Axioms for Neural Normalization

Adam Brandenburger
J.P. Valles Professor, NYU Stern School of Business
Distinguished Professor, NYU Tandon School of Engineering
Faculty Director, NYU Shanghai Program on Creativity + Innovation
Global Network Professor
New York University

“A basic presupposition of this book is that choice behavior is best described as a probabilistic, not an algebraic, phenomenon….

The presently unanswerable question is which approach will, in the long run, give a more parsimonious and complete explanation of the total range of phenomena.”

-- R. Duncan Luce (1959)

Luce, R.D., *Individual Choice Behavior: A Theoretical Analysis*, Wiley, 1959; http://img2.yardbarker.com/media/8/8/884c6ca6f1b4a25bc0597be7da8ea8d7c4c41d66/xl/Random-Numbers.jpg?stamp=1491141319

A Decision Model Based on Neural Normalization Findings

$$\rho(x, A) = \Pr\left(x = \underset{y \in A}{\operatorname{argmax}} \frac{v(y)}{F(A)} + \varepsilon_y\right)$$

Utility function

Probability of choosing alternative x from choice set A

Normalization factor

Noise

Levy, D., and P. Glimcher (2012), "The Root of All Value: A Neural Common Currency for Choice," *Current Opinion in Neurobiology*, 22 2012, 1027-1038; Heeger, D., "Normalization of Cell Responses in Cat Striate Cortex," *Visual Neuroscience*, 9, 1992, 181-197; Louie, K., L. Grattan, and P. Glimcher, "Reward Value-Based Gain Control: Divisive Normalization in Parietal Cortex," *Journal of Neuroscience*, 31, 2011, 10627-10639

What Kind of Optimization Process Is This?

Can we understand the normalization model in (some kind of) a **cost-benefit** framework?

Can we identify the **specific role played by the normalization factor** in such a framework?

We start with the (grounded) axiom that **behavior is stochastic**

Stochasticity can be reduced but we assume that reduction **involves a cost** (this can be justified on basic thermodynamic grounds)

Our decision maker therefore faces a **trade-off** between the (opportunity) cost of stochasticity and the (energetic) cost of determinism

We specify the cost of decreasing stochasticity to be proportional to the associated decrease in **Shannon entropy** (Shannon, 1948)

 Shannon, C., "A Mathematical Theory of Communication," *Bell System Technical Journal*, 27, 1948, 379-423

From Bounded Rationality to Grounded Rationality

A theory of decision making should be consistent with

"the access to information and the computational capacities that are actually possessed by the organism"
-- Herbert Simon (1955)

Simon, H., "A Behavioral Model of Rational Choice," *Quarterly Journal of Economics*, 69, 1955, 99-118; http://www.biografiasyvidas.com/biografia/s/fotos/simon_herbert_a.jpg

Decision-Making Framework

Let X be a finite set of alternatives

Let $\Im$ be the collection of all nonempty subsets ("choice sets") from X

A **random choice rule** is a function $\rho : X \times \Im \to [0,1]$ such that
$\rho(x, A) > 0$ if and only if $x \in A$ and

$$\sum_{x \in \mathrm{A}} \rho(x, A) = 1 \text{ for all } A \in \Im$$

The interpretation is that $\rho(x, A)$ is the probability that the DM chooses alternative x when faced with choice set A

Let $\wp$ denote the set of all random choice rules (for given X)

Information-Processing Model

Let $v(x)$ be the value of alternative x

The expected utility of choice rule ρ on choice set A is

$$\sum_{x \in A} \rho(x, A) v(x)$$

The associated Shannon entropy is

$$H(\rho, A) = -\sum_{x \in A} \rho(x, A) \ln \rho(x, A)$$

Let $H_{\max}(A)$ denote the maximum possible entropy (achieved when ρ is uniform)

We employ a proportionality factor $F(A)$ in our specification of the cost of choice rule ρ on choice set A, which is

$$F(A)(H_{\max}(A) - H(\rho, A))$$

Information-Processing Model contd.

A random choice rule ρ has an **information-processing formulation** if there exist functions $v: X \rightarrow (0,\infty)$ and $F: \Im \rightarrow (0,\infty)$ such that for all $A \in \Im$

$$\rho \in \underset{\hat{\rho} \in \wp}{\operatorname{argmax}} \left\{ \sum_{x \in \mathrm{A}} \hat{\rho}(x,A)v(x) - F(A)(H_{\max}(A) - H(\hat{\rho},A)) \right\}$$

Note: If we relax our definition of a random choice rule to allow for the possibility of assigning probability zero to an available alternative, this turns out never to be optimal. This is because the derivative of entropy become infinite as a probability approaches zero.

Information-Processing Foundation

We can prove that a random choice rule ρ has a neural-normalization form (when the noise is Gumbel-distributed) if and only if it has an information-processing formulation

The **normalization factor** in the first formulation becomes a part of the **cost function** in the second formulation, which gives a (partial?) normative foundation for this factor

At the technical level, the equivalence relies on:

1. a reinterpretation of standard math from thermodynamics (derivation of the Boltzmann distribution via free energy minimization)

2. an easy extension of standard math from estimation of choice models (McFadden, 1978)

Steverson, K., A. Brandenburger, and P. Glimcher, op.cit.; McFadden, D., "Modelling the Choice of Residential Location," in Karlqvist, S., L. Lundqvist, F. Snickars, and J. Weibull (eds.), *Spatial Interaction Theory and Planning Models*, Volume 673, North-Holland, 1978, 75-96

Comparison with Other Models

$$\rho(x,A) = \Pr\left(x = \operatorname*{argmax}_{y\in A} \frac{v(y)}{F(A)} + \varepsilon_y\right)$$

1. The factor $F(A)$ distinguishes this model from random utility

2. The Gumbel distribution arises as the asymptotic distribution of the maximum of a sequence of i.i.d. normal r.v.'s (can this fact be used to ground the model further?)

3. Neural studies have often taken $F(A)$ to be a sum (where σ is a constant):

$$F(A) = \sigma + \sum_{z\in A} v(z)$$

What Kind of Behavior Does Normalization Permit?

A random choice rule ρ obeys **regularity** if $\rho(x,B) \leq \rho(x,A)$ when $x \in A \subseteq B$, i.e., adding alternatives reduces existing probabilities

≈10% ≈90% ≈30% ≈30% ≈40%

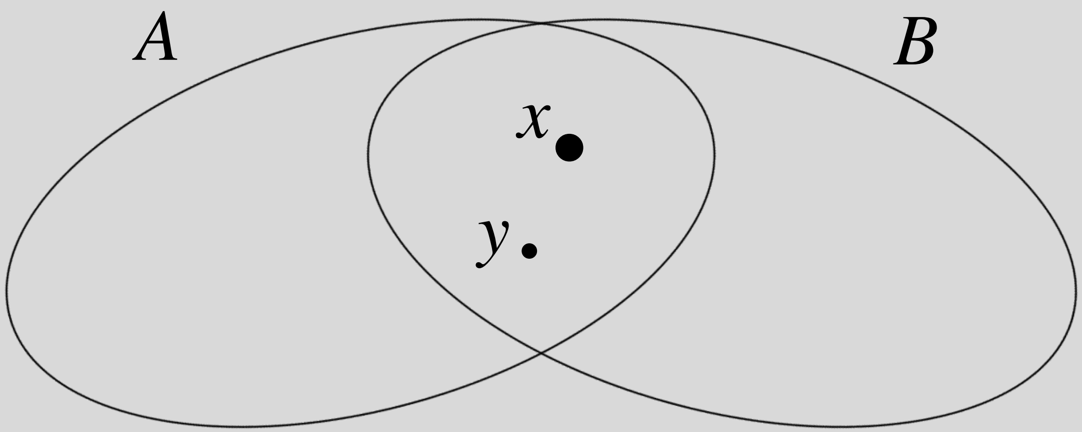
The classic **Luce rule** (axiomatized by independence from irrelevant alternatives) obeys regularity

Departures from regularity are routinely observed in experiments (as in the “attraction effect”)

Huber, J., J. Payne, and C. Puto, “Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis,” *Journal of Consumer Research*, 9, 1982, 90-98; Simonson, I., “Choice Based on Reasons: The Case of Attraction and Compromise Effects,” *Journal of Consumer Research*, 16, 1989, 158-174; images from *The Attraction Effect Explained*, MarketingExplained channel, https://www.youtube.com/watch?v=pybVPm5VU5A

Independence from Irrelevant Alternatives

The **IIA rule** characterizes the Luce model of stochastic choice (Luce, 1959)



$$\frac{\rho(x,A)}{\rho(y,A)} = \frac{\rho(x,B)}{\rho(y,B)}$$

Luce, R. D., *Individual Choice Behavior: A Theoretical Analysis*, Wiley, 1959

Relaxed IIA

Let's consider a relaxation of IIA that requires only

$$G_A\left(\frac{\rho(x,A)}{\rho(y,A)}\right) = G_B\left(\frac{\rho(x,B)}{\rho(y,B)}\right) \qquad (*)$$

where $G_A : (0,\infty) \to (0,\infty)$ is strictly increasing (and likewise for G_B)

Note: The "strictly increasing" condition implies that x is more likely to be chosen than y in A if and only if this is true in B

Say a collection of random choice rules $P \in \wp$ is **free** if for every $A \in \Im$ and full-support probability measure λ on A there is a $\rho \in P$ with $\rho(\cdot, A) = \lambda$

Note: The family of Luce rules is free

A family of functions $\{G_A\}_{A\in\Im}$ is **admissible** if the G_A are strictly increasing and the collection of random choice rules satisfying (*) for $\{G_A\}_{A\in\Im}$ is free

Behavioral Characterization

A random choice rule ρ is a **relaxed IIA rule** if there exists an admissible family of functions $\{G_A\}_{A\in\mathfrak{I}}$ with respect to which ρ obeys

$$G_A\left(\frac{\rho(x,A)}{\rho(y,A)}\right) = G_B\left(\frac{\rho(x,B)}{\rho(y,B)}\right) \qquad (*)$$

We can prove that a random choice rule ρ has a neural-normalization form (when the noise is Gumbel-distributed) if and only if it is a relaxed IIA rule

The equivalence is a new argument (using the Cauchy functional equation)

Relaxed IIA allows **departures from regularity** (via suitable choice of $F(\cdot)$)

There are closed-form axioms that characterize relaxed IIA, so that **behavioral testing** of the rule is possible

From Green Cheese to …

In computer science, it used to be said that the theory was independent of the physical substrate ("computers might as well be made of green cheese"*)

This view turned out to be wrong (thanks to the discovery of quantum speedup)

In decision theory, the traditional position appears to have been similar --- that the physical substrate does not matter

This view never made good sense, and now we can use inputs from the cognitive sciences to re-build decision theory … and game theory … and …

* I owe this nice phrasing to my colleague Samson Abramsky
http://www.istockphoto.com/photo/crowd-of-people-above-gm171159648-19164839