$$u(x, d) = x_d\beta - \alpha p_d + \xi_d + \epsilon(d) \equiv \delta_d + \epsilon(d)$$

where ξ_d is an *unobserved product characteristic* the mean (over consumers) of the unobserved utility components $\epsilon(d)$.

Despite this computational simplicity, the assumption that the utility function is additively separable into two terms, one determined entirely by the product characteristics (the δ_d) and one determined by the consumer characteristics ($\epsilon(d)$) is problematic. This is because it generates aggregate substitution patterns, and hence a set of (cross and own) price derivatives, as well as responses to the introduction of new products, that cannot possess many of the features that we expect them to have.

Problematic aspect due to IID $\epsilon(d)$

Thus, if we were using the specification to analyze an automobile market in which an inexpensive Yugo and an expensive Mercedes had the same market shares, then the parameter estimates would have to imply that the two cars have the same cross-price derivative with respect to any third car. In particular, the model would necessarily predict that an increase in the price of a BMW would generate equal increases in the demand for Yugos and for Mercedes. This contradicts the intuition which suggests that couples of goods whose characteristics are more 'similar' should have higher cross-price elasticities. We expect this to happen because the consumers who would have chosen a BMW at the old prices, but now do not, have a preference for large cars and are therefore likely to move to another large car. Similarly, when a new car enters the market, we expect it to have a large effect on the demand for cars with similar characteristics. Additive separability plus IID ϵ, on the other hand, imply that a consumer who substitutes away from any given choice will tend to substitute toward other popular products, not to other similar products and does not depend on the distribution for the $\epsilon(d)$ (e.g. logit).

Solution to the problem: random coefficients

- Suppose we also allow the β coefficients on the characteristics x_d of each alternative $d \in D$ to be *random variables.* This is a different source of *unobserved heterogeneity* that results in *correlation of preferences for items with similar characteristics.* For example if x_{dk} is the attribute "weight" of a car, a consumer who has a large positive β_k coefficient will prefer larger cars, other things equals.
- Let $P(d|x,\beta)$ be the probability of choosing item d given the product characteristics x and conditioning on the consumer type β. This might be an MNL model and conditional on β it suffers from the IIA and equal cross substitution problems give above
- But market demand is a *mixed logit* model given by

$$P(d|x) = \int_\beta P(d|x,\beta) f(\beta) d\beta$$

Why do random coefficients solve the problem?

BLP note:

The utility obtained from consuming good d can still be decomposed into a mean $\delta_d = x_d E\{\beta\} - \alpha p_d + \xi_d$ and a deviation from that mean $\nu_d = \epsilon(d) + x_d(\beta - E\{\beta\})$ but now ν_d depends on the interaction between consumer preferences and product characteristics. As a result, consumers who have a preference for size will tend to attach high utility to all large cars, and this will induce large substitution effects between large cars. Note, however, that though this specification allows for more realistic cross- price elasticities, it re-introduces the problem of computing the integral that defines market shares (via microaggregation of demand) as a function of the parameters of the model.

BLP: estimating demand models with aggregate data

- Most common type of data in IO applications: *aggregate market shares.* Much less common to have access to *micro data* i.e. data on choices made by individuals or households.
- Berry 1994 considered "the problem of 'supply and demand' on a cross section of oligopoly markets with differentiated products." That is, unit of analysis is *an individual market* so the observations are of *aggregate demand in each market* not the choices of individual households in these markets.
- The "traditional" approach to demand analysis posits an aggregate demand curve, that predicts the aggregate quantity of good j in a particular market, q_j as a function of its price and those of other goods. For example Berry considers a constant elasticity demand curve, i.e.

$$\log(q_j) = \alpha_j + \sum_{k=1}^{N} \eta_k \log(p_k) + \epsilon_j \tag{51}$$

Problems with traditional demand estimation

- "The well-known problem, however, is that a system of N goods gives N^2 elasticities to estimate, which is a very large number in many real world applications. For example in the automobile industry model of Breshnahan (1987), there are close to 100 distinct products, implying almost 10,000 separate elasticities."
- So Berry considers an alternative approach: *microaggregation* i.e. deriving aggregate market demand by "adding up" the demands of individual households
- "Discrete choice models are a common, tractable, and parsimonious method for obtaining the desired structure on demand. This parsimony comes at some cost, as the models rule out the purchase of multiple items and do not easily incorporate dynamic aspects of demand."

A discrete choice model of product demand

- Berry models the utility of an individual household i who purchases a single item j by

$$u_{ij} = x_j \tilde{\beta}_i - \alpha p_j + \xi_j + \epsilon_{ij} \tag{52}$$

where ϵ_{ij} is an extreme valued taste shock. $\tilde{\beta}_i$ are *random coefficients* for household i, and ξ_j captures *unobserved product characteristics* that are not captured by the *observed product characteristics* x_j.

- The market share s_j of good j can be computed by microaggegation, i.e. integrating over the distribution of households

$$s_j(x, \xi, p, \mu, \Sigma) = \int_\beta \left[\frac{\exp\{x_j \beta - \alpha p_j + \xi_j\}}{1 + \sum_{k=1}^N \exp\{x_k \beta - \alpha p_k + \xi_k\}} \right] f(d\beta | \mu, \Sigma) \tag{53}$$

where $f(\beta|\mu, \Sigma)$ is a $N(\mu, \Sigma)$ distribution of β coefficients, $p = (p_1, \ldots, p_N)$, $x = (x_1, \ldots, x_N)$, and $\xi = (\xi_1, \ldots, \xi_N)$ is the vector of unobserved product characteristics.

Market size and the outside good

- This model relies on two key assumptions: 1) there is an "outside good" denoted by choice index $j = 0$ whose utility is normalized to 0, and 2) we can observe the total size of the market, M.
- With these two assumptions we can use data on M and the quantities sold by the N firms to construct market shares $(s_0, s_1, \ldots, s_N)$ using the relations $q_j = Ms_j$ and $q_0 = Ms_0$.
- In many real world markets, we do not observe total market size M or the number of consumers choosing the outside good, q_0. Reality is also complicated by the fact that many consumers can be *multiple units* of the same good, something the simple discrete choice model does not handle well.
- **Example:** A local hotel market. Even if we observe the occupancies $(q_{1,t}, \ldots, q_{N,t})$ of the N hotels in the market on a given night t, we don't observe the total number of consumers M_t who "arrive" looking for a hotel, nor $q_{0,t}$ the number of consumers who didn't "arrive" or arrived and chose some other option (e.g. AirBnB).

The supply side and Bertrand Nash equilibrium

- Assume for simplicity that firm j has marginal cost of production $c_j = \exp\{w_j\gamma + \omega_j\}$, where w_j are observed "cost characteristics" (which include x_j) and ω_j is un observed cost-shifter (which could be correlated with ξ_j). The firm's profits (ignoring fixed costs) are

$$\pi_j(x, \xi, p, \mu, \Sigma) = Ms_j(x, \xi, p, p, \mu, \Sigma)[p_j - c_j] \tag{54}$$

- A Bertrand-Nash equilibrium is a vector of prices $p^* = (p_1, \ldots, p_N)$ such that each firm is maximizing its profits given the prices of its competitors. The first order condition for profit maximization implies this system of equations

$$p_j^* = c_j - \frac{s_j(x, \xi, p^*, \mu, \Sigma)}{\frac{\partial}{\partial p_j} s_j(x, \xi, p^*, \mu, \Sigma)}, \quad j = 1, \ldots, N \tag{55}$$

- The solution to this system of equation results in a vector of equilibrium prices $p^*(x, \xi, w, \omega, \mu, \Sigma)$ that is an implicit function of the vectors of observed and unbobserved characteristics (x, w, ξ, ω) as well as the parameters (μ, Σ) characterizing consumer heterogeneity. *The solution may not be unique!*

Formulating the model as a nonlinear regression

- Suppose we tried to use observed market shares $s_r = (s_{1,r}, \ldots, s_{N,r})$, on $r = 1, \ldots, R$ independent markets by non-linear least squares treating the unobserved characteristics of the products, ξ_r, in each market r as "error terms" whereas x_r are observed product characteristics and p_r are the prices in each market and $\theta = (\mu, \Sigma)$ are the parameters to be estimated to characterize demand.
- But to do this, we would have to make an assumption about the distribution of the ξ_r error terms, such as $f(\xi|\rho)$ that depends on some additional parameters ρ. Then we can define the conditional expectation $E\{s_j(x, \xi, p, \theta)|x, p, \rho\}$ given by

$$E\{s_j(x, \xi, p, \theta)|x, p, \rho\} = \int_{\xi} s_j(x, \xi, p, \theta) f(d\xi|\rho) \tag{56}$$

and if this model is correctly specified, we can write this system of regression equations for the vector of true market shares

$$s_j = E\{s_j(x, \xi, p, \theta)|x, p, \rho\} + \nu_j \quad j = 1, \ldots, N \tag{57}$$

Estimating the model using nonlinear least squares

- Add ρ to the parameters to be estimated $\theta = (\mu, \Sigma, \rho)$. The nonlinear least squares estimator $\hat{\theta}_R$ is given by

$$\hat{\theta}_R = \underset{\theta}{argmin} \sum_{r=1}^{R} \sum_{j=1}^{N} [s_{jr} - E\{s_{jr}(x_r, \xi_r, p_r | x_r, p_r, \theta\}]^2 \qquad (58)$$

  However a problem is that if Bertrand equilibrium holds in each market, the prices p_r in this regression will be functions of (x_r, ξ_r) and hence correlated with the "error terms" $\{\nu_{jr}\}$. Thus we have a problem of *price endogeneity* and just like the corresponding problem in a simple linear regression, we would not expect nonlinear least squares to be consistent. Berry provides examples such as Trajtenberg's (1989) study of the CT scanner market where "in some cases, prices appear to have a positive effect on demand."
- But using instrumental variables to estimate equation (58) won't work because the endogenous covariate vector p_r enters the regression functions $E\{s_{jr}(x_r, \xi_r, p_r | x_r, p_r)\}$ in a nonlinear fashion, but IV is only guaranteed to work for *linear regressions.*

What can we do? Invert market shares, then do IV!

- Consider first the simpler case without random coefficients, so market shares are given by the simple MNL model

$$s_j(x,\xi,p) = \frac{\exp\{x_j\beta - \alpha p_j + \xi_j\}}{1 + \sum_{k=1}^{N} \exp\{x_k\beta - \alpha p_k + \xi_k\}} \tag{59}$$

- Now recall the Hotz-Miller inversion: we can use it to analytically invert the true market shares s_j to get the following *linear regression equation* for the utility of good j

$$\log(s_j) - \log(s_0) \equiv \delta_j = x_j\beta - \alpha p_j + \xi_j \tag{60}$$

- Even though we do not "observe" consumer utilities, we can "observe" the left hand side of equation (60), i.e. the δ_j's, and use them as a basis for *linear regression* using ξ_j directly as the error term.
- Now we *can* legitimately use IV to deal with the problem of endogeneity in p_j, i.e. that equilibrium prices are functions of (x,ξ) *assuming we can find good instruments!*

Instruments for unobserved product characteristics

- BLP suggested using $z = (x, w)$ as the instruments, together with the following *orthogonality conditions*

$$\begin{aligned} E\{\xi | x, w\} &= 0 \\ E\{\omega | x, w\} &= 0. \end{aligned} \tag{61}$$

- Now, project the endogenous price variables p and unobserved product characteristics ξ onto the instruments $z = (x, w)$,

$$p = E\{p^*(x, \xi, w, \omega) | x, w\} + u \tag{62}$$

- Then using the orthogonality condition $E\{\xi_j | x, w\} = 0$ we can re-write the inverted market share regression equation (60) as

$$\log(s_j) - \log(s_0) \equiv \delta_j = x_j\beta - \alpha E\{p_j^*(x, \xi, w, \omega) | x, w\} - \alpha u_j + \xi_j \tag{63}$$

and the composite error term $\nu_j = (-\alpha u_j + \xi_j)$ is by construction mean-independent of x, i.e. $E\{\nu_j | x, w\} = 0$. So we could potentially consistently estimate β and α using non-parametric regression to estimate $E\{p_j^*(x, \xi) | x, w\}$ in the "first stage" and estimate (β, α) in a second stage regression using equation (63).

Estimating via GMM instead of 2SLS

- Let $H(x, w)$ be a $L \times 2N$ matrix that we use to construct moments for GMM estimation of $\theta^* = (\alpha^*, \beta^*, \gamma^*)$, noting that the conditional moment restrictions (61) imply the following unconditional moment restriction

$$E\{H(x, w)\epsilon\} = 0, \tag{64}$$

  where $\epsilon = (\xi, \omega)'$ is the $2N \times 1$ vector of unobserved product characteristics and cost shocks for the N products.
- Define $\epsilon(\theta) = [\delta - (x\beta - p\alpha), \log(c) - w\gamma]'$. Notice that if the model is correctly specified, $\epsilon(\theta^*) = [\xi, \omega]'$. We can write $\epsilon(\theta) = [x(\beta^* - \beta) - p(\alpha^* - \alpha) + \xi, w(\gamma^* - \gamma) + \omega]'$ so we have

$$E\{H(x, w)\epsilon(\theta)\} = 0 \quad \text{when } \theta = \theta^* \tag{65}$$

- Let W be a $L \times L$ positive definite weighting matrix, and note that equation (65) implies

$$\theta^* = \underset{\theta}{argmin}\, E\{H(x, w)\epsilon(\theta)\}' W E\{H(x, w)\epsilon(\theta)\}, \tag{66}$$

  For purposes of identification assume that θ^* is the *unique minimizer* of (66).

Definition of the GMM estimator, $\hat{\theta}_R$

- Define the empirical unconditional moments $\overline{H}_R(\theta)$ by

$$\overline{H}_R(\theta) = \frac{1}{R}\sum_{r=1}^{R} H(x_r, w_r)\epsilon_r(\theta), \tag{67}$$

where $\epsilon_r(\theta) = [\delta_r - (x_r\beta - p_r\alpha), \log(c_r) - w_r\gamma]'$ are the "residuals" for market r.

- Define the GMM estimator $\hat{\theta}_R$ by

$$\hat{\theta}_R = \underset{\theta}{argmin}\, \overline{H}_R(\theta)' W \overline{H}_R(\theta). \tag{68}$$

- If $\lim_{R\to\infty} \sup_\theta |H_R(\theta) - E\{H(x,w)\epsilon(\theta)\}| = 0$ with probability 1, then $\hat{\theta}_R$ will be a consistent estimator of θ^*.
- The first order conditions are for the minimization in (68) are

$$0 = \nabla \overline{H}_R(\hat{\theta}_R)' W \overline{H}_R(\hat{\theta}_R). \tag{69}$$

Asymptotic distribution of GMM estimator

- Note that $\nabla \overline{H}_R(\theta)$ is the $L \times D$ matrix of gradients of $\overline{H}_R(\theta)$ with respect to θ where $D = 1 + \dim(\beta) + \dim(\gamma)$ is the dimension of the θ vector. It is given by

$$\nabla \overline{H}_R(\theta) = \frac{1}{R} \sum_{r=1}^{R} H(x_r, w_r) \begin{bmatrix} -x_r \\ p_r \\ -w_r \end{bmatrix}. \tag{70}$$

- We can do a Taylor series expansion of $\overline{H}_R(\theta_R)$ about θ^* to get

$$\overline{H}_R(\hat{\theta}_R) = \overline{H}_R(\theta^*) + \nabla \overline{H}_R(\overline{\theta}_R)(\hat{\theta}_R - \theta^*) \tag{71}$$

  for a vector $\overline{\theta}_R$ (where we use coordinate-wise mean value theorem to define this vector).
- Substituting this into the first order condition (69) and solving for $\sqrt{R}(\hat{\theta}_R - \theta^*)$ we get

$$\sqrt{R}(\hat{\theta}_R - \theta^*) = -\left[\nabla \overline{H}_R(\hat{\theta}_R)' W \overline{H}_R(\overline{\theta}_R)\right]^{-1} \left[\nabla \overline{H}_R(\hat{\theta}_R)' W \sqrt{R} \overline{H}_R(\theta^*)\right] \tag{72}$$

Asymptotic distribution of GMM estimator, cont

- Under suitable regularity conditions, a central limit theorem can be established to show

$$\sqrt{R}\overline{H}_R(\theta^*) \Longrightarrow N(0, \Gamma), \tag{73}$$

where Γ is an $L \times L$ covariance matrix given by

$$\Gamma = E\left\{H(x, w)\Omega H(x, w)'\right\}, \tag{74}$$

where Ω is the $2N \times 2N$ covariance matrix of $\epsilon(\theta^*)$, i.e.

$$\Omega = \left[\begin{array}{cc} \mathrm{cov}(\xi, \xi) & \mathrm{cov}(\xi, \omega) \\ \mathrm{cov}(\omega, \xi) & \mathrm{cov}(\omega, \omega) \end{array} \right]. \tag{75}$$

- Using these results, equation (72), and the continuous mapping theorem, we can show

$$\sqrt{R}[\hat{\theta}_R - \theta^*] \Longrightarrow N(0, \Lambda(\theta^*)), \tag{76}$$

Asymptotic distribution of GMM estimator, cont

- where $\Lambda(\theta^*)$ is a $D \times D$ symmetric matrix that has the "sandwich" form, $\Lambda(\theta^*) = \Lambda_1(\theta^*)\Lambda_2(\theta^*)\Lambda_1(\theta^*)'$ where

$$\begin{aligned} \Lambda_1(\theta^*) &= [[\nabla E\{H(x,w)\epsilon(\theta^*)\}]' W [\nabla E\{H(x,w)\epsilon(\theta^*)\}]]^{-1} \\ \Lambda_2(\theta^*) &= [\nabla E\{H(x,w)\epsilon(\theta^*)\}]' W \Gamma W [\nabla E\{H(x,w)\epsilon(\theta^*)\}], \end{aligned} \tag{77}$$

  where

$$\nabla E\{H(x,y)\epsilon(\theta^*)\} = E\left\{ H(x,w) \begin{bmatrix} -x \\ p \\ -w \end{bmatrix} \right\}. \tag{78}$$

- The optimal weighting matrix for the moments $H(x,w)$ is W^* given by $W^* = \Gamma^{-1}$, assuming Γ is invertible. Then the asymptotic covariance matrix similifies to

$$\Lambda(\theta^*) = \left[[\nabla E\{H(x,w)\epsilon(\theta^*)\}]' \Gamma^{-1} [\nabla E\{H(x,w)\epsilon(\theta^*)\}] \right]^{-1} \tag{79}$$

  This is the smallest asymptotic covariance matrix of any GMM estimator that uses $H(x,w)$ to form unconditional moments.

Comments on BLP's Moment Restriction

- The restriction $E\{\omega|x,w\}=0$ is plausible if $w_j\gamma$ is a good approximation to $E\{\log(c_j)|x,w\}$

$$\log(c_j)=w_j\gamma+\omega_j. \tag{80}$$

But BLP's other *a priori* moment condition, that $E\{\xi|x,w\}=0$, is more problematic. Why should unobserved car characteristics ξ be mean independent of other observed characteristics x and variables characterizing costs of production, w?

- It only makes sense if the observed characteristics if very complete so that $x_j\beta$ provides a very good approximation of the utility consumers get from product j and thus the utility due to unobserved characteristics ξ_j is akin to "random noise"
- What if we relax the assumption $E\{\xi|x,w\}=0$. What can we do instead? Well, we can "model the error term" using a parametric function $g_j(x,w,\psi)$ for the conditional expectation of ξ

$$\xi_j=E\{\xi_j|x,w\}+\zeta_j=g_j(x,w,\psi)+\zeta_j \tag{81}$$

were $E\{\zeta_j|x,w\}=0$. Now the unknown parameters are $\theta=(\alpha,\beta,\gamma,\psi)$.

Nonlinear 2SLS when $E\{\xi_j|x,w\} \neq 0$

- Re-write equation (63) using equation (81) as

$$\delta_j = x_j\beta - \alpha p_j + g_j(x,w,\psi) + \zeta_j \tag{82}$$

- We can estimate (α,β,ψ) by nonlinear 2SLS if we replace the endogenous price covariate p_j by a non-parametric prediction of it, $E\{p_j^*(x,\xi,w,\omega)|x,w\}$ in equation (82).

$$\delta_j = x_j\beta - \alpha E\{p_j^*(x,\xi,w,\omega)|x,w\} + g_j(x,w,\psi) - \alpha u_j + \zeta_j \tag{83}$$

- Note that the error term in equation (83) is $\nu_j = -\alpha u_j + \zeta_j$. But since $E\{u_j|x,w\} = 0$ by (62) and $E\{\zeta_j|x,w\} = 0$ from (81), it follows that $E\{\nu_j|x,w\} = 0$ so equation (82) is a valid regression function, so "nonlinear 2SLS" will consistently estimate (α,β,ψ).
- This is an example of how to control for endogeneity by "modeling" the error term ξ_j. Note that if we assumed $g_j(x,w,\psi) = x_j\psi$, then we have multi-collinearity of $x_j\beta$ and $x_j\psi$ so we can only identity the sum of coefficients $(\beta+\psi)$, so the BLP coefficients of x_j are a combination of the "true effects" of x_j on utility, plus a "distortion" ψ since the model also uses x_j to control as best it can for the unobserved characteristic ξ_j.

BLP when $E\{\xi_j|x,w\} \neq 0$

- Instead of 2SLS, we can estimate θ via GMM using the same approach that BLP did, under the moment restriction $E\{\xi|x,w\} = 0$. Our new moment restriction when we relax $E\{\xi_j|x,w\} = 0$ is $E\{\zeta_j|x,w\} = 0$, where ζ_j is the residual in the potentially nonlinear regression $\xi_j = g_j(x,w,\psi) + \zeta_j$.
- So now define a new residual vector $\epsilon(\theta)$ by

$$\epsilon(\theta) = [\delta - (x\beta - p\alpha + g(x,w,\psi)), \log(c) - w\gamma]' \qquad (84)$$

where we have stacked the product specific components, $x = (x_1,\ldots,x_N)'$, $w = (w_1,\ldots,w_N)'$, $p = (p_1,\ldots,p_N)'$, $c = (c_1,\ldots,c_N)'$, and $g(x,w,\phi) = (g_1(x,w,\psi),\ldots,g_N(x,w,\psi))'$.

- We repeat the same GMM approach that BLP used to estimate (α,β,γ) under the restriction that $E\{x_j|x,w\} = 0$ but now we also have to estimate the additional parameter ψ when we relax this restriction.
- Question: instead of estimating $g_j(x,w,\phi)$ why not just directly estimate the ξ_j as fixed effects (i.e. coefficients of product-specific dummy variables)?

Issues with aggregate market share data

- Clearly we cannot estimate ξ_j directly since δ_j is already itself "the fixed effect" or "sufficient statstic" for the market share of good j, s_j! If we treat ξ_j as a parameter or fixed effect, it is perfectly collinear with $x_j\beta$ and even $-\alpha p_j$ so we can longer separately estimate (α, β).
- If we tried to do BLP using only $R = 1$ market (e.g. the US auto market), then we would have $2N$ "observations, (p, s) (i.e. the prices and market shares of the N products excluding the outside good), but $2N + K + 1$ parameters (α, β) if we also treated ξ and the marginal costs c as fixed effects. So BLP are forced to parametrize the model more parsimoniously, explaining the $2N$ "data points" (p, s) in terms of $(1 + K + J)$ parameters (α, β, γ) where J are the number of w variables.
- However if we have access to *micro data* then we have many more data points (observations of choices of each of M consumers), so can we estimate the ξ_j as fixed effects, along with (α, β) in this case?

Estimating the model with micro cross section data

- Actually, in a *homogeneous market where all consumers face the same prices,* we still can't identify the N fixed effects ξ and the $1 + K$ additional parameters (α, β) even using micro data.
- Why? The reason is that despite the micro observations on choices, in a large market of identical consumers who all face goods with the same characteristics x and the same prices p, then their choice probabilities are also the same. Call this common probability vector $P(x, p, \xi, \alpha, \beta)$ which in equilibrium also equals the $N \times 1$ vector of market shares s (excluding the outside good), so there are really only N pieces of information or "knowns". But there are a total of $1 + K + N$ parameters (α, β, ξ), so we cannot identify all of the parameters: *more unknowns than equations!*
- By Hotz-Miller the most we can do is invert the choice probability vector to get the $N \times 1$ δ vector from the observed market shares (excluding the outside good) i.e. $\delta = P^{-1}(s)$. But then

$$P^{-1}(s) = \delta = x\beta - \alpha p + \xi \tag{85}$$

which is linear system of N equations in $(1 + K + N)$ unknowns.

What do we need to identify (α, β, ξ)?

- To identify (α, β, ξ) we would need some sort of *individual specific price and attribute variation* so each individual i in the micro sample faces different (x_i, p_i) values.
- However this is incompatible with the standard setup, which considers consumers in a single market who all face the same products (and hence observed attributes x) and prices p. The simple model in Berry (1994) and BLP (1995) do not allow for such variation.
- If we assume *multiple markets* and allow (x_r, p_r) to differ across markets, then we can potentially identify (α, β, ξ) *provided we assume that the unobserved characteristics* ξ *are the same in all* R *markets.* But why allow x_r to vary over r but not ξ?
- This is implausible and in such cases, better to take a *random effects* approach where ξ_r is an *unobserved random variable* that differs across markets just as we allow observed characteristics x_r to vary over markets.

Exploiting cross-market variation to identify (α, β)

- If we posit some restrictions on the distribution of ξ (as BLP do), the model allows for variation in p_r over markets since in each market r there will be an equilibrium price $p_r^*(x_r, \xi_r, w_r, \omega_r)$ and variation in costs (w_r, ω_r) also contribute to the "exogenous variation" in price enables us to identify (α, β). So one way to identify (α, β) is BLP under their moment restrictions $E\{\xi|x, w\} = E\{\omega|x, w\} = 0$.
- In BLP's original paper, they analyzed the US auto market as divided into $R = 20$ regional markets. However they relied *large market asymptotics* since there were $N = 100$ types of cars sold in each of these markets.
- So they conjectured that some sort of law of large numbers and central limit theorem might hold as the total number of products $N \to \infty$. But Bertrand-Nash equilibrium leads to correlation since all N prices p are set in a common market equilibrium and hence equals $p(x, \xi, w, \omega)$ in each market. So prices of different products p_j and p_i should be correlated and will not be independent.

"Large market asymptotics" of the BLP estimator

- Armstrong (2016) studied the asymptotics of the BLP estimator when R is treated as fixed but N is large, so as $N \to \infty$.
- He notes that "Since the BLP instruments are correlated with prices only through equilibrium markups, their validity in this setting depends crucially on the nature of competition in markets with many products. If the dependence of markups on characteristics of other products disappears as the number of products increases and does so quickly enough, the BLP instruments will lose power in large markets and estimates based on them will be inconsistent when asymptotics are taken with respect to the number of products per market."
- "I find that, in certain cases, the dependence of prices on product characteristic instruments through markups disappears at a fast enough rate that the BLP instruments lead to inconsistent estimates when asymptotics are taken in the number of products per market. In particular, this is the case with the logit and random coefficients logit models when the number of products increases with the number of markets and products per firm fixed." Based on this, we focus on asymptotics as $R \to \infty$ treating N fixed.

Relaxing BLP's Moment Restriction $E\{\xi|x, w\} = 0$

- Are there other ways to relax the restriction $E\{\xi|x, w\} = 0$?
- Yes, but they require even more explicit *parametric* modeling of the unobservables, (ξ, ω).
- Suppose we do as previously suggested, *assume a parametric probability distribution for* (ξ, ω). That is, let $f(\xi, \omega|x, w, \rho)$, be a joint density over (ξ, ω) that depends on the observed (x, w) and a vector of parameters ρ to be estimated.
- We sketch two different approachs for estimating the parameters $\theta = (\alpha, \beta, \rho)$:
  1. maximum likelihood, or
  2. method of moments

  where, like BLP, both methods use "market cross section data" $\{s_r, p_r\}$ on markets $r = 1, \ldots, R$.
- We will show these methods extend to the case of random coefficients, both on β as well as allowing heterogeneity in the price parameter, α. But to start simple, suppose all consumers havae the same (α, β) parameters.

Maximum Likelihood, relaxing $E\{\xi|x,w\}=0$

- A continuous joint density $f(\xi,\omega|x,w,\rho)$ induces a continuous conditional distribution over prices and market shares, $(p^*(x,\xi,w,\omega,\theta^*), s(x,\xi,p^*(x,\xi,\omega),\theta^*))$, that depends on the parameters $\theta^*=(\alpha^*,\beta^*,\rho^*)$, assuming we have correctly specified the model and so there are *R IID* equilibria in the cross section of markets, each with its own unobserved (ξ_r,ω_r) from $f(\xi,\omega|x_r,w_r,\rho^*)$ and observed (x_r,w_r) from some other distribution $g(x,w)$ that we don't need to specify. We assume that this is the "true data generating process".
- Suppose we can calculate the Bertrand-Nash equilibrium (and in case of multiple equilbria specify a measurable mapping to select one of them for each (x,w,ξ,ω) and derive the implied joint density $h(p,s|x,w,\theta)$ implied by the speification above. Then let $\hat{\theta}$ be the maximum likelihood estimator

$$\hat{\theta}=\underset{\theta}{argmax}\sum_{r=1}^{R}\log\left(h(p_r,s_r|x_r,w_r,\theta)\right). \tag{86}$$

MSM, relaxing $E\{\xi|x,w\} = 0$

- Maximum likeihood may be infeasible since it may be difficult to derive the induced density, $h(p,s|x,w,\theta)$, implied by the model and joint density over unobserved shocks $f(\xi,\omega|x,w,\rho)$.
- It might be possible to do *simulated empirical likelihood* by using a bivariate kernel density estimated using many simulations $\{(p_j,s_j)|x_r,w_r\}$ of the model for each observed market r and value of θ. Using this we can calculate an "simulated empirical likelihood" $\hat{h}(p,s|x_r,w_r,\theta)$ that could be used to estimate θ^*.
- But it may be far easier to use McFadden's *Method of Simulated Moments* (MSM) estimator.
- The MSM estimator is based on the following two *conditional moment restrictions.* If the model is correctly specified then in each market r the observed price and market shares are $\tilde{p}_r = p^*(x_r,\xi_r,w_r,\omega_r,\theta^*)$ and $\tilde{s}_r = s(x_r,\xi_r,p_r,\theta^*)$ for some unobserved realized (ξ_r,ω_r). So at $\theta = \theta^*$ we have

$$\begin{aligned} E\left\{\tilde{p} - p^*(x,\tilde{\xi},w,\tilde{\omega},\theta)|x,w\right\} &= 0 \\ E\left\{\tilde{s} - s^*(x,\tilde{\xi},w,p^*(x,\tilde{\xi},w,\tilde{\omega}),\theta)|x,w\right\} &= 0. \end{aligned} \tag{87}$$

A MSM estimator that relaxes $E\{\xi|x,w\}=0$

- Let $H(x,w)$ be a $K \times 2N$ matrix valued, bounded measurable function of (x,w) (so its moments are finite). Define the $2N \times 1$ vector of residuals $(\epsilon_s(\theta), \epsilon_p(\theta))'$ where N is the number of products (assumed to be the same in all markets in the sample), defined by

$$\begin{aligned} \epsilon^s(\theta) &= s - s^*(x,\xi,p^*(x,\xi,w,\omega,\theta),\theta) \\ \epsilon^p(\theta) &= p - p^*(x,\xi,w,\omega,\theta). \end{aligned} \tag{88}$$

- So if (p_r, s_r) are the *actual* prices and market shares in market r, and (x_r, w_r) are the observed product and producer characteristics, then for a simulated draw of unobservables (ξ_r, ω_r) we can calculate, using the model, simulated prices and market shares $s^*(x_r,\xi_r,p^*(x_r,\xi_r,w_r,\omega_r,\theta),\theta)$ and $p^*(x_r,\xi_r,w_r,\omega_r,\theta)$, and hence the corresponding residuals $\epsilon_r(\theta) = [\epsilon_r^s(\theta), \epsilon_r^p(\theta)]$.
- Define the $2N \times 2N$ conditional covariance matrix $\Omega(x,w)$ by

$$\Omega(x,w) = E\left\{\epsilon_r(\theta^*)\epsilon_r(\theta^*)'|x,w\right\}. \tag{89}$$

Defining the MSM estimator

- The conditional moment restrictions (87) imply the following $K \times 1$ *unconditional* moment restrictions

$$E\{H(x,w)\epsilon_r(\theta^*)\} = 0. \tag{90}$$

- Now, suppose we draw S realizations (ξ_{sr}, ω_{sr}), $s = 1, \ldots, S$ from $f(\xi, \omega | x_r, w_r, \rho)$ for each market $r = 1, \ldots, R$ in our sample. Using these, and the observed (x_r, w_r) we can compute simulated prices and market shares in each market r and hence simulated residuals, $\epsilon_{sr}(\theta)$. Define the corresponding $K \times 1$ *simulated moments* $\overline{H}_R(\theta)$ by

$$\overline{H}_{SR}(\theta) = \frac{1}{R}\sum_{r=1}^{R}\left[\frac{1}{S}\sum_{s=1}^{S} H(x_r, w_r)\epsilon_{sr}(\theta)\right]. \tag{91}$$

- Let W be a $K \times K$ positive definite weighting matrix. Then we can define $\hat{\theta}_{SR}$, the MSM estimator, as

$$\hat{\theta}_{SR} = \underset{\theta}{argmin}\, \overline{H}_{SR}(\theta)' W \overline{H}_{SR}(\theta). \tag{92}$$

Asymptotics of the MSM estimator

- Assume that observations from different markets are *IID.* This may not be a good assumption in many cases, but we do it for simplicity to start.
- Using a Taylor series expansion, we can show that as $R \to \infty$, the normalized MSM estimator, $\sqrt{R}[\hat{\theta}_{SR} - \theta^*]$, converges in distribution to a normal distribution, $\sqrt{R}[\hat{\theta}_{SR} - \theta^*] \Longrightarrow N(0, \Lambda(W))$ where $\Lambda(W) = \Lambda_1(W)^{-1})\Lambda_0(W)\Lambda_1(W)^{-1}$ where

$$\begin{aligned} \Lambda_1(W) &= [\nabla H(\theta^*)' W \nabla H(\theta^*)] \\ \Lambda_0(W) &= \left(1 + \frac{1}{S}\right)[\nabla H(\theta^*)' W [E\{H\Omega(\theta^*)H'\}] W \nabla H(\theta^*)] \end{aligned} \tag{93}$$

where

$$\begin{aligned} E\{H\Omega(\theta^*)H'\} &= E\left\{H(x,w)\epsilon(\theta^*)\epsilon(\theta^*)'H(x,w)'\right\} \\ \nabla H(\theta)^* &= E\left\{H(x,w)\nabla\epsilon(\theta^*)\right\}. \end{aligned} \tag{94}$$

- Given the choice of moments, H, the optimal weighting matrix is

$$W = [E\{H\Omega(\theta^*)H'\}]^{-1}. \tag{95}$$

Comments on the MSM estimator

- MSM relaxes the $E\{\xi|x,w\}=0$ restriction, but at the cost of having to make parametric assumptions about $f(\xi,\omega|x,w,\rho)$.
- It is more computationally intensive than BLP, since we must solve for equilibrium prices and market shares (p_r,s_r) in each market $r=1,\ldots,R$ each time θ is updated. Special care must be taken in case a market has *multiple equilibria* so a sensible *equilibrium selection rule* should be employed.
- The beauty of MSM is it needs only *a single simulated draw* (ξ_r,ω_r) *per market* since "simulation noise" averages out like sampling error as $R\to\infty$, so the penalty for doing S draws is to increase asymptotic parameter variance by a factor $(1+\frac{1}{S})$.
- We have a *smooth simulator* since $\overline{H}_{SR}(\theta)$ is a smooth function of θ. For $S=1$, pick $\{u_{1r},u_{2r}\}$ *common random uniform draws* at the start of the estimation *and keep them fixed throughout the search for* $\hat{\theta}_R$. Factor the joint CDF as $F(\xi,\omega|x,w,\rho)$ as $F_2(\xi|\omega,x,w,\rho)F_1(\omega|x,w,\rho)$ and apply the probability integral transform sequentially

$$\omega_r=F_1^{-1}(u_{1r}|x_r,w_r,\rho)\quad \xi_r=F_2^{-1}(u_{2r}|\omega_r,x_r,w_r,\rho). \tag{96}$$

Allowing for random coefficients heterogeneity

- Returning to BLP's original approach under the assumption $E\{\xi|x, w\} = 0$, let's now consider how to extend Berry 1994 to the case where $\beta \sim N(\overline{\beta}, \Sigma)$. Note *BLP assume there is no heterogeneity in* α.
- BLP write the realized random coefficient for consumer i as $\beta_i = \overline{\beta} + \nu_i$ where $\nu_i \sim N(0, \Sigma)$. Thus, the utility of consumer i from purchasing item j, u_{ij} is given by

$$u_{ij} = x_j\overline{\beta} - \alpha p_j + (x\nu_{ij} + \epsilon_{ij}), \tag{97}$$

  which is nearly the same expression as the utility $u_{ij} = x_j\beta - \alpha p_j + \epsilon_{ij} = \delta_j + \epsilon_{ij}$ for the non-random coefficient case, except that now we have a "composite" error term $\eta_{ij} = x\nu_{ij} + \epsilon_{ij}$ and the pattern of correlation in these error terms result in more flexible patterns of cross-substitutions, and thus more realistic substitution patterns in the derived market shares, equation (53).
- So we can write

$$u_{ij} = \delta_j + \eta_{ij} \tag{98}$$

Hotz-Miller Inversion to the Rescue!

- Notice the utility representation (98) is just another random utility model, so it will have choice probabilities of the form $P(j|\delta_1, \ldots, \delta_N)$, $j = 1, \ldots, N$. We can write this more compactly as $P(\delta)$. Suppose the vector of observed market shares is the $N \times 1$ vector s, then the Hotz-Miller Inversion Theorem guarantees the existence of a unique $N \times 1$ vector δ that solves the equation

$$P(\delta) = s. \tag{99}$$

- How do we numerically invert this equation to get the δ implied by s? One obvious way is *Newton's Method* for solving the nonlinear system $P(\delta) = s = 0$ for $\delta = P^{-1}(s)$. Starting from any initial guess δ_0 (e.g. $\delta_0 = 0$) Newton's method generates a sequence $\{\delta_t\}$ given by

$$\delta_{t+1} = \delta_t - [\nabla P(\delta_t)]^{-1} [P(\delta_t) - s], \tag{100}$$

- Problem: unless δ_0 is in a *domain of attraction* of the true solution $\delta = P^{-1}(s)$, Newton iterations (100) may not coverge to δ.

Contraction-Mapping Theorem to the Rescue!

- Fortunately, BLP proved the following result for the case where $P(\delta)$ is given by the mixed logit model, (53). Define the mapping $T : R^N \to R^N$ by

$$T(\delta) = \delta + \log(s) - \log(P(\delta)) \tag{101}$$

  where $\log(s)$ is the element by element logrithm of the market share vector s, and similarly for $\log(P(\delta))$.
- Notice that $T(\delta) = \delta$ (i.e. T has a fixed point) if and only if $s = P(\delta)$. BLP proved that T is a *contraction mapping,* i.e. there exists a scalar $\lambda \in (0, 1)$ such that for any $\delta, \delta' \in R^N$ we have

$$\|T(\delta') - T(\delta)\| \leq \lambda \|\delta' - \delta\| \tag{102}$$

Contraction Mapping Theorem

If T is a contraction mapping it has a unique fixed point $T(\delta) = \delta$ which can be computed by the method of successive approximations $\delta_{t+1} = T(\delta_t)$ starting from any initial guess $\delta_0 \in R^N$.

Combining the best of both worlds

- While successive approximation iterations for a contraction mapping T *always converges from any starting point* δ_0, the rate of convergence is only *geometric* i.e. we have

$$\|\delta_{t+1} - \delta_t\| = \|T(\delta_t) - T(\delta_{t-1})\| \leq \lambda \|\delta_t - \delta_{t-1}\| \tag{103}$$

  so if λ is close to 1, successive approximations can be quite slow. However once successive approximations results in a δ_t sufficiently close to the true fixed point $\delta = P^{-1}(s)$ that it gets into a *domain of attraction* we can switch to Newton's method since these iterations converge *quadratically*

$$\|\delta_{t+1} - \delta_t\| \leq K(T)\|\delta_t - \delta_{t-1}\|^2 \tag{104}$$

  where $K(T)$ is a positive constant that depends on the curvature of the mapping T.
- So a better algorithm is a *polyagorithm*
  1. First start with contraction iterations $\delta_{t+1} = T(\delta_t)$
  2. Then switch to Newton iterations to rapidly converge to the solution, $\delta_{t+1} = \delta_t - [\nabla P(\delta_t)]^{-1}[P(\delta_t) - s]$

Calculating the modulus λ of T

- Following the proof of the contraction mapping property of T in the appendix of BLP, we have that the *modulus* of T, λ is given by

$$\lambda = \max_{j=1,\ldots,N} \sup_{\delta} \left[1 - \frac{\int_\nu P_j(\delta + x\nu) P_0(\delta + x\nu) \phi(d\nu|\Sigma)}{\int_\nu P_j(\delta + x\nu) \phi(d\nu|\Sigma)} \right] \tag{105}$$

where $\phi(\nu|\Sigma)$ is the mutivariate normal distribution of the $K \times 1$ vector of random coefficient residuals ν, and $P_j(\delta + x\nu)$ is the MNL "kernel" of the mixed logit model

$$P_j(\delta + x\nu) = \frac{\exp\{\delta_j + x_j\nu\}}{1 + \sum_{k=1}^{N} \exp\{\delta_k + x_k\nu\}} \tag{106}$$

and $P_0(\delta + x\nu)$ is the probability of choosing the outside good, given (δ, x, ν).

- Thus, if the predicted probability of the outside good $P_0(\delta + x\nu)$ is large in a neighborhood of the fixed point $\delta = P^{-1}(s)$, then λ will be small and successive approximations using T should converge quickly.

How does mixed logit affect BLP?

- Unlike the simpler "homogeneous logit" case, when we have a mixed logit choice probability model for market shares, the Hotz-Miller inversion will now depend on the parameters Σ of the normal mixing distribution.
- Specifically, using the expression for the "MNL kernel" (106) we get the following expression for market share s_j

$$s_j = \int_\nu \frac{\exp\{\delta_j + x_j\nu\}}{1 + \sum_{k=1}^{N} \exp\{\delta_k + x_k\nu\}} \phi(d\nu|\Sigma) \tag{107}$$

- So we can stack the market shares and write the market share inversion equation as

$$s = P(\delta, x, \Sigma) \tag{108}$$

and so by the Hotz-Miller inversion theorem the inverse $\delta(x, \Sigma) = P^{-1}(s, x, \Sigma)$ exists, *but is an implicit function of the observed characteristics x and the covariance matrix* Σ.

- With mixed logit, *we must repeatedly invert (108) each time the "outer" optimization algorithm changes* Σ.

Nested numerical solution version of BLP

- We can define the residuals for each market r, $\epsilon_r(\theta)$ as for the homogeneous MNL case

$$\epsilon_r(\theta) = [\delta_r(s_r, x_r, \Sigma) - [x_r\overline{\beta} - \alpha p_r], \log(c_r) - w_r\gamma]' \tag{109}$$

- Using these residuals, an $L \times 2N$ matrix function $H(x, w)$ for the relevant moments and an $L \times L$ weighting matrix W the "outer" GMM minimization problemm is

$$\hat{\theta}_R = \underset{\theta}{argmin}\, \overline{H}_R(\theta)' W \overline{H}_R(\theta) \tag{110}$$

  where $\overline{H}_R(\theta)$ is the $L \times 1$ vector of moments given by (67).
- So in contrast to the non-mixed logit case where the δ_j only need to be calculated only once at the start of the optimization process, for mixed logit, *each time we evaluate* $\overline{H}_R(\theta)$ *we must invert the market shares* $s_r = P(\delta_r, x_r, \Sigma)$ *in each of the R markets in the sample.* This is why BLP with mixed logit can be computationally intensive.

MPEC: A faster way to do BLP?

- Instead of the "nested fixed point" (NFP) algorithm BLP proposed to compute $\hat{\theta}_R$, Dubé, Fox and Su (2012) suggest *Mathematical Programming with Equilibrium Constraints.*
- Instead of numerically inverting $s_r = P(\delta_r, x_r, \Sigma)$ for all $r = 1, \ldots, R$ each time Σ changes, MPEC solves the *constrained minimization program* to calculate $\hat{\theta}_R$

$$\hat{\theta}_R = \underset{\theta}{\operatorname{argmin}} \overline{H}_R(\theta)' W \overline{H}_R(\theta) \tag{111}$$

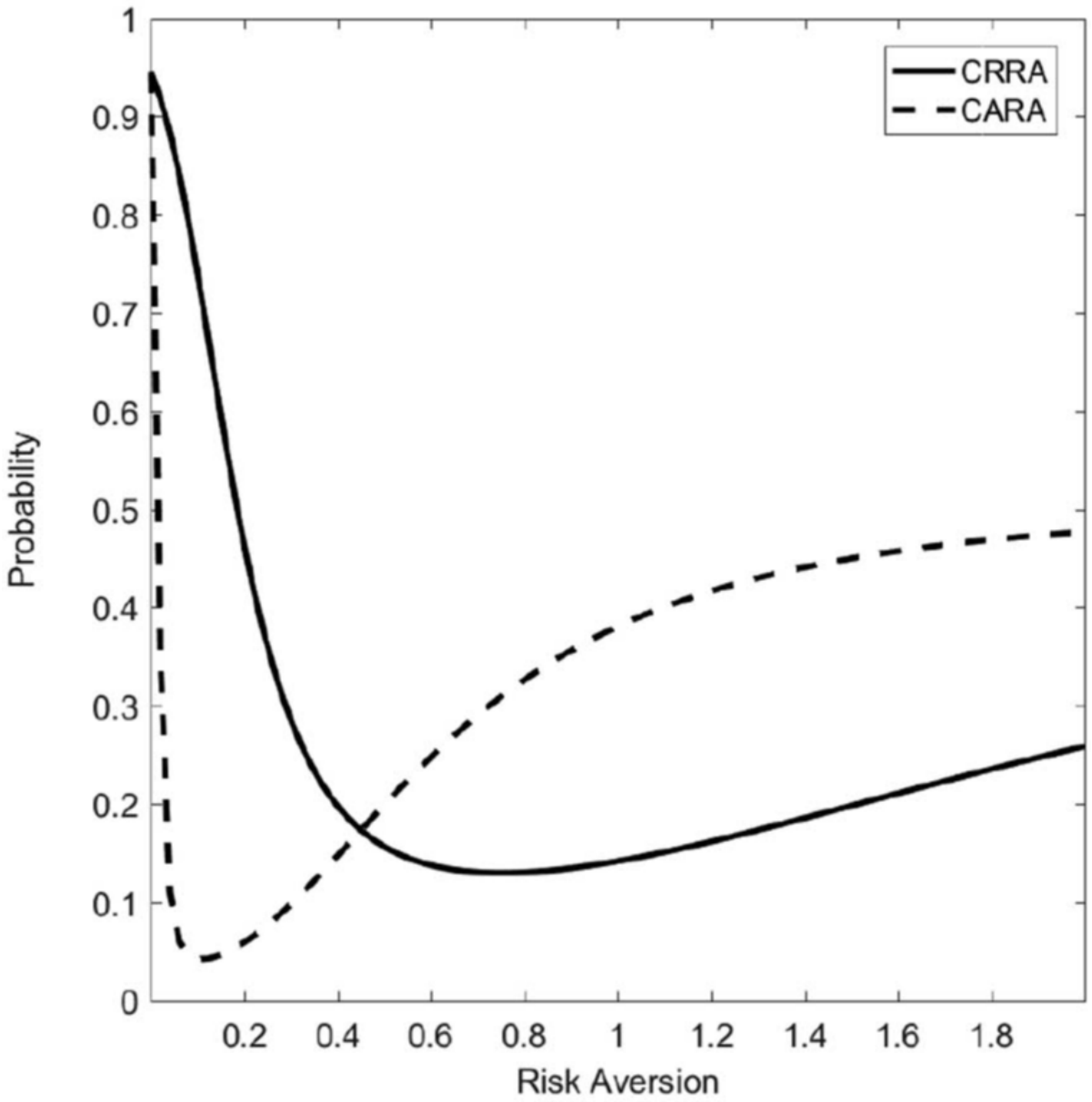
subject to:

$$s_r = P(\delta_r, x_r, \Sigma), \quad r = 1, \ldots, R. \tag{112}$$

Dubé, Fox and Su's conclusion

MPEC produces good estimates relatively quickly for most of the data-generating processes that we consider. Its speed is invariant to the Lipschitz constant of the inner-loop contraction mapping used in NFP, as expected. In the case of a very dense, large-dimensional problem with few markets and many products, we lose MPEC's speed advantage over NFP.

Another problem: non-monotonic choice probabilities



Discrete choice over gambles

- Figure above taken from Apesteguia and Ballester 2018 *JPE* "Monotone Stochastic Choice Models: The Case of Risk and Time Preferences"
- They consider an experiment where a risk averse subject choose between two gambles:
  - d_x is a gamble that pays \$1 with probability $p = .9$ and \$60 with probability $1 - p = .1$
  - d_y is the choice of receiving \$5 with probability 1.
- Assuming CRRA utility $u(c) = c^{1-\theta}/(1-\theta)$, the expected utilities of these two gambles are

$$\begin{aligned} E\{u(d_x)\} &= .9\frac{1^{1-\theta}}{1-\theta} + .1\frac{60^{1-\theta}}{1-\theta} \\ E\{u(d_y)\} &= \frac{5^{1-\theta}}{1-\theta} \end{aligned}$$

- Why the non-monotoncity in θ? Higher θ corresponds to more risk averse so we should expect the probability of choosing d_x should decrease in θ

What's going on?

- With scale value σ the logit RUM model predicts $P(d_x|\theta)$ to be

$$P(d_x|\theta) = \frac{\exp\{E\{u(d_x)\}/\sigma\}}{\exp\{E\{u(d_x)\}/\sigma\} + \exp\{E\{u(d_y)\}/\sigma\}}$$

- Note that as $\theta \to \infty$ we have

$$\lim_{\theta\to\infty} E\{u(d_x)\} = \lim_{\theta\to\infty} E\{u(d_y)\} = 0$$

  which implies that

$$\lim_{\theta\to\infty} P(d_x|\theta) = \frac{1}{2}$$

- So even though a risk neutral person prefers d_x with high probability (i.e. when $\theta = 0$) and the choice probability initially declines in θ (as risk aversion increases), the RUM specification implies the counterintuitive non-monotonicity in the choice probability when θ gets sufficiently large and choice probabity converges to $1/2$.
- Is this a problem with RUM or should there be a correlation between σ and θ so $\sigma \to 0$ as $\theta \to \infty$?

RPM vs RUM

- Apesteguia and Ballester show that "every RUM using either CRRA or CARA utilities, and not exclusively the logit RUM, violates monotonicity, and this happens for every pair of gambles in which one is riskier than the other, showing that the problem is ubiquitous."
- But should we conclude that "that their use in preference estimation may be problematic. They may pose identification problems and could yield biased estimations. We then establish that the alternative random parameter models are always monotone." ?
- The RPM assumes away the additive preference shocks $\epsilon(d)$ but assumes a distribution $f(\theta)$ over risk aversion. Then we have

$$P(d_x) = \int_\theta P(d_x|\theta) f(\theta) d\theta$$

where

$$P(d_x|\theta) = I\{E\{u(d_x, \theta)\} \geq E\{u(d_y, \theta)\}\}$$

Combining RPM and RUM

- One implication of RPM is that unless the realization of θ changes from choice to choice, it will predict perfect correlation between successive risky choices, which may be easily violated such as in experimental data settings. But the idea of *IID* variation in risk aversion in choices only seconds apart does not seem appealing either. Is there any compromise?
- Why not the idea of mixed logit similar to BLP? We can also parameterize σ as a function of θ so $\sigma(\theta)$ tends to 0 as $\theta \to \infty$

$$P(d_x) = \int_\theta P(d_x|\theta) f(\theta)$$

where $P(d_x|\theta)$ is the logit choice probability above.
- This sort of compromise solution can share the monotonity advantages of RPM but (by assuming that θ is fixed across choices but $\epsilon(d)$ vary across choices) allows more flexibility and better fit to the data, since it reduces the perfect correlation across successive choices of gambles that the pure RPM model would predict (unless we allow *IID* variation in θ across successive choices)

Application of RPM: Tomás Jagelka

- His paper, "Are Economist's Preferences Psychologist's Personality Traits?" analysed responses to hypothetical gambles and time preference questions posed to 1224 Canadian high school students
- The questions ask the students to evaluate a sequence of *hypothetical gambles* and choose the gamble they preferred. These gambles are similar to the ones analyzed by Apesteguia and Ballester above.
- Tomás used the pure RPM framework (no RUM) and estimated *subject-specific distributions* for $f(\theta)$ which we denote by $f_i(\theta)$ where i denotes subject i. Tomás assumed that $\theta_i \sim N(\mu_i, \sigma_i^2)$ where the parameters (μ_i, σ_i^2) are subject-specific.
- He calculates a threshold for each gamble pair (x, y), call it $\theta(x, y)$ which is a risk aversion level that makes a subject with risk aversion parameter $\theta(x, y)$ indifferent between gambles x and y. Then if x is the riskier gamble, we have

$$P(d_x|\mu_i, \sigma_i^2) = \int_\theta I\{\theta \leq \theta(x, y)\} f_i(\theta) d\theta = \Phi\left(\frac{\theta(x, y) - \mu_i)}{\sigma_i}\right)$$

Adding "trembles" to explain choices

- In addition, Tomás allows for *trembles* — he assumes a subject makes a mistake and chooses his/her preferred gamble only with probability λ_i. So the overall likelihood for subject *i* is

$$P(d_x|\mu_i, \sigma_i^2, \lambda_i) = (1 - \lambda_i)P(d_x|\mu_i, \sigma_i^2) + \lambda_i[1 - P(d_x|\mu_i, \sigma_i^2)]$$

- What does he find? First analyzing the 1224 subject-specific estimates $(\mu_i, \sigma_i^2, \lambda_i)$ he finds the mean risk aversion is $\frac{1}{N}\sum_{i=1}^{1224} \hat{\mu}_i = 1.14$ and $\frac{1}{N}\sum_{i=1}^{1224} \hat{\sigma}_i = 0.59$ so there seems to be big trial to trial variability in risk aversion. The average rate of trembles is $\frac{1}{N}\sum_{i=1}^{1224} \hat{\lambda}_i = 0.05$.
- He then regresses these subject-specific responses on observable characteristics including self-reported *personality traits* and *unobserved types.* He finds that "there is significant heterogeneity in risk and time preferences in the population and also in their individual-level stability."

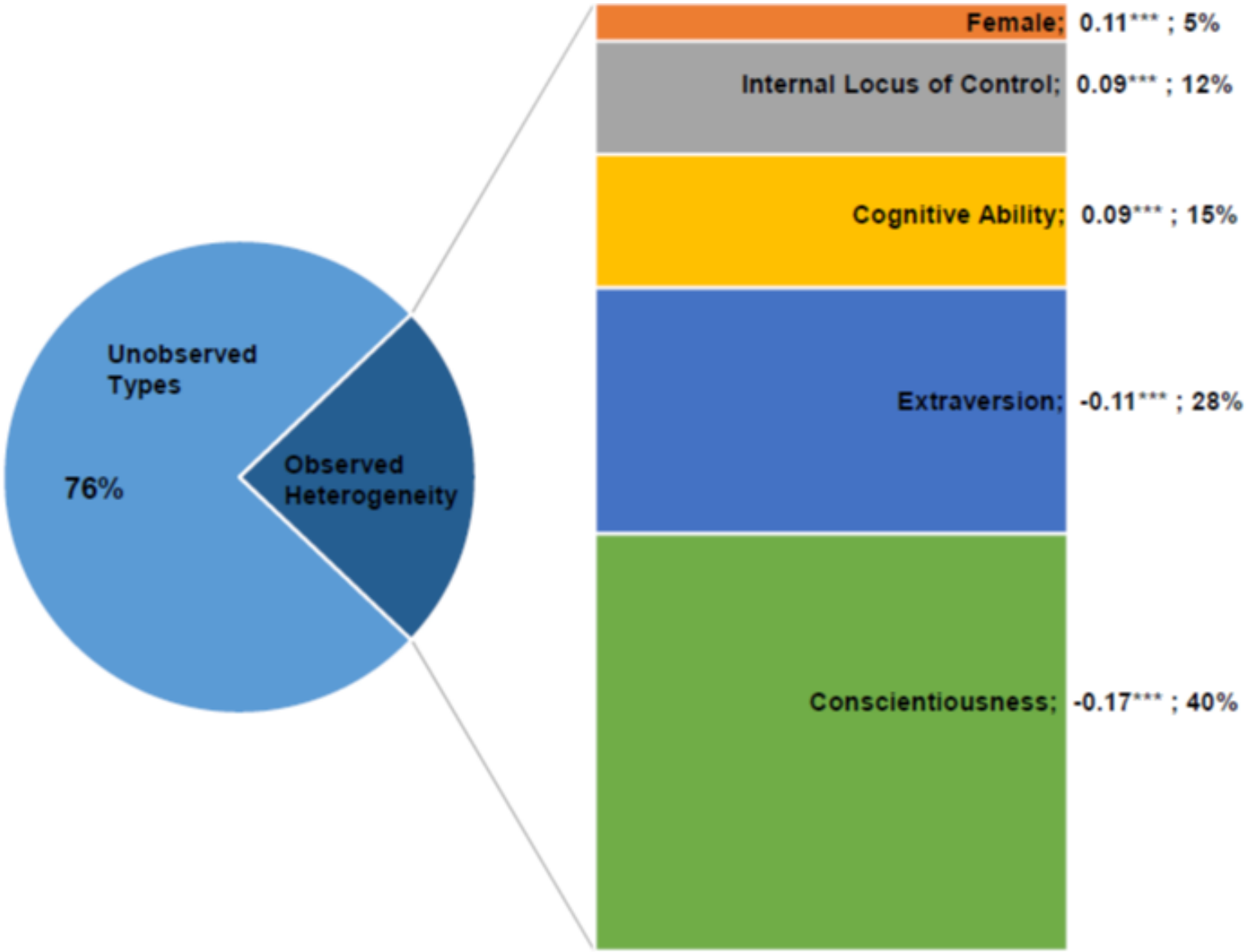
Explaining individual level heterogeneity

Jagelka used factor analysis and unobserved types to try to see how much of the variability in preferences he could "explain."

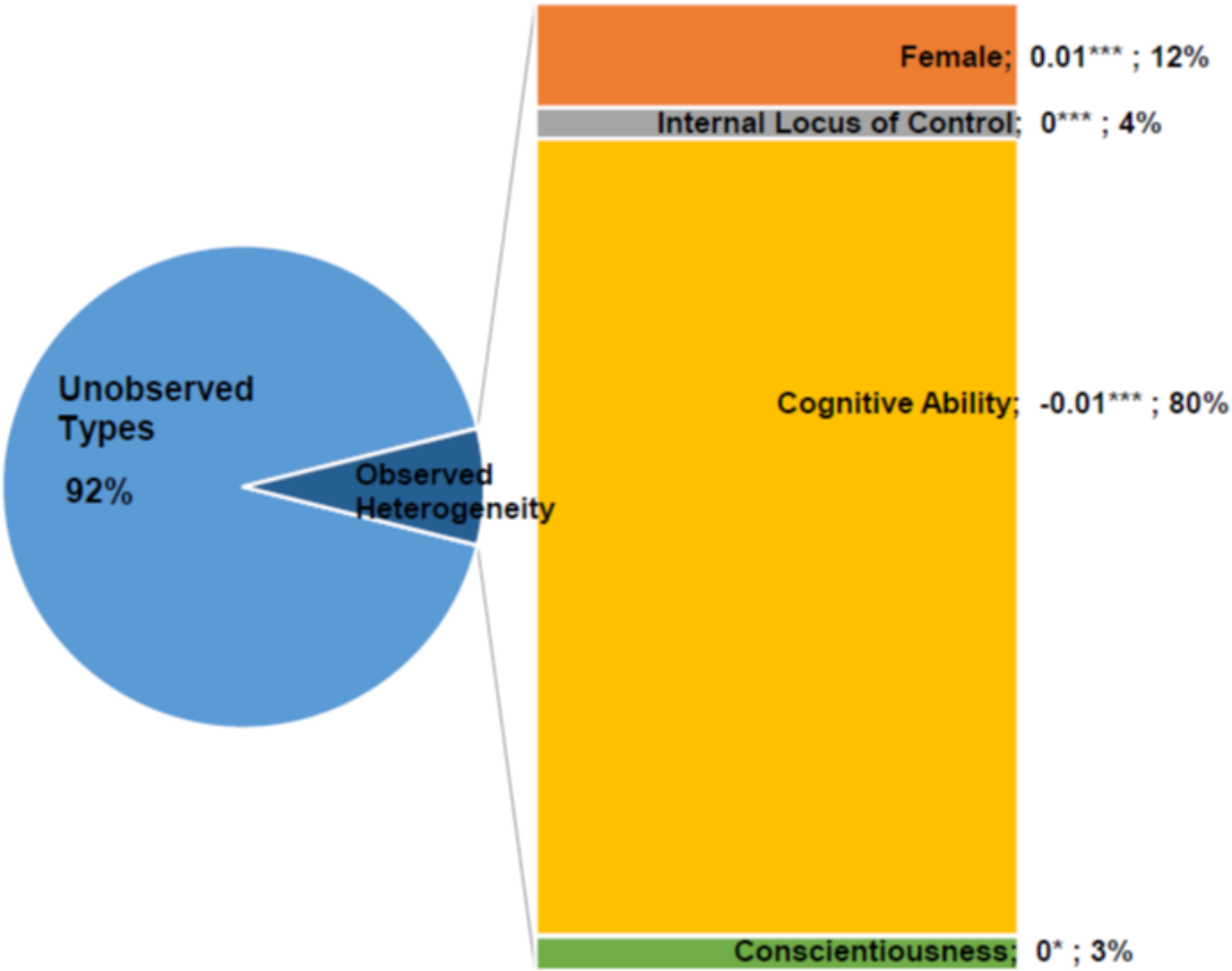
Jagelka's conclusion

Depending on the parameter in question, up to 50% of the variation in risk aversion, discount rates, and parameters governing their stability and individuals' rationality can be explained by cognitive ability and personality traits. Conscientiousness is the trait with the highest overall explanatory power, in line with previous results on the predictive potential of personality traits on real-world outcomes. It explains 45% of the cross-sectional variation in discount rates, 10% of the variation in risk aversion, and 20% of the variation in their individual-level stability. Nevertheless, individuals' preferences, their stability, and people's propensity to make mistakes remain to a large part a function of unobserved heterogeneity. One can thus conclude that economists' preferences and psychologists' personality traits are related but distinct concepts.

Breakdown of heterogeneity in risk aversion



Breakdown of heterogeneity in trembles



Another application of trembles: El-Gamal and Grether

- "Are People Bayesian? Uncovering Behavioral Strategies" *JASA* 1995
- Analyzed decision making experiment on 257 college students in California. Each subject was asked to guess which of two bingo cages a sample of 6 balls were drawn from (with replacement)
- Cage A had 4 N balls and 2 G balls, and Cage B had 3 N balls and 3 G balls. Priors of drawing each cage were communicated in a credible, intuitive way. Some subjects were paid if they made correct guesses, others were not paid.
- There are 2^{21} possible decision rules for assigning a guessed cage for each possible realized sample. El-Gamal and Grether restricted to attention to a smaller set of 512 *cutoff rules* choose cage A if there are at least $\overline{c}$ N balls in the sample. Three priors were announced: $\pi = 1/3, 1/2, 2/3$ of choosing cage A.
- Bayes rule corresponds to cutoffs of $\overline{c} = 4$, $\overline{c} = 3$ and $\overline{c} = 2$ N balls for the 3 priors, respectively.

Trembles in El-Gamal and Grether

- Let $c = (\overline{c}_1, \overline{c}_2, \overline{c}_3)$ be a cutoff rule (possibly not Bayes rule). Let X_c^s be the number of trials for subject s that are consistent with cutoff rule c, so the other $t_s - X_c^s$ trials the subject chose something else not predicted by the cutoff rule c. These are considered to be "errors" or "trembles"
- Let λ be a common error rate: then the likelihood for subject s for cutoff rule c is

$$f(t^s, X_c^s, c) = (1-\lambda)^{X_c^s} \lambda^{t_s - X_c^s}$$

- Then the maximum likelihood estimates are $(\hat{\lambda}, \hat{c})$ are given by

$$(\hat{\lambda}, \hat{c}) = \underset{\lambda, c}{argmax} \prod_{s=1}^{S} f(t^s, X_c^s, c)$$

- Estimated multiple types of decision rules used by different subjects via an "estimation-classification" algorithm (EC-algorithm)

Results

- Error rates overall were $\hat{\lambda} = .10$ for the for-pay subjects, and $\hat{\lambda} = .15$ for the no-pay subjects.
- Bayes Rule was the most common cutoff rule used by nearly 50% of UCLA subjects and 30% of Pasadena Community College students. Subjects in the for-pay experiments were more likely to use Bayes rule than students in the no-pay experiments

El-Gamal and Grether's conclusion

The response of economists and psychologists to the discovery of anomalous violations of standard models of statistical decision theory has mainly been to devise new theories that can accommodate those apparent violations of rationality. The enterprise of finding out what experimental subjects actually do (instead of focusing on what they do not do; i.e., violations of standard theory) has not progressed to the point that one would hope. As a first step in that direction, we propose a general estimation/classification approach to studying experimental data. The procedure is sufficiently general in that it can be applied to almost any problem.

Summary of findings

El-Gamal and Grether;s conclusion

Our results seem robust, and the most prominent rules that our algorithm selected are reasonable rules. The most prominent rule in most cases is the Bayes updating rule. 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 the Bayes rule would imply. We believe that given the flexibility of our approach, and given the strong results that it generated in our particular application, its positive usefulness for uncovering the rules used by experimental subjects can be quite substantial.

Recent contributions to discrete choice modeling

- Matejka and McKay *AER* 2015 "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model"

From abstract

Individuals must often choose among discrete actions with imperfect information about their payoffs. Before choosing, they have an opportunity to study the payoffs, but doing so is costly. This creates new choices such as the number of and types of questions to ask. We model these situations using the rational inattention approach to information frictions. We find that the decision maker's optimal strategy results in choosing probabilistically in line with a generalized multinomial logit model, which depends both on the actions' true payoffs as well as on prior beliefs.

Two stage formulation of the discrete choice problem

- "The decision problem has two stages. In the first stage, the decision maker selects an information strategy to refine his belief about the state. The second stage is a standard choice under uncertainty with the beliefs generated in the first stage." Second stage problem is

$$V(\pi) = \int \max_{d \in D} v_d \pi(v_1, \ldots, v_D)$$

  where $\pi(v_1, \ldots, v_D)$ is the posterior probability over the utilities of the D choices the decision making is considering choosing, resulting from the the solution to first step "information gathering problem"
- First stage problem is to choose an information gathering strategy that results in signals s that the decision maker observes and uses to form a posterior distribution $\pi(v_1, \ldots, v_D|s)$ over the utilities of the choices. This is framed as a choice of joint distribution $\pi(v, s)$ over values and signals ($v = (v_1, \ldots, v_D)$) so that $\pi(v|s)$ is the posterior and $\pi(v) = \int_s \pi(v, s)ds$ is the decision maker's prior. Let $c(\pi)$ be a cost function for acquiring these signals.

First stage information acquistion problem

- The first stage problem is

$$\max_{\pi(s,v)} \int_v \int_s V(\pi(\cdot|s))\pi(s|v)\pi(v)dsdv - c(\pi(\cdot,\cdot))$$

- A generalized version of the MNL results when c is given by the entropy cost function

$$c(\pi) = \lambda[H(\pi) - \int_s H(\pi(\cdot|s))\pi(s|z)]$$

where

$$H(\pi) = -\int_z \log(\pi(v))\pi(v)dv$$

- While it is important to account for the "cost of information" this problem does not account for the "cost of computation" of solving the infinite-dimensional first stage optimization problem to determine the optimal first stage information gathering strategy. Progress can be made by more realistcally modeling how people actually gather information and engage in "mental evaluation of alternatives" when making decisions.

Second recent contribution: neural dynamics of choice

- Ryan Webb (2018) "The (Neural) Dynamics of Stochastic Choice" *Management Science*

From intro

This article proposes a neurobiological foundation to a random utility model, grounded explicitly in the prevailing framework for modeling the dynamics of decision making found in the psychology and neuroscience literature. Over a half century of psychometric research has established empirically the joint distribution between perceptual judgements and response time in a wide variety of sensory domains

- Thus, the focus is on the joint disribution of the item chosen and the time needed to make a decision, $Pr\{d^* = d, t* \leq t\}$ where d^* is the chosen alternative and t^* is the time to make a decision to choose d^* from a finite set of alternatives D.

Human brains are slow relative to computers

- Models of neural processing indicate that unlike digital computers, neurons are actually rather slow: they are "wet computers" and nerve impulses actually travel at relatively slow speeds down axons, and neurotransmitters do take time (milliseconds) to travel between synapes between axons and dendrites, the gaps over which different neurons communicate with each other

A neuron outputs discrete, electrical impulses, and this activity of individual neurons in cortex is highly irregular. This stochasticity arises, at least in part, from the small-scale thermodynamic processes involved in the flow of ions between neurons

- Webb models the state of decision making over n alternatives by a continuous time diffusion process $\{Z(t)\}$ where $Z(t)$ is an $n \times 1$ vector of the "current accumulations" (or evaluations) of the n alternatives while the individual is considering which of the alternatives to choose.

Bounded accumulation models of choice (BAM)

- **Accumulators** In an accumulation, each alternative d in the choice set is associated with a decision variable. In empirical practice, the decision variable is widely taken to be the activity level of a population of neurons associated with a given alternative
- $Z(t)$ evolves according to a vector-valued diffusion process

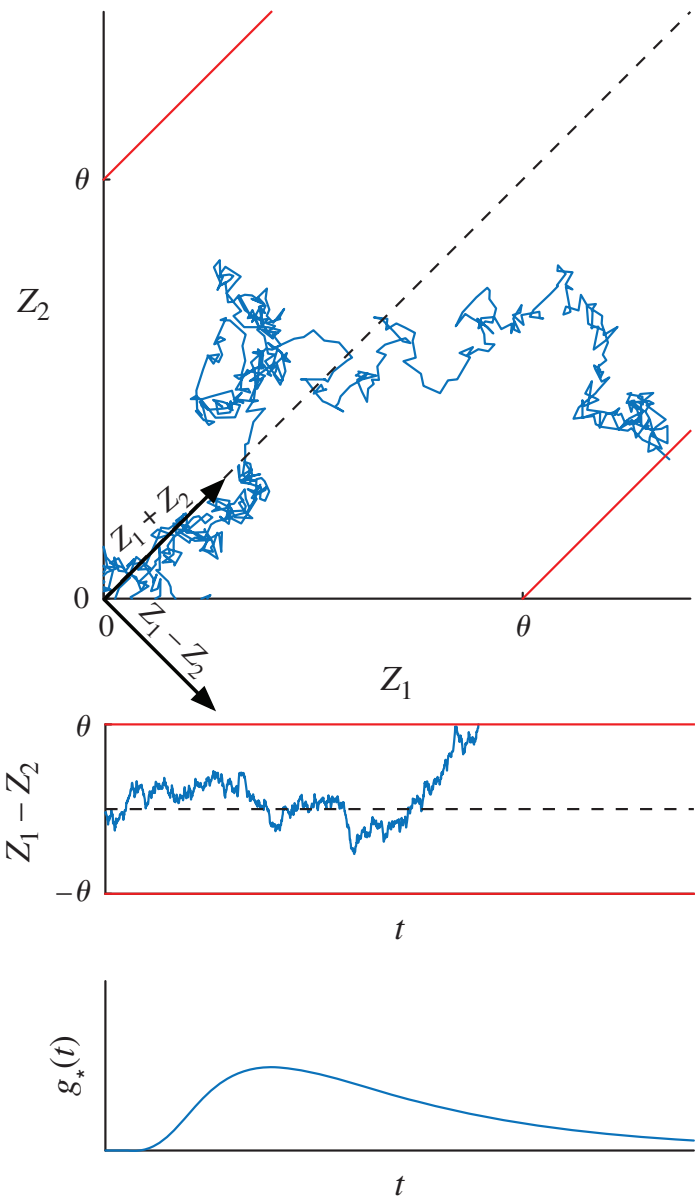
$$dZ(t; v) = [v + c(t) + \Gamma(t)Z(t; v)]dt + \sigma(t)dB(t)$$

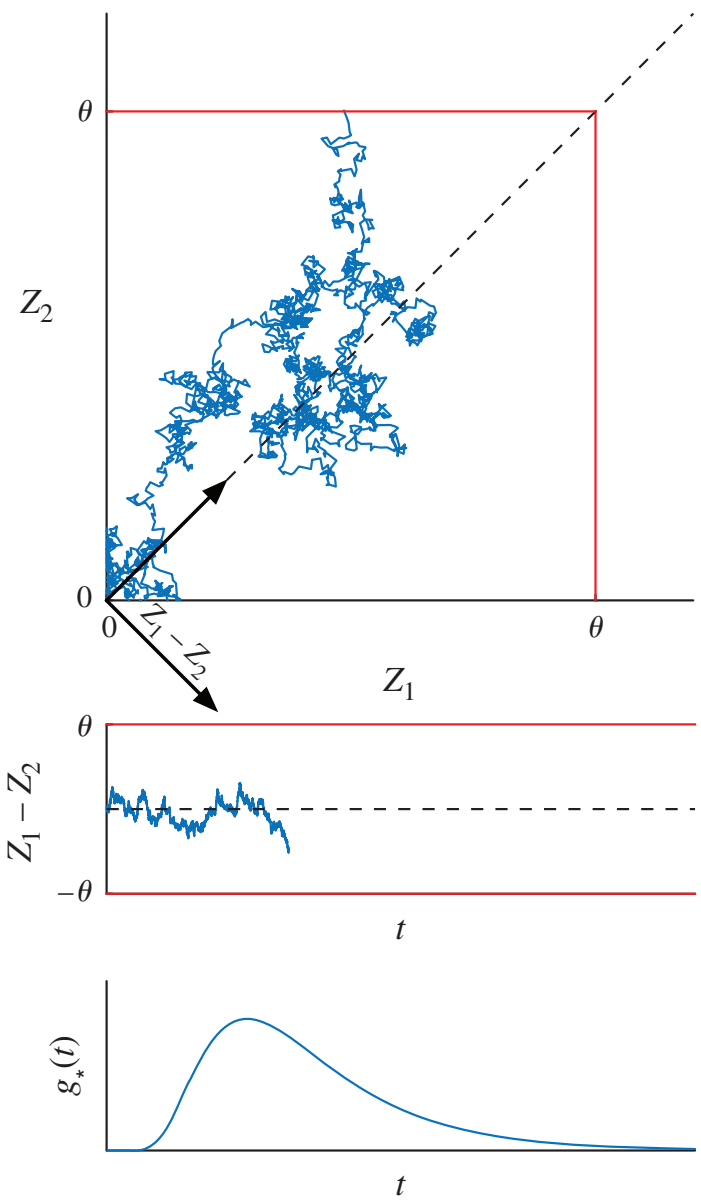
where $B(t)$ is an $n \times 1$ Brownian motion process and the additional drift term $c(t)$ "is included to capture a universal ramping of neural activity over the course of a decision, particularly observed in cases of time pressure, and is often termed an 'urgency' signal"

- **Decision times** The second element of a BAM is a stopping rule that determines when/whether each accumulator (or a function of each accumulator) has reached a decision threshold. In practice, the stopping rule provides a functional equation for the response time t^*.

$$t^* = \inf \{t | Z(t; v) \notin C(t)\}$$

Simulations of accumulators and decisions





Choice criterion: highest value at the stopping time

- The choice is the largest accumulator at the decision stopping time t^*

$$d^* = \underset{d}{argmax}\{Z_d(t^*; v)\}$$

Theorem

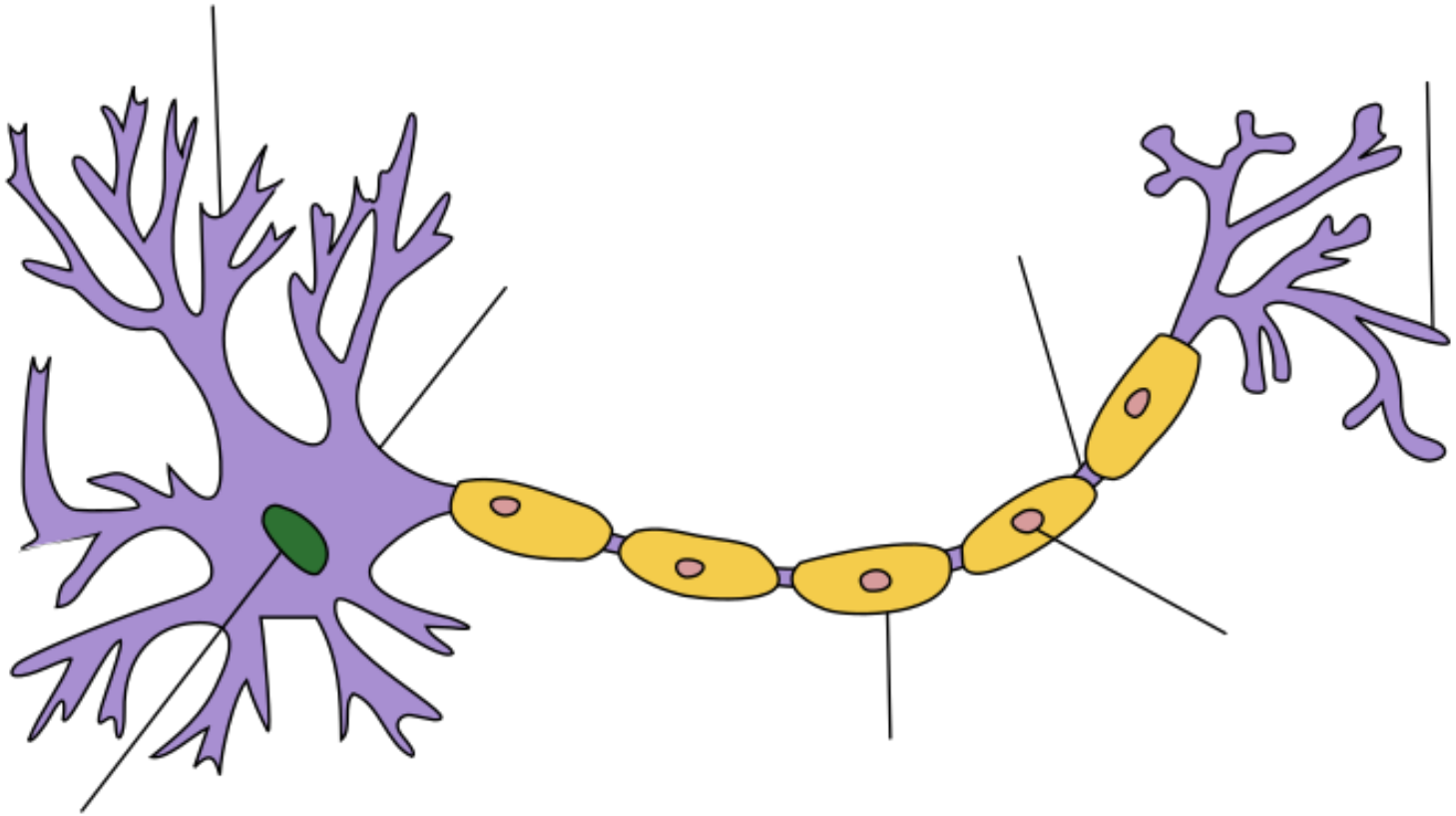
Let $Z_d(t; v) = v_d dt + \sigma dB_d(t)$ be a Gaussian process for alternative d accumulator, which are "uncoupled" and evolve independently as Brownian motions, where the continuation region $C(t)$ is the subset of values (v_1, v_2) such that $|v_1 - v_2| \leq \theta$. Then the choice probabilities at the random decision time t^* are logit

$$P(1|v_1, v_2) = \frac{1}{1 + \exp\{-2(v_1 - v_2)\theta/\sigma^2\}}$$

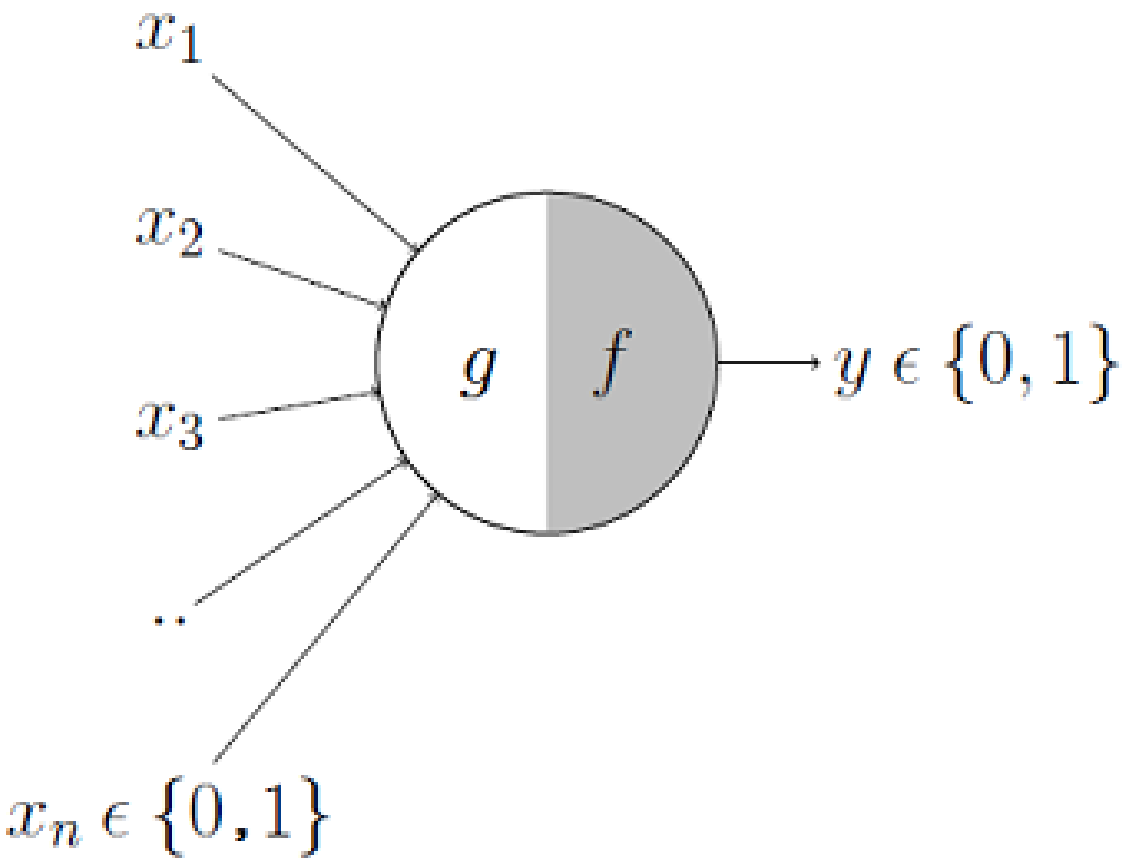
- Webb applied his framework to estimate a risk aversion parameter α in simple gambles from an experiment by Holt and Laury (2002).

Relationship to neural network literature

A schematic of a real neuron



An artificial neuron

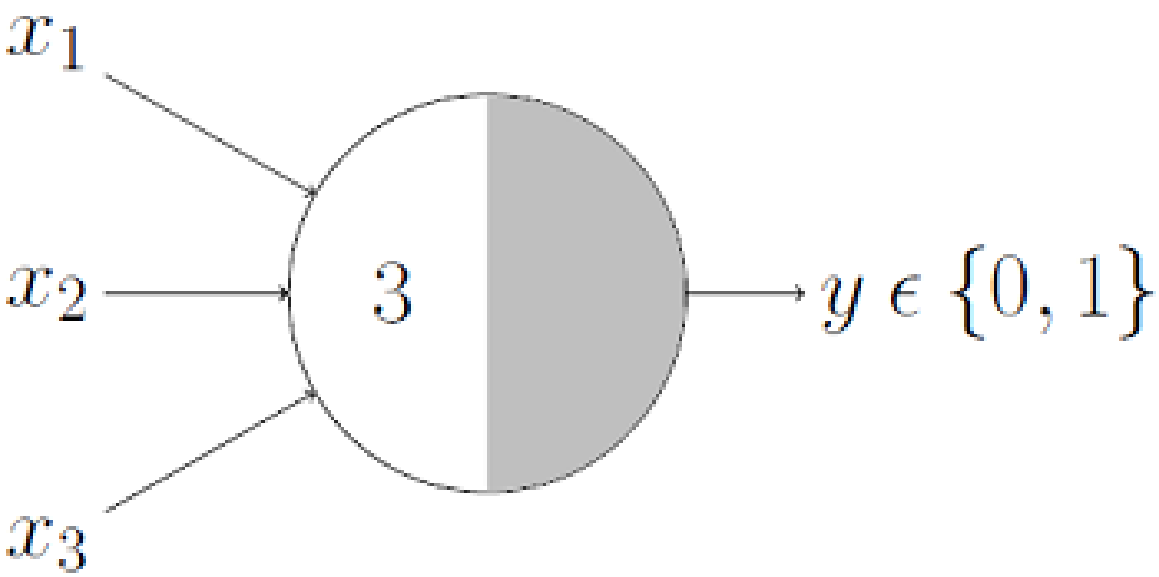


Example of g and activation function f

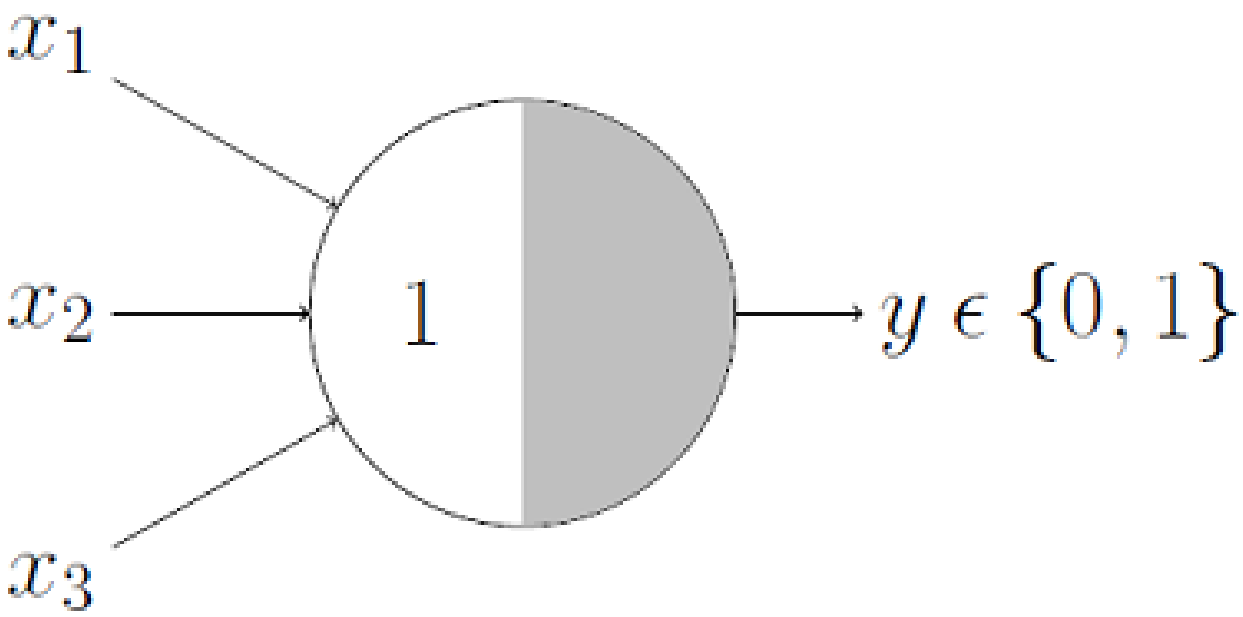
$$g(x_1, x_2, x_3, \ldots, x_n) = g(\mathbf{x}) = \sum_{i=1}^{n} x_i$$

$$\begin{aligned} y = f(g(\mathbf{x})) &= 1 \quad if \quad g(\mathbf{x}) \geq \theta \\ &= 0 \quad if \quad g(\mathbf{x}) < \theta \end{aligned}$$

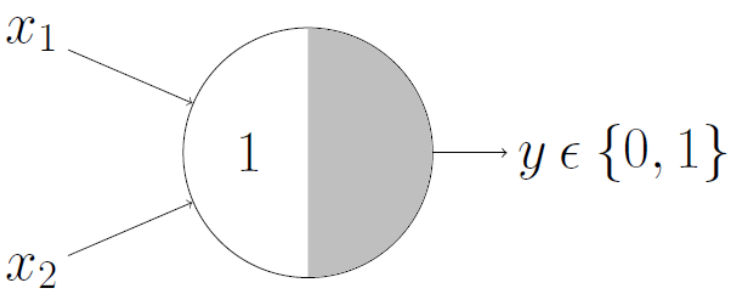
Implementing the logical “and” function



Implementing the logical “or” function

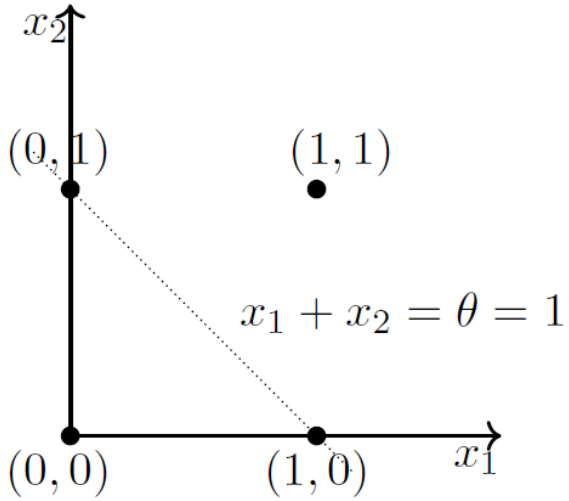


Logical “or” function as a classifier

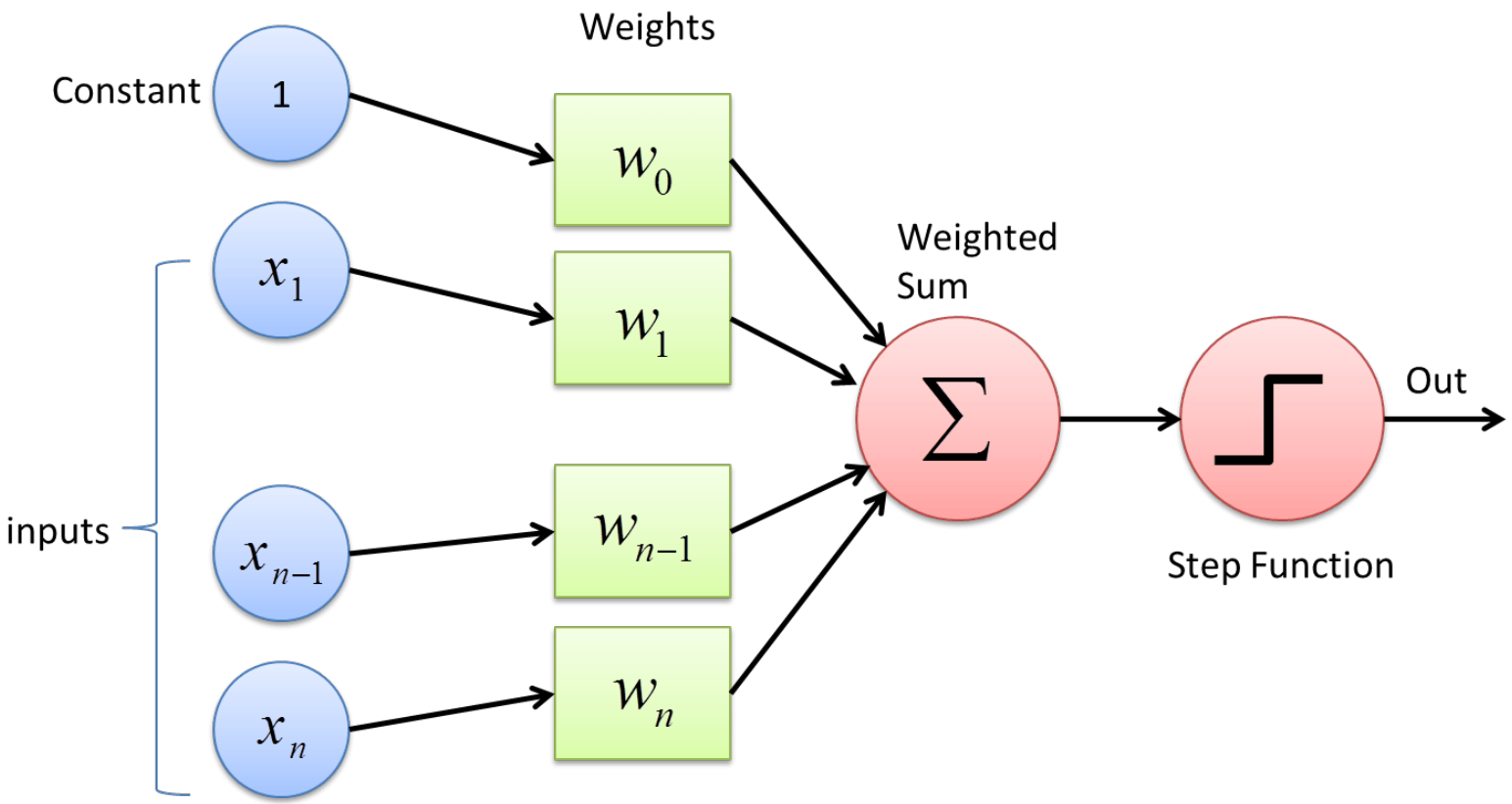


OR function

$$x_1 + x_2 = \sum_{i=1}^{2} x_i \geq 1$$



The perceptron, a single layer neural net



The perceptron, a single layer neural net

$$y = f\left(\sum_{i=0}^{n} w_i x_i\right)$$

where the activation function is

$$f(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Loistic activation (squashing) function

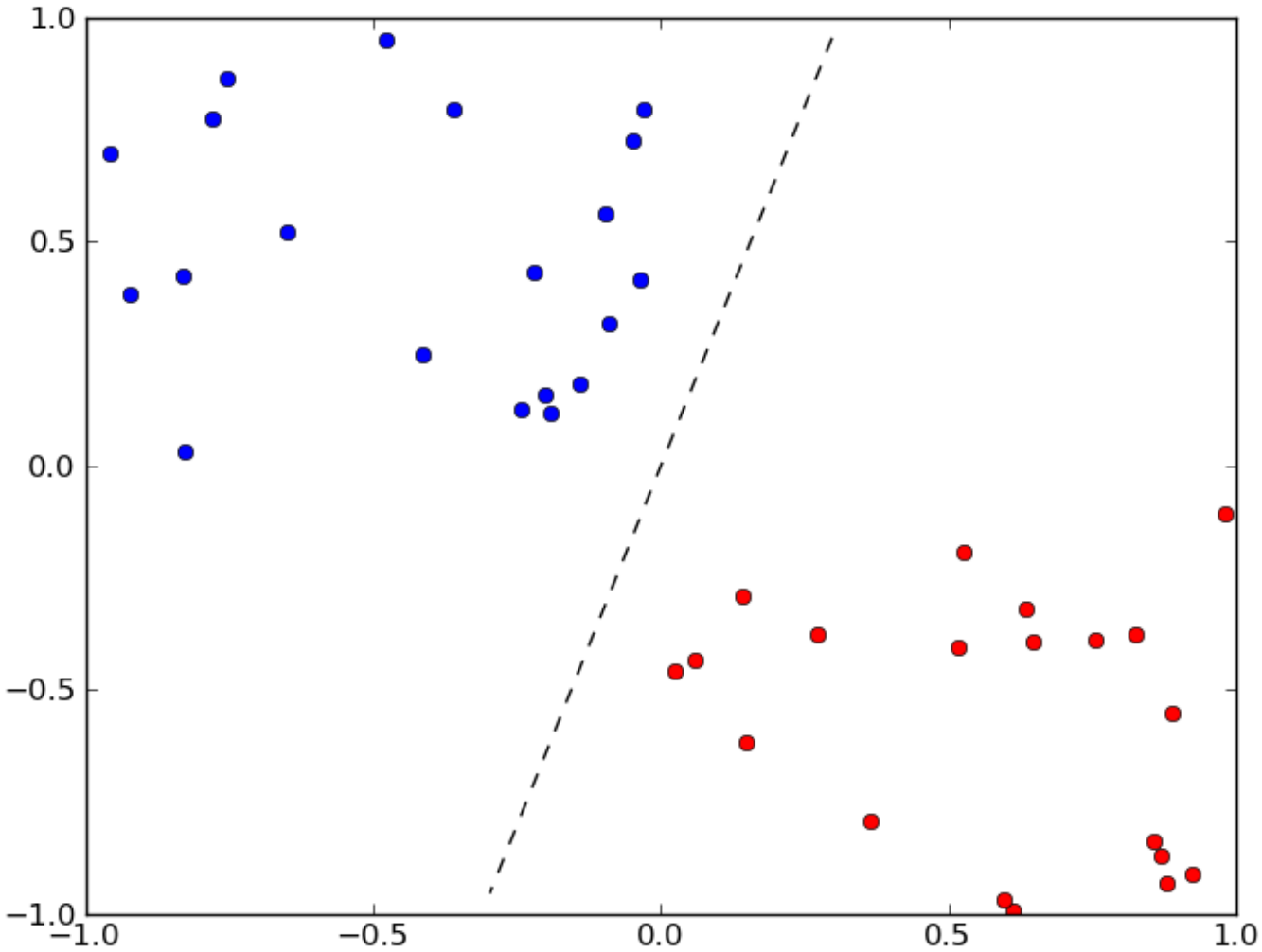
$$f_\sigma(x) = \frac{1}{1 + \exp\{x/\sigma\}}$$

Note that $\lim_{\sigma \to 0} f_\sigma(x) = f(x) = I\{x \geq 0\}$, so the logistic squashing function approximates the indicator activation function. In the neural net literature, the MNL model is known as the *softmax function* given by

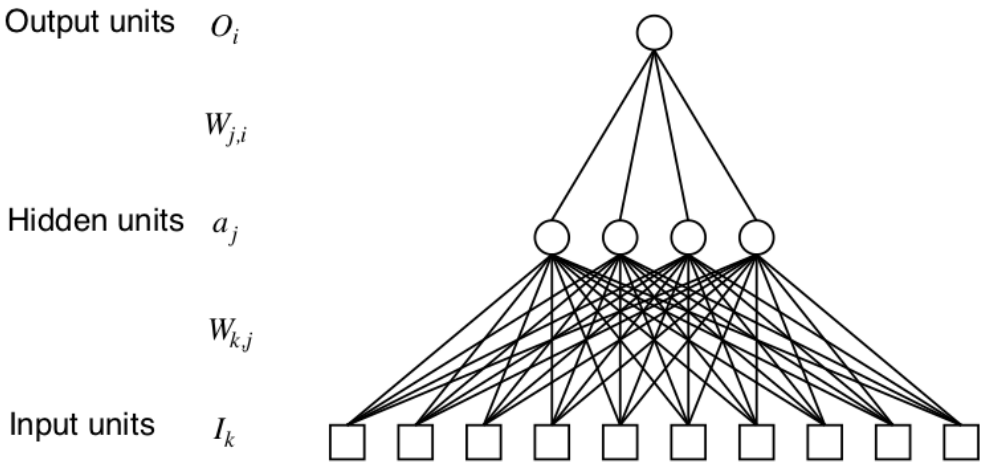
$$P_\sigma(x_i | x_1, \ldots, x_n) = \frac{\exp\{x_i/\sigma\}}{\sum_{j=1}^{n} \exp\{x_j/\sigma\}}$$

and $\lim_{\sigma \to 0} P_\sigma(x_1 | x_1, \ldots, x_n\} = I\{x_i = \max(x_1, \ldots, x_n)\}$.

Using a neural net for classificiation



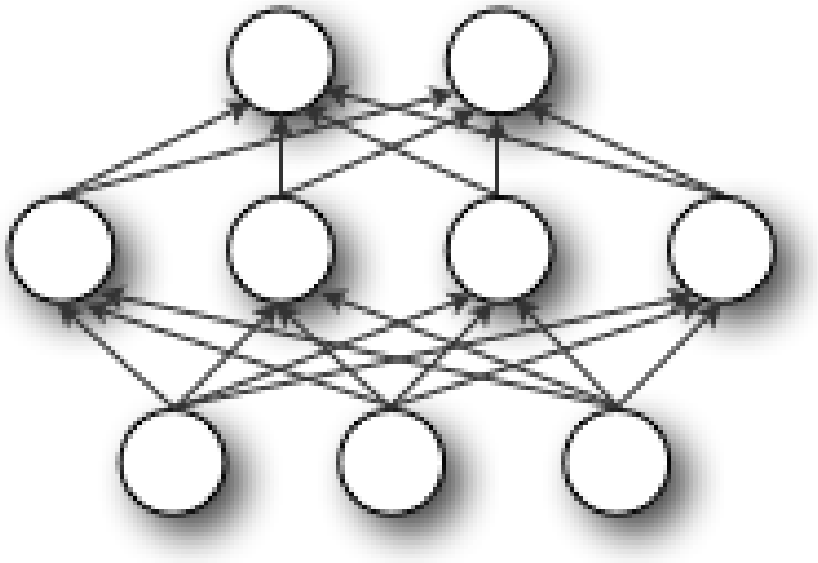
Two layer network with “hidden units”



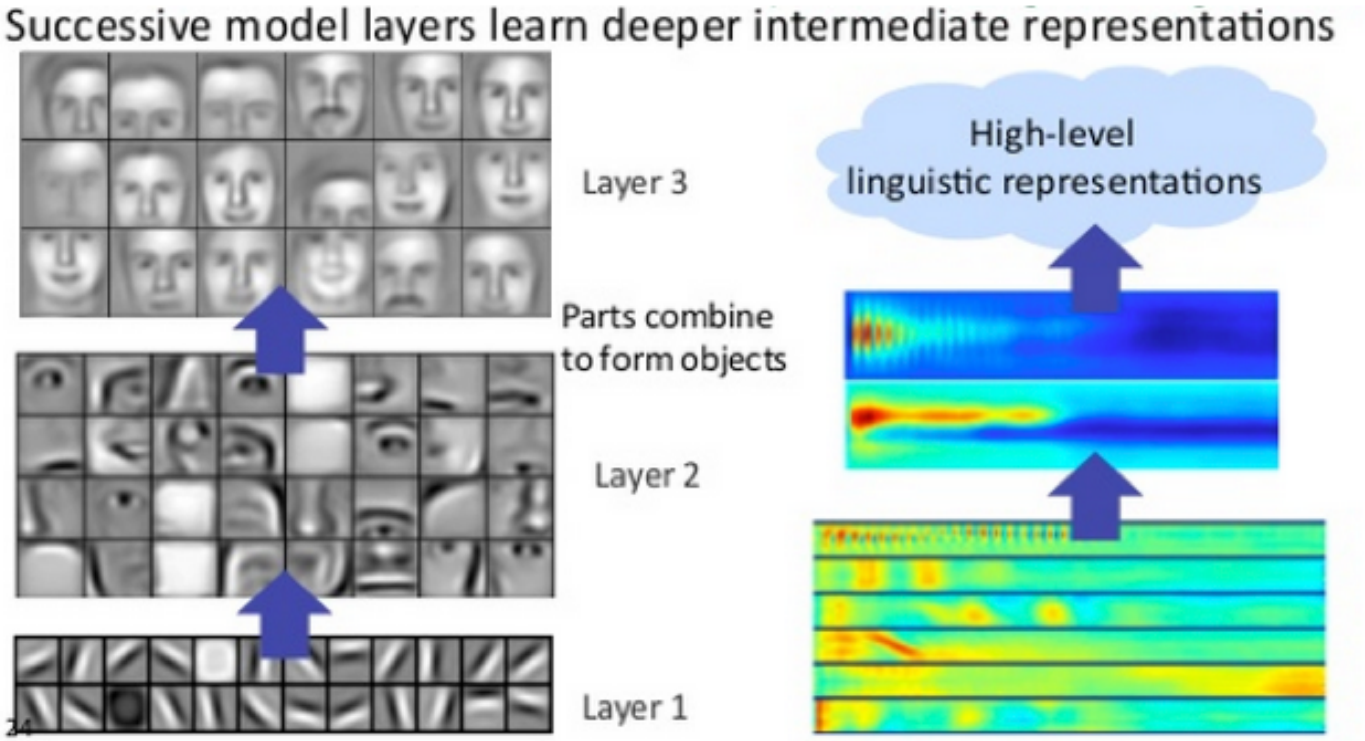
$$O_i = g(\sum_j W_{j,i} a_j)$$

$$a_j = g(\sum_k W_{k,j} I_k)$$

Multilayer nets with multiple outputs

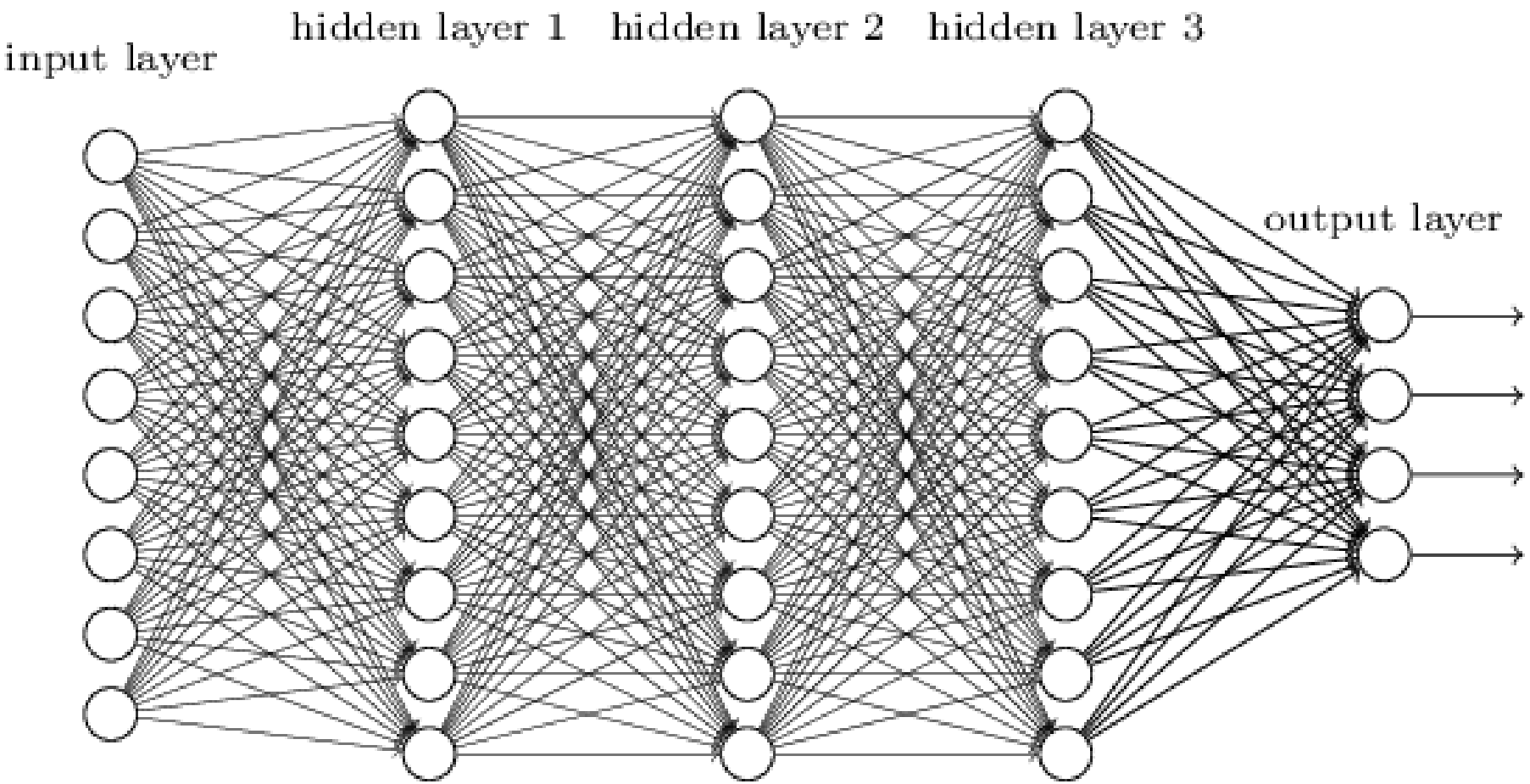


Using “deep nets” to capture more abstract concepts



Prior: underlying factors & concepts compactly expressed w/ multiple levels of abstraction

A fully connected "convolutional net"



Deep Q Networks

- A deep Q-network (DQN) is a type of deep learning model that combines a deep CNN with Q-learning, a form of reinforcement learning. Unlike earlier reinforcement learning agents, DQNs can learn directly from high-dimensional sensory inputs

From DeepMind

For artificial agents to be considered truly intelligent they should excel at a wide variety of tasks that are considered challenging for humans. Until this point, it had only been possible to create individual algorithms capable of mastering a single specific domain. With our algorithm, we leveraged recent breakthroughs in training deep neural networks to show that a novel end-to-end reinforcement learning agent, termed a deep Q-network (DQN), was able to surpass the overall performance of a professional human reference player and all previous agents across a diverse range of 49 game scenarios. This work represents the first demonstration of a general-purpose agent that is able to continually adapt its behavior without any human intervention, a major technical step forward in the quest for general AI.

Road Map: Dynamic structural models

1. Review of dynamic programming and some example applications

2. Review of key solution methods for solving dynamic programs
3. Overview of alternative methods for structural estimation of DPs
4. Maximum likelihood of dynamic discrete choice models

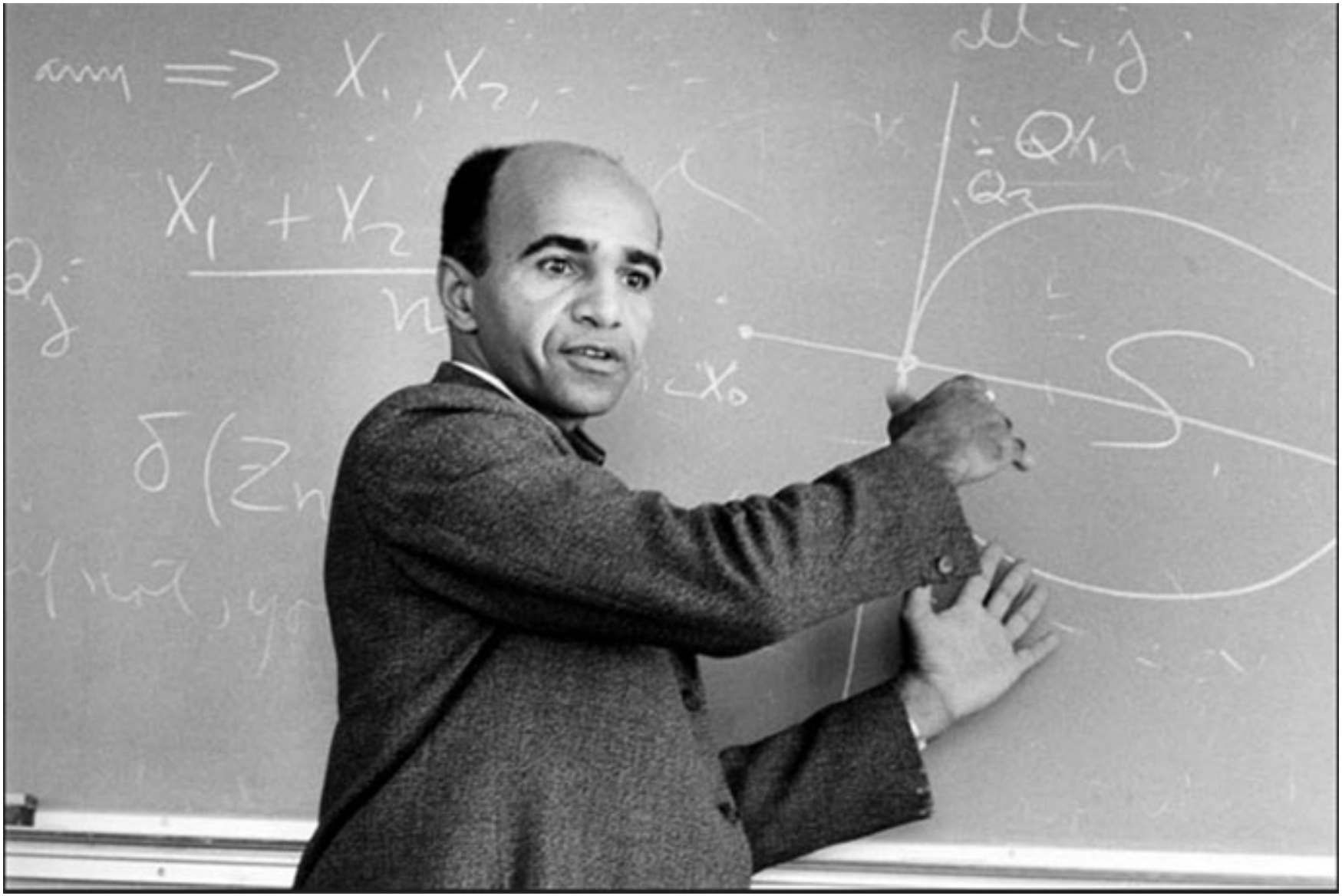
Origins of Dynamic Programming

- We assume you all know what it is: a recursive way of solving dynamic optimization problems
- If not, see Rust (2008) "dynamic programming" *New Palgrave Dictionary of Economics* 2nd edition
- Origins: Massé 1946 *Les réserves et la régulation de l'avenir* (optimal regulation of hydro reservoirs)
- Sequential analysis and statistical decision theory, Arrow, Blackwell and Girshick 1949 *Econometrica* "Bayes and Minimax Solution of Sequential Decision Problems"
- they were "trying to understand systematically sequential analysis from a decision-theoretic viewpoint. We wrote a paper in which was clearly displayed the recursive nature of the optimization problem. That is, each stage involved a choice of actions and a random event, which together changed the prospects for the future." (Arrow, 2002)

Pierre Massé 1898–1927

Kenneth A. Arrow b. 1921

David Blackwell 1919–2010



Richard Ernest Bellman, 1920–1984

Origin of the term "Dynamic Programming"

- from his 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.”

Bellman's "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."*
- These solutions can be computed via *backward induction on the game tree* where the "game tree" is an extensive-form representation of the game against nature.
- Bellman equation: the embodiment of the recursive way of formulating and solving these problems

Phelp's problem: a canonical problem

- Phelps' problem: Suppose an individual is retired and has no pension and earns no further labor income but can invest their wealth either in a risk-free bond with rate of return $R = (1 + r) \geq 1$.
- The individual has utility of consumption $u(c) = \log(c)$ and will live T more years, and has initial wealth W.
- How should the retiree save, consume and invest over his/her remaining lifetime to maximize their expected discounted utility? By backward induction, starting with $V_0(w) = \log(w)$, using Bellman equation

$$V_t(w) = \max_{0 \leq c \leq w} \left[\log(c) + \beta V_{t-1}\left(R(w - c)\right) \right].$$

for $t = 1, \ldots, T$.

The retirement problem

- Now, consider a *worker* who earns a non-stochastic annual salary y while working, and experiences an additive disutility of effort so the worker's utility of consuming c is $\log(c) - \alpha$ where α is the *disutility of work.*
- Let $V_t(w, 1)$ be the value function of a worker and $V_t(w, 0)$ be the value function of a retiree with wealth w. Assuming *retirement is an absorbing state* the Bellman equation for a worker is

$$V_t(w,1) = \max \left[V_t(w,0), \max_{0 \le c \le w} \left[\log(c) - \alpha + \beta V_{t-1}\left(R(w-c)+y, 1\right)\right] \right].$$

The retirement problem

- **Homework exercise:** solve Phelp's Problem and the Retirement Problem by Dynamic Programming, and derive the optimal retirement, consumption, and investment decision rules.
- The retirement problem is an example of a *discrete/continuous choice problem* where investment and consumption are the continuous choices and retirement is the discrete choice.
- Generally continuous choice problems are harder than discrete choice problems, and discrete-continuous choice problems are harder to solve than either purely continuous or purely discrete problems. The discrete max operator can generate *kinks* in the value function that can result in *discontinuities* in the continuous decision rules.

Dynamic discrete choice: bus engine replacement

- Let *Harold Zurcher* be the head of maintenance of the Madison (WI) bus company. Harold needs to decide when to replace a bus engine with a new or rebuilt bus engine.
- Assume that the number of miles a bus drives each much is an exponential distribution with parameter $\lambda \geq 0$.
- Assume the expected monthly maintenance cost is $c(x)$ and K is the cost of replacing the bus engine.
- Let x be the miles the bus has run *since last engine replacement* and let $V(x)$ be the mimimized expected discounted cost of following an optimal bus engine replacement strategy. The Bellman equation is

$$V(x) = \min \left[c(0) + K + \beta \int_0^\infty V(y)\lambda e^{-\lambda y} dy, \right.$$
$$\left. c(x) + \beta \int_0^\infty V(x+y)\lambda e^{-\lambda y} dy. \right]$$

Optimal investment and dividend policy, public firm

- Consider a *publicly traded firm.* It has capital stock k and a *cash flow production function* $f(k)$.
- Capital is *putty-clay:* that is, new capital can be bought using cash *investments* but once installed capital cannot be sold. It only *depreciates* at a constant rate $\delta \in (0, 1)$.
- Suppose that the firm has no debt and cannot borrow. The firm can either use its case flow to *invest* (i) or *pay dividends* (d) subject to the *budget constraint* $i + d \leq f(k)$.
- The market discounts the dividends of the firm at rate $\beta = 1/(1 + r)$ where r is the risk-free market interest rate. What investment and dividend policy maximizes the value of the firm?

$$V(k) = \max_{0 \leq i \leq f(k)} [f(k) - i + \beta V(k(1 - \delta) + i)]$$

Optimal investment and dividend policy, private firm

- Now consider a *private firm* owned by a single individual who also manages the firm. The owner has utility function $u(c)$ and the only source of income is from the dividends the owner pays to him/herself.
- Suppose the owner is infinitely lived and discounts future utilities at rate $\beta_p \in (0,1)$.
- What investment and dividend policy maximizes the owner's discounted utility?

$$U(k) = \max_{0 \leq i \leq f(k)} [u(f(k) - i) + \beta_p U(k(1-\delta) + i)]$$

- How does the dividend and investment policy of a private firm differ from that of a public firm?

How and when to go public

- Now consider a firm that is initially a privately held firm. The owner can choose to do an *IPO* and sell of some or all of his/her firm converting a privately held firm into a publicly held one.
- "A Simple Theory of Why and When Firms Go Public" (with Sudip Gupta, Indiana Kelley Business School)
- Let $P(k, \alpha, \omega)$ be the *gross IPO proceeds* when an owner chooses to sell a fraction α of his/her ownership stake in the firm and "cash out" a fraction ω of the total IPO proceeds.

$$P(k, \alpha, \omega) = (1 - \alpha)V\left(k + (1 - \omega)[P(k, \alpha, \omega)(1 - \rho) - F]\right)$$

where ρ is the proportional *underwriting fee* and F is the *fixed cost* of doing an IPO.

Optimal timing of an IPO

- If the owner of private firm with capital k did an IPO and converted his/her cash out and ownership into an annuity, the annual payment from this annuity would be

$$a(k, \alpha, \omega) = (1 - \beta)\left[\omega + \alpha/(1 - \alpha)\right] P(k, \alpha, \omega)$$

- The optimal (α, ω) is given by

$$(\alpha^*(k), \omega^*(k)) = \underset{\alpha \in [0,1]}{\textit{argmax}} \; \underset{\omega \in [0,1]}{\textit{argmax}} \; a(k, \alpha, \omega)$$

- $U(k, 0)$ be the value of keeping the firm private. Let $U(k, 1)$ be the value of going public with the optimal cash out and reinvestment

$$U(k, 1) = u\left(a(k, \alpha^*(k), \omega^*(k))\right)/(1 - \beta_p)$$

so the discrete IPO decision is determined from the solution to

$$U(k) = \max\left[U(k, 0), U(k, 1)\right]$$

Solving Dynamic Programs

- DPs sometimes admit *closed form, analytical solutions*
- But it only takes slight twiddles to a problem formulation to destroy them.
- Fortunately, there is a large body of work on *numerical dynamic programming* that shows how to numerically approximate the solutions to dynamic programs. See Rust "Numerical Dynamic Programming in Economics" (1994) *Handbook of Computational Economics*
- In finite horizon problems we do *backward induction* using *numerical quadrature* (to approximate expectation operators), *function approximation* (to approximate value functions), and *numerical optimization* (to approximate the optimal decision rule).

The Bellman Operator is a Contraction Mapping

- In *stationary infinite horizon problems* we need to solve the *Bellman equation* by solving for V as a *fixed point of the Bellman operator* $V = \Gamma(V)$ where

$$\Gamma(V)(s) = \max_{a \in A(s)} \left[u(s,a) + \beta \int_{s'} V(s') p(ds'|s,a) \right]$$

- In *bounded* stationary, infinite horizon problems, the Bellman operator can be shown to be a *contraction mapping* under fairly general conditions

$$\|\Gamma(V) - \Gamma(W)\| \leq \beta \|V - W\|$$

where $\|V\| = \sup_{s \in S} |V(s)|$.

Successive approximations = value function iteration

- When Γ is a contraction mapping the standard method of *successive approximations* can be used to find V, the fixed point of Γ starting from any initial guess V_0

$$V_{t+1} = \Gamma(V_t) \quad t = 0, 1, 2 \ldots$$

- Thus, the method of successive approximations is *globally convergent* from any initial guess V_0.
- Often $V_0 = 0$ and successive approximations is equivalent to *approximating the solution to an infinite horizon problem by solving a finite horizon problem with a sufficiently large horizon T*

Policy Iteration = Newton's Method

- Successive approximations can be slow when β is close to 1, and in problems with short time intervals ∇t we have $\beta = \exp\{-r\nabla t\}$ where r is the discount rate expressed as an annual interest rate. So $\beta \to 1$ as $\nabla t \to 0$.
- A much more effective method, which usually converges in a *very small number of iterations,* is the method of *policy iteration* introduced by Howard (1960).
- Policy iteration can be shown to be equivalent to *Newton's method* and in finite state, finite action problems it is also globally convergent

$$V_{t+1} = V_t - [I - \Gamma'(V_t)]^{-1}[V_t - \Gamma(V_t)] \quad t = 0, 1, \ldots,$$

- Main cost of policy iteration is inverting the linear operator $[I - \Gamma'(V_t)]$ where $\Gamma'(V_t)$ is the *Fréchet* or *directional derivative* of the Bellman operator Γ

$$\Gamma'(V)(W) = \lim_{\delta \downarrow 0} \frac{\Gamma(V + \delta W) - \Gamma(V)}{\delta}$$

Complexity of DP

- Main approximation method is to solve a *discrete approximation* to continuous state DP problems over a *finite grid* of points in the state space, $\{s_1, \ldots, s_N\}$.
- When we do this, we obtain an *approximate Bellman operator* $\Gamma_N : R^N \to R^N$, and $\Gamma'(V)$ is a *linear operator* on R^N.
- It follows that each Policy iteration step (Newton step) requires $O(N^3)$ computations to solve a linear system of equations.
- If there are d continuous state variables and we discretize each variable into N grid points, we have multidimensional grid with a total of N^d grid points.
- To find a uniform ϵ-approximation to the true value function we typically need $O(1/\epsilon)$ grid points in each dimension which implies that the number of grid points is $N = O(1/\epsilon^d)$ and total effort is $O(1/\epsilon^{3d})$.
- This is what Bellman referred to as *The Curse of Dimensionality.*

Breaking the Curse of Dimensionality

- Generally we can't: the complexity lower bound for both *deterministic and random algorithms* of continuous state, multi-dimensional DPs with continuous decisions increases is $\Omega(1/\epsilon^d)$.
- However Rust (1997) "Using Randomization to Break the Curse of Dimensionality" used *empirical process theory* to show that for *random grids* it is possible to construct a *random Bellman operator* $\tilde{\Gamma}_N$ which converges uniformly at rate $\sqrt{N}$ to the true Bellman operator Γ.
- This implies that the fixed point of the random Bellman operator, $\tilde{V}_N = \tilde{\Gamma}_N(\tilde{V}_N)$ converges at rate $\sqrt{N}$ to $V = \Gamma(V)$, the true solution.
- This implies that Rust's *random multigrid algorithm* breaks the curse of dimensionality, i.e. it bounds the (randomized) complexity of the problem from above by $O(1/\epsilon^4)$ independent of the dimension d.

Numerical Issues

- In many economic problems, value function is *unbounded* so there are issues of how to approximate an unbounded function over a finite grid (or set of basis functions)
- If the value function is unbounded, the Bellman operator may not be a *contraction mapping*
- Although numerical solution of DPs has extended the range of problems we can solve, and we can break the curse of dimensionality in some cases, still numerical methods are not a *panacea*
- There are many large scale DPs we would like to solve, but still do not have the computational power to solve them to a sufficient degree of reliability/accuracy: problems with 10 or more continuous state variables and multidemsional continuous choices are generally outside the frontier.
- A big problem: $d = 24$, see Brumm and Scheidegger (2014) "Using Adaptive Sparse Grids to Solve High-Dimensional Dynamic Models"

From solution to inference

- Once we have solved DP models, it is relatively easy to *simulate* them
- The limitless variety of DP models combined with the rich, detailed simulated behavior they can generate, made DP attractive an attractive basis for *empirical work* in economics
- *dynamic structural models* refer to both single agent DP and multi-agent game-theoretic and equilibrium models that have arisen in economics since the late 1970s.
- Early pioneers in macroeconometrics: linear quadratic models, Hansen and Sargent, 1978
- Early pioneers in microeconometrics: Gotz and McCall (retirement from the airforce, 1979), Wolpin (1984), Pakes (1986)
- The early micro work focused on *discrete choice* and used *unobserved state variables* to derive *conditional choice probabilities* and thus a *likelihood function* implied by the DP model.

Other estimators/methods of inference

- *Method of Simulated Moments* and *Indirect inference* — "if you can simulate it, you can estimate it"
- *Moment Inequality Estimation* uses a non-parametric first stage and "forward simulation" approach
- These methods extended the range of dynamic structural models substantially — to problems where it is difficult or impossible to derive a likelihood function for the data.
- They skirt problems of *statistical degeneracy.* That is, without a sufficiently rich specification of *unobserved state variables* a dynamic structural model can result in *zero likelihood problem* — i.e. there can be observations in the data that have a zero probability of occurring for *any* values of the structural parameters of the model.
- In macro, there is a much less formal method of *calibration* advocated by Edward Prescott and his disciples, but these models can often be reformulated and estimated using GMM or MSM.

The dynamic discrete choice model

- Now, let's take the static discrete choice model, the static RUM, and extend it to a dynamic discrete choices.
- Simplest case, a two period model. The agent has discount factor β and makes discrete choices $d_1 \in D_1(x_1)$ in period 1 and $d_2 \in D_2(x_2)$ to maximize expected discounted utility, where there is a transition probability $p_2(x_2, \epsilon_2 | x_1, \epsilon_1, d_1)$ for the observed and unobserved state variables $s_1 = (x_1, \epsilon_1)$ and $s_2 = (x_2, \epsilon_2)$.

$$\max_{d_1, d_2} E\left\{u_1(x_1, \epsilon_1, d_1) + \beta u_2(x_2, \epsilon_2, d_2)\right\}$$

- We make 3 key assumptions to make the solution tractable

Assumption EV

ϵ_1 and ϵ_2 are multivariate Type 1 extreme value distributions with components equal to the number of elements in the choice sets $D_1(x_1)$ and $D_2(x_2)$

The dynamic discrete choice model

Assumption AS

The utility functions have the additively separable representations

$$u_i(x_j, \epsilon_j, d) = u_j(x_j, d) + \epsilon_j(d) \quad d \in D_j(x_j)\ j \in \{1, 2\}$$

Assumption CI

The transition probability $p_2(x_2, \epsilon_2 | x_1, \epsilon_1, d_1)$ factors as

$$p_2(x_2, \epsilon_2 | x_1, \epsilon_1, d_1) = \pi_2(x_2 | x_1, d_1) q(\epsilon_2 | x_2)$$

- Let's do dynamic programming, starting from period 2. We have

$$V_2(x_2, \epsilon_2) = \max_{d_2 \in D_2(x_2)} u_2(x_2, d_2) + \epsilon_2(d_2)$$

- Now, we solve the problem at period 1 using Bellman's equation

$$V_1(x_1, \epsilon_1) = \max_{d_1 \in D_1(x_1)} [u_1(x_1, d_1) + \epsilon_1(d_1) + \beta E \{ V_2(x_2, \epsilon_2) | x_1, \epsilon_1, d_1 \}]$$

The dynamic discrete choice model

- Notice that we have used the AS assumption. Now we use the EV and CI assumptions to write

$$
\begin{aligned}
& E\{V_2(x_2,\epsilon_2)|x_1,\epsilon_1,d_1\} \\
= & \int_{\epsilon_2}\int_{x_2} V_2(x_2,\epsilon_2)p_2(x_2,\epsilon_2|x_1,\epsilon_1,d_1)dx_2 d\epsilon_2 \\
= & \int_{x_2}\int_{\epsilon_2} V_2(x_2,\epsilon_2)q_2(\epsilon_2|x_2)p_2(x_2|x_1,d_1)d\epsilon_2 dx_2 \\
= & \int_{x_2}\left[\int_{\epsilon_2}\max_{d_2}[u_2(x_2,d_2)+\epsilon_2(d_2)]q_2(\epsilon_2|x_2)d\epsilon_2\right]p_2(x_2|x_1,d_1)dx_2 \\
= & \int_{x_2}\sigma\log\left(\sum_{d_2}\exp\{u_2(x_2,d_2)/\sigma\}\right)p_2(x_2|x_1,d_1)dx_2
\end{aligned}
$$

The dynamic discrete choice model

- Notice that we also have these representations of the value functions

$$\begin{aligned} V_2(x_2, \epsilon_2) &= \max_{d_2 \in D_2(x_2)} [u_2(x_2, d_2) + \epsilon_2(d_2)] \\ V_1(x_1, \epsilon_1) &= \max_{d_1 \in D_1(x_1)} [v_1(x_1, d_1) + \epsilon_1(d_1)] \end{aligned}$$

  where

$$v_1(x_1, d_1) = u_1(x_1, d_1) + \beta \int_{x_2} \sigma \log \left(\sum_{d_2} \exp\{u_2(x_2, d_2)/\sigma\} \right) p_2(x_2|x_1, d_1)$$

- This implies that in each period, choices are defined by a standard RUM model, but with $v_j(x_j, d_j)$ instead of $u_j(x_j, d_j)$

The dynamic discrete choice model

- This immediately implies that the choice probabilities in each period have the MNL form

$$
\begin{aligned}
P_2(d_2|x_2) &= \int_{\epsilon_2} I\left\{ d_2 = \underset{d \in D_2(x_2)}{argmax}[u_2(x_2,d) + \epsilon_2(d)] \right\} q_2(\epsilon_2|x_2) d\epsilon_2 \\
&= \frac{\exp\{u_2(x_2,d_2)/\sigma\}}{\sum_{d \in D_2(x_2)} \exp\{u_2(x_2,d)/\sigma\}} \\
P_1(d_1|x_1) &= \int_{\epsilon_1} I\left\{ d_1 = \underset{d \in D_1(x_1)}{argmax}[v_1(x_1,d) + \epsilon_1(d)] \right\} q_1(\epsilon_1|x_1) d\epsilon_1 \\
&= \frac{\exp\{v_1(x_1,d_1)/\sigma\}}{\sum_{d \in D_1(x_1)} \exp\{v_1(x_1,d)/\sigma\}}
\end{aligned}
$$

The likelihood for the dynamic discrete choice model

- Suppose the utility function depends on parameters θ_1 and the transition probability depends on parameters θ_2, so the full parameter to be estimaed is $\theta = (\beta, \theta_1, \theta_2)$
- Suppose we have panel data on choices of N individuals, $\{x_{ti}, d_{ti}\}$, then the full likelihood is

$$L_N(\theta) = \prod_{i=1}^{N} P_2(d_{2i}|x_{2i}, \theta) p_2(x_{2i}|x_{1i}, d_{1i}, \theta_2) P_1(d_{1i}|x_{1i}, \theta)$$

- We can also estimate in two stages, estimating θ_2 in the first stage using the following *partial likelihood*

$$L_N^p(\theta_2) = \prod_{i=1}^{N} p_2(x_{2i}|x_{1i}, d_{1i}, \theta_2)$$

- then estimate the remaining parameters (β, θ_1) using this partial likelihood

$$L_N^p(\beta, \theta_1, \hat{\theta}_2) = \prod_{i=1}^{N} P_2(d_{2i}|x_{2i}, \beta, \theta_1, \hat{\theta}_2) P_1(d_{1i}|x_{1i}, \beta, \theta_1, \hat{\theta}_2)$$

Extending to infinite horizon problems

- It is straightforward to see how to generalize the two period setup above to T period discete dynamic programming problems.
- What about stationary infinite horizon problems? The Bellman equation is

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

where we have used the AS and CI assumptions. Now define the function $v(x,d)$ by

$$v(x,d) = u(x,d)+\beta\int_{x'}\int_{\epsilon'}V(x',\epsilon')q(d\epsilon'|x')p(x'|x,d)d\epsilon' dx'$$

- Then it is clear that $V(x,\epsilon)$ has the representation

$$V(x,\epsilon) = \max_{d\in D(x)}[v(x,d)+\epsilon(d)]$$

Extending to infinite horizon problems

- Now substitute the latter representation for $V(x,\epsilon)$ into the formula for its conditional expectation, using the properties of the EV distribution to derive

$$\begin{aligned} EV(x,d) &= \int_{x'}\int_{\epsilon'} V(x',\epsilon')q(\epsilon'|x')d\epsilon' p(x'|x,d)dx' \\ &= \int_{x'}\left[\int_{\epsilon'} \max_{d'\in D(x')}[v(x',d')+\epsilon'(d')]q(\epsilon'|x')d\epsilon'\right] p(x'|x,d) \\ &= \int_{x'} \sigma \log\left(\sum_{d'\in D(x')} \exp\{v(x',d')/\sigma\}\right) p(x'|x,d) \end{aligned}$$

- Substituting this into the equation defining $v(x,d)$ we obtain the following functional equation

$$v(x,d) = u(x,d)+\beta \int_{x'} \sigma \log\left(\sum_{d'\in D(x')} \exp\{v(x',d')/\sigma\}\right) p(x'|x,d)$$

Nested fixed point algorithm

- We can write v as fixed point of the operator Ψ, $v = \Psi(v)$ and it is not hard to show Ψ is a contraction mapping so v is the unique solution.
- The choice probabilities are given by the MNL formula

$$P(d|x) = \frac{\exp\{v(x,d)/\sigma\}}{\sum_{d' \in D(x')} \exp\{v(x',d')/\sigma\}}$$

- If u depends on parameters θ_1 and p on θ_2, then $\theta = (\beta, \theta_1, \theta_2)$, and we can show (using the Implicit functio theorem) that v is a smooth function of θ if u is a smooth function of θ_1 and p is a smooth function of θ_2.
- Given panel data, (possibly unbalanced) $\{(x_{ti}, d_{ti}) | t \in \{\underline{T}_i, \ldots, \overline{T}\}, i \in \{1, \ldots, N\}\}$, the full likelihood function is

$$L_N(\theta) = \prod_{i=1}^{N} \prod_{t=\underline{T}_i}^{\overline{T}_i} P(d_{ti}|x_{ti}, \theta) p(x_{ti}|x_{t-1,i}, d_{t-1,i}, \theta_2) \tag{113}$$