Choice-Theoretic Foundations of the Divisive Normalization Model*

Kai Steverson[†]
Adam Brandenburger[‡]
Paul Glimcher[§]

First version: September 10, 2016
This version: November 9, 2018

Abstract

Recent advances in neuroscience suggest a utility-like calculation is involved in how the brain makes choices, and that this calculation may use a computation known as divisive normalization. While this tells us *how* the brain makes choices, it is not immediately evident *why* the brain uses this computation or exactly *what* behavior is consistent with it. In this paper, we address both of these questions by proving a three-way equivalence theorem between the normalization model, an information-processing model, and an axiomatic characterization. The information-processing model views behavior as optimally balancing the expected value of the chosen object against the entropic cost of reducing stochasticity in choice. This provides an optimality rationale for *why* the brain may have evolved to use normalization. The axiomatic characterization gives a set of testable behavioral statements equivalent to the normalization model. This answers *what* behavior arises from normalization. Our equivalence result unifies these three models into a single theory that answers the "how", "why", and "what" of choice behavior. JEL Codes: D87, D81.

*Financial support from NIH grant R01DA038063, NYU Stern School of Business, NYU Shanghai, and J.P. Valles is gratefully acknowledged. We have greatly benefited from discussions with Andrew Caplin, Faruk Gul, Kenway Louie, Anthony Marley, Paula Miret, Wolfgang Pesendorfer, Doron Ravid, Shellwyn Weston, Jan Zimmerman, members of the Glimcher Lab at NYU, and seminar audiences at Columbia, HKUST, NYU Shanghai, PHBS Shenzhen, Princeton, Yale, and Zurich University.

[†] Department of Neuroscience, New York University, New York, 10003, U.S.A. ksteverson@nyu.edu, https://sites.google.com/site/ksteverson/

[‡] Stern School of Business, Tandon School of Engineering, NYU Shanghai, New York University, New York, NY 10012, U.S.A., adam.brandenburger@stern.nyu.edu, http://www.adambrandenburger.com

[§] Department of Neuroscience, New York University, NYU School of Medicine, New York, 10016, U.S.A. paul.glimcher@nyu.edu, http://www.neuroeconomics.nyu.edu/people/paul-glimcher/

1 Introduction

Choice is often modeled as behavior that seeks to maximize a utility function. Advances in neuroscience over the past few decades have pointed to a discrete set of brain areas apparently dedicated to representing a quantity that functions much like a utility representation (Platt and Glimcher, 1999; Knutson et al., 2001; Fehr and Rangel, 2011; Glimcher, 2011). These brain areas produce different levels of neural activity for different choice alternatives, where higher associated activity indicates a higher probability that the alternative in question is chosen. This "utility-like" process of the brain is referred to by neuroscientists as the subjective value function. Its stochastic relationship with choice can be modeled by a utility function with an additive noise term (Webb et al., 2013), which adapts some of the building blocks of classic stochastic choice theory (Luce, 1959; McFadden, 1973) to a modern neuroscientifically-motivated setting.

Another important finding from neuroscience about the subjective value function is that it appears to employ a computation known as *divisive normalization* (Louie, Grattan, and Glimcher, 2011). Originally identified in the visual domain, divisive normalization has been argued to be a canonical neural computation (Carandini and Heeger, 2012) in which the neural activity level generated by a particular stimulus (whether visual or other) is re-scaled in a way that depends on neighboring stimuli, that is, on the context. In choice behavior, divisive normalization works by re-scaling the value of each option by a function that depends on the values of all available alternatives (Louie et al., 2015). While this re-scaling function can take different forms, it is typically assumed to be proportional to the sum of the values of the items currently in the choice set plus a constant, which is the form we adopt throughout this paper.

More recently, divisive normalization has been successfully used to explain human choice behavior. Louie, Khaw, and Glimcher (2013), Itthipuripat et al. (2015), and Khaw, Glimcher, and Louie (2017) conduct choice experiments that confirm different predictions of the divisive normalization model regarding context effects in choice. Webb, Glimcher, and Louie (2014) show that divisive normalization explains choice involving departures from the classic Luce rule —which is equivalent to the Independence of Irrelevant Alternatives — better than other proposals such as Multinomial Probit. Glimcher and Tymula (2018) document how divisive normalization can re-produce many of the behaviors typically associated with prospect theory. Landry and Webb (2017) show how a variant of the divisive normalization model can accommodate a range of context effects. Divisive normalization can also accommodate the violation of the Regularity Property often seen in the well-known Attraction Effect (Huber, Payne, and Puto, 1982; Simonson, 1989). Its ability to violate Regularity demonstrates that divisive normalization is not a member of the "random utility" class of models (Thurstone, 1927; Block and Marschak, 1960), since these latter obey this condition.

While stochastic choice with divisive normalization tells us *how* the brain makes choices, it is not immediately evident *why* the brain uses this particular computation or exactly *what* behavior is consistent with it. In this paper, we address both of these questions by proving a three-way equivalence theorem between the divisive normalization model, an information-processing model, and an axiomatic characterization. The information-processing model views behavior as optimally balancing the expected value of the chosen object against the entropic cost of reducing stochasticity in choice. This provides an optimality rationale for why the brain may have evolved to use divisive normalization. The axiomatic characterization gives a set of testable behavioral statements equivalent to the divisive normalization model. This gives a precise answer to the question of what behavior arises from divisive normalization. Our equivalence result unifies the three models into a single theory that can simultaneously address the how, why, and what of choice behavior.

Divisive normalization, as a functional form, has been used by neuroscientists for over a decade to model how neural data, and, in particular, the neurally measurable subjective value function, determines choices. However, whether neural data can be precisely predicted (identified cardinally) from choices has not yet been addressed. For neuroscientists, this has proven to be a stark limitation, since it has meant neural observables have only been inferred from choice via the fitting of arbitrary functional forms to choice data. In this paper, we prove a result showing that choice data alone can be used to uniquely specify neural observables within the framework of the normalization model. In other words, we show there is a one-to-one relationship between neural and choice observables.

This result is important to empirical neuroscientists, since it allows for novel and precise predictions which link neural and choice data. This result is also important for economists working only with choice data since it allows the application of insights from neural analysis without neural data. For example, it has been suggested that the subjective value function can be used as a direct measure of welfare and happiness (Loewenstein, Rick, and Cohen, 2008; Glimcher, 2011). Without taking any stance on this controversial question, we merely point out that the one-to-one relationship we establish means that this measurement can be made from choice data alone – and it suggests that inter-individual comparisons may also be possible. We prove the one-to-one relationship as a corollary to our more general uniqueness result that shows that the parameters of the divisive normalization model are behaviorally identified up to a multiplicative scaling. This level of identification is similar to other theories of stochastic choice, such as the Luce rule, and is enough to rank the alternatives by their values and to rank the choice sets by the expected value the agent receives when facing that set.

We next use our equivalence result to study how the divisive normalization model handles context effects. We use a standard notion of context effects as departures from the Independence of Irrelevant Alternatives (IIA) property developed by Luce (1959). Violations of IIA have been widely documented empirically, including in the well-known Compromise and Attraction Effects (Huber, Payne, and Puto, 1982; Simonson, 1989) as well as in a wide variety of other settings. (For a survey, see Rieskamp, Busemeyer, and Mellers 2006.) We apply our equivalence result to find analogs to IIA violations in our divisive normalization and information-processing models. Specifically, we show that IIA violations are equivalent to differences in the divisive factor and the marginal cost of reducing stochasticity, in the divisive normalization and information-processing models, respectively. However, divisive normalization also places limits on context effects. It does not allow (stochastic) preference reversals: If one alternative is chosen more often than a second in one choice set, then it is chosen more often than the second in every choice set. Also, we provide a novel testable restriction by showing that any choice sets, which share at least two items, can be unambiguously ranked in terms of stochasticity of the choices made on those sets.

We now offer more detail on each of the three models in our equivalence result. The divisive normalization model works by first assigning a set-independent value to each possible choice alternative. Within a particular choice set, these values are then re-scaled by a factor that depends on the available alternatives. Specifically, this set-dependent factor is equal to the sum of the values of the items in the choice set plus a constant. These re-scaled values are then multiplied by a constant that serves to set upper and lower bounds on the set of possible choice probabilities. Both constants are assumed to be set-independent. A random error term is added to each re-scaled value and the alternative with the highest sum of re-scaled value plus error is then chosen. The error term accounts for observed neurobiological stochasticity in the representation of value.

Our information-processing model views choices as optimally balancing the cost and benefit of decreasing stochasticity in choice. Decreasing stochasticity in choices decreases information entropy, which, on physical grounds, must be costly (Landauer, 1961). We model this cost as an increasing function of the associated decrease in Shannon entropy (Shannon, 1948), where the specific functional form can depend on the choice set faced. Decreasing stochasticity in choices benefits the decision maker by increasing the chance of selecting the highest valued alternative. Therefore, a decision maker faces a trade-off between the cost and benefit of reducing stochasticity, which our information-processing model optimally balances. (Other work in decision theory models this feature as a preference for hedging or stochasticity; see, for example, Machina, 1989, or Agranov and Ortoleva, 2017.) Our general information-processing analysis identifies a broad family of divisive normalization models. We also identify the specific functional form for the cost of reducing entropy that corresponds to the additive normalization factor which is most often used in the empirical neuroscience literature, thereby providing an information-processing foundation for the latter.

Our axiomatic characterization consists of a nested set of six behaviorally testable axioms that together are equivalent to the divisive normalization model. The axioms are layered in a way that allows for different versions of the model to be independently tested. The first two axioms by themselves characterize a version of the divisive normalization model where the divisive term can be any strictly positive choice set-dependent factor. Adding in the next two axioms ensures that the divisive term is equal to the sum of the values of the items in the choice set plus a constant. The final two axioms ensure the values and the two constants in the model are strictly positive.

Returning to our equivalence result, we think of its three components as answering, respectively, the "how," the "why," and the "what" of choice behavior. The divisive normalization model explains *how* the brain makes choices, namely through the normalization computation. The information-processing formulation provides some insight into *why* that computation is used. The axiomatic characterization outlines exactly *what* behavior arises. In this way, our theory provides a unified answer to the "how," the "why," and the "what" of choice behavior.

2 Literature Review

In addition to the divisive normalization literature reviewed in the introduction, our paper relates to the literature on random choice following the Luce rule (Luce, 1959) as well as the literature that employs Shannon entropy in models of decision making.

Our use of the Gumbel error term connects the divisive normalization model to the Luce rule — without the divisive re-scaling step, our model would be equivalent to the Luce rule. The connection between the Gumbel error and the Luce rule is well known (see Luce and Suppes, 1965, who attribute the result to Holman and Marley). However, our model does not contain the Luce rule as a special case since our divisive re-scaling factor cannot be constant across sets. The divisive normalization model with Gumbel error is an instance of the more general Set-Dependent-Luce model (Marley, Flynn, and Louviere, 2008), which is equivalent to a version of our model where the re-scaling factor is allowed to be any strictly positive set-dependent term. We also relate to the larger literature generalizing the Luce rule to accommodate a wider range of empirical phenomena (e.g., Echenique, Saito, and Tserenjigmid, 2014; Gul, Natenzon, and Pesendorfer, 2014; Ravid, 2015; Echenique and Saito, 2015; Tserenjigmid, 2016). Our paper is similar to these in that we can also accommodate a wider range of behavior, but we differ in that we use a neurobiological motivated functional form.

Our paper connects to previous work that employs Shannon entropy in models of decision making. Some of these previous papers also use entropy to model the cost of reducing stochasticity (Mattsson and Weibull, 2002; Fudenberg et al., 2015), while others have used entropy to model a taste for variety (Anderson et al., 1992; Swait and Marley, 2013). Several of these papers trace out a similar mathematical connection between entropy and the probability formulas as do we. In fact, this mathematical connection goes back much further to the physics connecting Helmholtz free energy to the Boltzmann distribution (see Mandl, 1988). Of particular note is Swait and Marley (2013), who studied a model equivalent to the Set-Dependent Luce model discussed above. Thus, the model of Swait and Marley (2013) is equivalent to a divisive normalization model with any strictly positive divisive factor.

The use of Shannon entropy in our information-processing formulation is also a point of connection with the rational-inattention literature initiated by Sims (1998, 2003). Recently, the rational-inattention notion has been applied to stochastic choice settings with uncertain values for the alternatives (Matějka and McKay, 2014; Caplin and Dean, 2015). This leads to an information-processing task: Determine the optimal cost to incur in learning about these uncertain values. By contrast, the information-processing task we consider is an efficient reduction in the intrinsic stochasticity of choosing among alternatives.

Finally, our paper relates to the efficient coding literature from neuroscience, which argues that neural processes should be efficiency-promoting (Attneave, 1954; Barlow, 1961). Within this literature, a number of papers advance efficiency arguments for the divisive normalization computation (see Carandini and Heeger, 2012 for a review). However, this literature differs from our paper by being concerned with applying divisive normalization to sensory processing and how the brain can efficiently store and represent sensory information. For example, one thread of this literature shows how divisive normalization can be used to de-correlate the activity of different neurons to reduce redundancy in how sensory input is represented (see for example, Schwartz and Simoncelli, 2001). We see our information-processing model as building a parallel efficiency argument, with a focus on the choice domain.

3 Foundations of the Normalization Model

We now provide two foundations for the divisive normalization model: an information-processing model and an axiomatic characterization. The information-processing model provides insight into why the normalization computation is used — namely, because it optimally balances the costs and benefits of reducing stochasticity. The axiomatic characterization pins down what behavior the normalization model allows by providing a set of testable restrictions. The main result of this section is an equivalence theorem uniting all three models in terms of the behavior they imply. For example, any behavior that arises from the normalization model optimally solves the information-processing model and obeys our axioms. The equivalence works in all directions.

We begin with the formal framework, which we maintain throughout. Let X be a finite set consisting of all the alternatives from which the decision maker may be able to choose, which we assume contains at least four items. Let $\mathcal{A} = 2^X \backslash \emptyset$ be the collection of all non-empty subsets of X, to be thought of as the possible choice sets the decision maker may face. As a convenient shorthand, for any function $f : X \to \mathbb{R}$ and $A \in \mathcal{A}$, we set $f(A) = \sum_{x \in A} f(x)$.

The choice behavior of the decision maker is described by a random choice rule ρ that assigns a full-support probability measure to every choice set $A \in \mathcal{A}$. We limit attention to full-support measures because the divisive normalization model uses a full-support error term that implies all available options have a non-zero chance of being chosen. Formally, for any choice set A, define

$$\Delta A := \left\{ p : A \to [0,1] \mid \sum_{x \in A} p(x) = 1 \right\},$$

which is the set of all probability measures on A. A random choice rule is then any function $\rho : X \times \mathcal{A} \to [0,1]$ such that $\rho(x, A) > 0 \iff x \in A$ and $\rho(\cdot, A) \in \Delta A$ for each $A \in \mathcal{A}$.[1] The interpretation is that $\rho(x, A)$ is the probability that the decision maker chooses alternative x when faced with choice set A. To avoid degenerate cases, we assume ρ assigns at least three distinct probabilities in choice set X. In other words, there exists $x, y, z \in X$ such that $\rho(x, X)$, $\rho(y, X)$, and $\rho(z, X)$ are all distinct.

We now define the divisive normalization model using the standard functional form widely employed in the neuroscience literature (Carandini and Heeger, 2012). The normalization model generates a stochastic and set-dependent utility for each alternative, and the highest utility alternative is then chosen. The stochastic utility of alternative x in set A is

$$\gamma \frac{v(x)}{\sigma + v(A)} + \varepsilon_x.$$

The function v provides a set-independent value of each alternative. These set-independent

[1] More precisely, by $\rho(\cdot, A) \in \Delta A$ we mean that the restriction of $\rho(\cdot, A)$ to A is in ΔA. Throughout, we will often find it convenient to treat $\rho(\cdot, A)$ as a function from A to $[0,1]$, since the values of $\rho(\cdot, A)$ outside of A are always zero.

values are then divisively re-scaled by the factor $\sigma + v(A)$ which is a constant plus the sum of values of items in the choice set. The term γ is a strictly positive constant that sets an upper and lower bound on achievable choice probabilities, with a larger γ corresponding to more relaxed bounds. We will make these claims about γ more precise at the end of this section. Lastly, ε_x is a random noise term that is i.i.d. across the alternatives and that we assume follows a Gumbel distribution with location 0 and scale 1.

We define the normalization model as the set of choice probabilities that can be generated using this utility form. We state this more formally as follows.

Definition 1. A random choice rule ρ has a **divisive normalization representation** if there exists $v : X \to \mathbb{R}_{++}$, $\sigma > 0$, and $\gamma > 0$, such that for any $A \in \mathcal{A}$ and $x \in A$

$$\rho(x, A) = \Pr\left(x \in \arg\max_{y \in A} \gamma \frac{v(y)}{\sigma + v(A)} + \varepsilon_y\right),$$

where ε_y is distributed i.i.d. Gumbel (0,1).

The divisive normalization model is distinct from random utility. On the one hand, the presence of the divisive factor $\sigma + v(A)$ allows a set dependence absent in a standard random utility model. On the other hand, random utility allows for more general assumptions on the error term. The Gumbel distribution we assume does arise in a number of settings — in particular, as the asymptotic distribution of the maximum of a sequence of i.i.d. normal random variables. See, e.g., David and Nagaraja (2003). Also note that the continuous nature of the Gumbel distribution ensures there there is always a strict utility maximizing element, so we do not have to worry about ties.

Information-Processing Model

In our information-processing model, the decision maker balances the expected utility of a given choice rule against the cost involved in reducing stochasticity in choices. The value of alternative x is given by $v(x)$ and the expected utility of choice rule ρ on choice set A is

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

The information-processing costs of a particular rule ρ come from the reduction in Shannon entropy (Shannon, 1948) relative to the fully stochastic case. Shannon entropy measures the degree of stochasticity in behavior, where a higher degree of stochasticity implies higher entropy. We will find it useful to define entropy generally for any function $f : A \to \mathbb{R}_+$, as follows:

$$H(f) := -\sum_{x\in A} \frac{f(x)}{f(A)} \ln \frac{f(x)}{f(A)},$$

where $0\ln 0$ is understood to equal zero. For a probability measure $p \in \Delta A$, $H(p)$ equals the associated Shannon entropy of p on A. The maximum entropy of any function defined on set A is $\ln|A|$, which is achieved by the uniform measure that assigns the same probability to each alternative in A. Therefore, the entropy reduction achieved by any function $f : A \to \mathbb{R}_+$ is

$$\Delta H(f) := \ln|A| - H(f).$$

The total cost of random choice rule ρ on choice set A will be a strictly increasing function of the entropy reduction achieved by ρ on A, where the shape of the function can depend on the choice set faced. We impose standard regularity conditions that the function is continuously differentiable and convex. Combining this with the expected value of a choice rule yields our definition of optimal behavior with costly information processing.

Definition 2. A random choice rule ρ has an **information-processing representation** if there exists function $v : X \to \mathbb{R}_{++}$ and, for each $A \in \mathcal{A}$, a strictly increasing, convex, and continuously differentiable function $C_A : \mathbb{R} \to \mathbb{R}$ such that for any $A \in \mathcal{A}$

$$\rho(\cdot, A) \in \arg\max_{p\in\Delta A} \sum_{x\in A} p(x)\, v(x) - C_A(\Delta H(p)). \tag{1}$$

A choice rule ρ having an information-processing representation is actually equivalent to a more general normalization computation where the divisive factor can be any strictly positive set-dependent function[2]. To get our specific neurobiologically motivated functional form for normalization, we impose additional restrictions on C_A. First, for any $A \in \mathcal{A}$, $v : X \to \mathbb{R}_{++}$, $\sigma > 0$, and $\gamma > 0$, define

$$\delta(A; v, \sigma, \gamma) := \Delta H\left(\exp\left(\gamma \frac{v(x)}{\sigma + v(A)}\right)\right).$$

In words, $\delta(A; v, \sigma, \gamma)$ equals the entropy reduction achieved by the function that maps x to $\exp\left(\gamma\frac{v(x)}{\sigma+v(A)}\right)$. This turns out to be equal to the entropy reduction achieved on A by the normalization computation using (v, σ, γ).

We say an information-processing representation $\left(v, \{C_A\}_{A\in\mathcal{A}}\right)$ obeys the Marginal Cost

[2] For details, see online appendix at http://www.adambrandenburger.com/articles/papers

Condition (MCC) if there exist $\sigma > 0$ and $\gamma > 0$ such that, for each $A \in \mathcal{A}$,

$$C'_A\left(\delta\left(A; v, \sigma, \gamma\right)\right) = \frac{\sigma + v\left(A\right)}{\gamma}$$

for each $A \in \mathcal{A}$. The MCC places a restriction on the marginal cost of entropy reduction at the choice probabilities generated by the divisive normalization model. Specifically, the marginal cost has to vary linearly with the total value of items in the set. This places a neurobiologically motivated restriction on the functional forms in the information-processing model.

Axiomatic Characterization

Our axiomatic characterization gives six testable behavioral restrictions, arranged into three nested groups, which are jointly equivalent to the divisive normalization model. The axioms are layered in a way that allows for different versions and aspects of the normalization model to be independently tested.

To state the axioms compactly, it will be useful to define a few terms. We say the pair (x, y) is distinguishable in A if $x, y \in A$ and $\rho\left(x, A\right) \neq \rho\left(y, A\right)$. For any (x, y) distinguishable in A, we define

$$R_{xy}\left(A\right) := \left(\ln \frac{\rho\left(x, A\right)}{\rho\left(y, A\right)}\right)^{-1} \ln \frac{\rho\left(x, X\right)}{\rho\left(y, X\right)}.$$

The number $R_{xy}\left(A\right)$ measures how the choice probability ratio between x and y differs across choice sets A and X. Larger $R_{xy}\left(A\right)$ means that this ratio is closer to 1 in A than in X. This suggests $R_{xy}\left(A\right)$ is related to the divisive factor, since a larger divisive factor pushes the re-scaled values, and hence the choice probabilities, closer together. In fact, if ρ has a divisive normalization representation (v, σ, γ), and (x, y) is distinguishable in A, then

$$R_{xy}\left(A\right) = \frac{\sigma + v\left(A\right)}{\sigma + v\left(X\right)}. \tag{2}$$

We will discuss the proof of this fact in our discussion around Equation (3) below.

We are now ready to state our axioms.

Axiom 1 (Order)**.** *Let $A, B \in \mathcal{A}$ and $x, y \in A \cap B$. Then $\rho\left(x, A\right) \geq \rho\left(y, A\right)$ if and only if $\rho\left(x, B\right) \geq \rho\left(y, B\right)$.*

Our first axiom requires a set-independent ordinal ranking of the alternatives, in the sense that whether x is chosen more often than y is consistent across all choice sets. In the divisive normalization model, this ordinal ranking follows the ranking given by v, a fact we explore further in the next section.

Axiom 2 (Divisive Factoring)**.** *If (x, y) and (x', y') are distinguishable pairs in A, then*

$$R_{xy}(A) = R_{x'y'}(A).$$

Our second axiom states that the value of $R_{xy}(A)$ does not depend on the specific x, y pair used. This is an immediate implication of Equation (2) and captures the fact that the divisive factor in the normalization model depends only on the choice set and not on the particular item being re-scaled.

Together, our first two axioms characterize a version of the normalization model where the divisive factor is allowed to be any strictly positive set-dependent function.[3]

Axiom 3 (Additive Separability)**.** *For any $z \in A$*

$$R_{xy}(A) - R_{xy}(A \setminus \{z\}) = R_{xy}(X) - R_{xy}(X \setminus \{z\}),$$

where (x, y) is distinguishable in all four sets used in the above equation.

Our third axiom says that the effect of removing an item on the divisive factor does not depend on the other alternatives in the choice set. This captures the additive separability of the divisive factor across the items.

Axiom 4 (Separability by Values)**.** *Suppose the pairs (x, z), (x', z'), (y, z), and (y', z') are each distinguishable in the set that contains only that pair and that (x', y') is distinguishable in X. Then*

$$\frac{R_{xz}(\{x, z\}) - R_{yz}(\{y, z\})}{R_{x'z'}(\{x', z'\}) - R_{y'z'}(\{y', z'\})} = \ln \frac{\rho(x, X)}{\rho(y, X)} \left(\ln \frac{\rho(x', X)}{\rho(y', X)} \right)^{-1}.$$

We can interpret the ratio $\rho(x, X) / \rho(y, X)$ as providing a measure of x's value relative to y's, since we expect more valuable items to be chosen more often. Under this interpretation, the fourth axiom relates the relatives values of x, y and x', y' to the divisive factor involving those alternatives. This ensures the divisive factor is additively separable using the values of the alternatives.

Our first four axioms together characterize a version of normalization where v, σ, and γ are not necessarily positive.[4] For this, we require two additional axioms.

[3] See online appendix for the proof.

[4] The sum $\sigma + v(A)$ would have to have the same sign for all $|A| \geq 2$ to avoid violating Axiom 1.

Axiom 5 (Strictly Positive v and γ)**.** *Suppose $A \in \mathcal{A}$ contains distinguishable pair (x, y), and let $z, z' \in A\backslash\{x, y\}$ be such that $\rho(z, X) > \rho(z', X)$. Then*

$$R_{xy}(A) > R(A\backslash\{z'\}) > R(A\backslash\{z\}).$$

Our fifth axiom implies two facts: (1) the divisive factor strictly increases when adding alternatives to the choice set, and (2) the divisive factor increases by more when adding alternatives with a higher choice probability. From Equation (2), we can easily see that the first fact corresponds to $v > 0$. The second follows from Equation (2) under the assumption that alternatives with a higher choice probability have a higher value for v. This assumption requires $\gamma > 0$, since $\gamma < 0$ would allow the re-scaled value $\gamma v(x) / (\sigma + v(A))$ to decrease in $v(x)$. Therefore, Axiom 5 corresponds to v and γ being strictly positive.

Axiom 6 (Strictly Positive σ)**.** *If (x, y) is distinguishable in both A and $A \cup B$, and (x', y') is distinguishable in B, then*

$$R_{xy}(A) + R_{x'y'}(B) > R_{xy}(A \cup B).$$

Our final axiom imposes strict subadditivity on the R_{xy} function, which captures the fact that $\sigma > 0$. To see why, note that, applying Equation (2), the difference between the left and right-hand sides of the inequality equals

$$\frac{\sigma + v(A \cap B)}{\sigma + v(X)},$$

which must be positive because v and σ are strictly positive.

Equivalence Result

The main result of this paper establishes a three-way equivalence uniting the divisive normalization model, our information-processing model, and our axiomatic characterization. The unification of the three models works on the level of behavior. Any choice probabilities that fit into one of the three models necessarily must fit into all three.

Theorem 1. *For any random choice rule ρ the following are equivalent:*

1. *ρ has a divisive normalization representation,*

2. *ρ has an information-processing representation that obeys the MCC,*

3. *ρ obeys Axioms 1-6.*

Proof. See the Appendix. □

We think of the three models in our equivalence result as answering, respectively, the "how," the "why," and the "what" of choice behavior. According to existing work in neuroscience, the divisive normalization model explains *how* the brain makes choices, namely through the normalization computation. The information-processing formulation provides some insight into *why* that computation is used. The axiomatic characterization outlines exactly *what* behavior arises.

The proof of Theorem 1 proceeds by establishing that all three parts are equivalent to the statement that there exists a function $v : X \to \mathbb{R}_{++}$ and constants $\sigma > 0$ and $\gamma > 0$ such that

$$\rho\left(x,A\right) = \frac{\exp\left(\gamma \frac{v(x)}{\sigma+v(A)}\right)}{\sum_{y\in A}\exp\left(\gamma \frac{v(y)}{\sigma+v(A)}\right)} \tag{3}$$

for all $A \in \mathcal{A}$ and $x \in A$.

We can use Equation (3) to prove the claims we made about the role of γ. First, rewrite Equation (3) as

$$\rho\left(x,A\right) = \frac{1}{1+\sum_{y\in A\setminus\{x\}}\exp\left(\gamma \frac{v(y)-v(x)}{\sigma+v(A)}\right)}.$$

Since all the parameters are strictly positive, whenever $x, y \in A$,

$$-\gamma \leq \gamma \frac{v\left(x\right)-v\left(y\right)}{\sigma+v\left(A\right)} \leq \gamma.$$

Combining these inequalities with our rewritten version of Equation (3) yields

$$\frac{1}{1+\left(|A|-1\right)\exp\left(\gamma\right)} \leq \rho\left(x,A\right) \leq \frac{1}{1+\left(|A|-1\right)\exp\left(-\gamma\right)}.$$

These inequalities confirm that γ determines an upper and lower bound on the possible choice probabilities. As $\gamma \to \infty$ these inequalities give only the trivial statement $\rho\left(x,A\right) \in [0,1]$, and as $\gamma \to 0$ they force $\rho\left(x,A\right) = \frac{1}{|A|}$ for all x, A.

We can also use Equation (3) to establish our claim regarding $R_{xy}\left(A\right)$ in Equation (2). Equation (3) implies that, for all $x, y \in A$

$$\frac{\rho\left(x,A\right)}{\rho\left(y,A\right)} = \exp\left(\gamma\frac{v\left(x\right)-v\left(y\right)}{\sigma+v\left(A\right)}\right),$$

from which,

$$R_{xy}\left(A\right) = \left(\gamma\frac{v\left(x\right)-v\left(y\right)}{\sigma+v\left(A\right)}\right)^{-1}\left(\gamma\frac{v\left(x\right)-v\left(y\right)}{\sigma+v\left(X\right)}\right),$$

which simplifies to Equation (2).

4 Identifying Neural and Behavioral Parameters

Divisive normalization has been used by neuroscientists to model how neural data determines choices. Whether neural data can be identified from choices has not yet been addressed. In this section we prove a result showing that choice alone can be used to uniquely specify neural observables within the framework of the normalization model. In more detail, the re-scaled values in the divisive normalization model are used to match the neurally observable subjective value function, experimentally measured as the number of action potentials per second (or "firing rate") of individual neurons. We prove that the re-scaled values are fully identified from choice behavior alone, and conversely, that the re-scaled values fully determine the (stochastic) choice behavior. In other words, there exists a one-to-one relationship between the neurally measurable subjective value function and the behavior it generates. This result is important to empirical neuroscientists. It allows novel and precise predictions linking neural and choice data. This result is also important for economists, since it allows the application of insights from neural analysis without neural data.

We prove the one-to-one relationship as a corollary to our more general uniqueness result on the parameters of the divisive normalization model. We start this section by presenting this more general identification result. We then discuss its implications for neural and choice parameters in the form of two corollaries. We end by providing the proof of the identification result, which builds on the notation and logic (notably Equation 2) from the axiomatic characterization in the previous section.

Proposition 1. *Suppose ρ has divisive normalization representation (v, σ, γ). Then (v', σ', γ') is also a divisive normalization representation of ρ if and only if $\gamma = \gamma'$ and there exists $\alpha > 0$ such that $(v, \sigma) = \alpha (v', \sigma')$.*

Proposition 1 establishes that the v and σ parameters in the divisive normalization model are jointly unique up to a strictly positive multiplicative constant, while γ is fully unique.

A transformation of these parameters of particular interest is the re-scaled value of each alternative x in choice set A, that is

$$\gamma \frac{v(x)}{\sigma + v(A)}.$$

As discussed above, these re-scaled values model the neurally measurable subjective value function in the divisive normalization framework. An immediate corollary of Proposition

1 is that these re-scaled values are fully identified from choice behavior. Conversely, the definition of the divisive normalization model immediately implies that these re-scaled values fully determine the choice probabilities.

Corollary 1. *Suppose ρ has divisive normalization representation (v, σ, γ). Then (v', σ', γ') is also a divisive normalization representation of ρ if and only if for every (x, A) in $X \times \mathcal{A}$:*

$$\gamma \frac{v(x)}{\sigma + v(A)} = \gamma' \frac{v'(x)}{\sigma' + v'(A)}.$$

Corollary 1 establishes two facts. First, if (v, σ, γ) and (v', σ', γ') have the same re-scaled values, then they represent the same behavior. Second, if (v, σ, γ) and (v', σ', γ') represent the same behavior, then they must have the same re-scaled values. This creates an exact one-to-one relationship between choice behavior and the neurally observable subjective value function, within the divisive normalization model. This allows for precise predictions on neural data from choice data alone, enabling new types of experimental hypotheses for empirical neuroscientists. This result is also relevant for data sets containing choices alone, since it justifies the application of insights based on neural analysis without neural data. For example, some researchers have suggested that the neurally measured subjective value function is the correct value level for welfare analysis (Loewenstein, Rick, and Cohen, 2008), and our result suggests this analysis can be performed with choice data alone.

Corollary 1 works because the re-scaling step uses the values of the alternatives. We can measure $v(x)$ by how much the choice probabilities of other alternatives change when x is added to the set. A larger $v(x)$ causes more re-scaling which pushes the values and choice probabilities closer together. If, instead, the re-scaling was done via a set-dependent factor that did not depend on the values then the re-scaled values would not be unique.

For example, suppose we assigned stochastic utility to alternative x in set A of

$$\frac{v(x)}{F(A)} + \varepsilon_x,$$

where $F(A)$ is any strictly positive set-dependent function and ε_x is an i.i.d. random variable. Now define $v'(x) := v(x) + \alpha$ for some constant $\alpha \neq 0$. Then for any choice set A and $x, y \in A$,

$$\frac{v(x)}{F(A)} - \frac{v(y)}{F(A)} = \frac{v'(x)}{F(A)} - \frac{v'(y)}{F(A)}.$$

This is enough to ensure that (v, F) and (v', F) deliver the same choice probabilities, while having different re-scaled values. By contrast, in the divisive normalization model, changing from v to v' also changes the re-scaling factor which impacts the choice probabilities.

The second set of parameters we are interested in identifying consists of the untransformed values without re-scaling. Proposition 1 shows that these are unique only up to a multiplicative constant, as is the case in other stochastic choice models, such as the Luce rule. This degree of uniqueness is enough to determine a unique ordinal ranking over the choice alternatives and choice sets. Define

$$E_A\left[v\left(x\right)|\rho\right] := \sum_{x\in A}\rho\left(x,A\right)v\left(x\right),$$

which is the expected value from choice rule ρ on set A using values v.

Corollary 2. *Suppose ρ has two divisive normalization representations (v,σ,γ) and (v',σ',γ). Then:*

1. *$v\left(x\right)\geq v\left(y\right)$ if and only if $v'\left(x\right)\geq v'\left(y\right)$ for all $x,y\in X$.*
2. *$E_A\left[v\left(x\right)|\rho\right]\geq E_B\left[v\left(x\right)|\rho\right]$ if and only if $E_A\left[v'\left(x\right)|\rho\right]\geq E_B\left[v'\left(x\right)|\rho\right]$*

Corollary 2 says that the divisive normalization model uniquely ranks the alternatives by their values and choice sets by their expected values. In this sense, the divisive normalization model provides a well-defined ordinal preference over alternatives and choice sets.

Proof of Proposition 1. The "if direction" is obvious. For the other direction, suppose that (v,σ,γ) and (v',σ',γ') are both divisive normalization representations of ρ. If choice set A contains a distinguishable pair (x,y), then we know

$$\frac{\sigma+v\left(A\right)}{\sigma+v\left(X\right)}=\frac{\sigma'+v'\left(A\right)}{\sigma'+v'\left(X\right)} \tag{4}$$

since, using Equation (2), both sides of the equation equal $R_{xy}\left(A\right)$.

Now define

$$\alpha := \frac{\sigma+v\left(X\right)}{\sigma'+v'\left(X\right)}.$$

It is clear that $\alpha>0$ since all the terms are strictly positive. By our assumptions on ρ, we can find $x,y,w\in X$ such that $\rho\left(x,X\right)$, $\rho\left(y,X\right)$, and $\rho\left(w,X\right)$ are all distinct. By Theorem 1, we know ρ obeys Axiom 1, which implies all pairs from $\{x,y,w\}$ are distinguishable in every set that contains them. Hence, whenever $x,y\in A$, we can rearrange Equation (4) to get

$$v\left(A\right)=\alpha v'\left(A\right)+\alpha\sigma'-\sigma.$$

Therefore, whenever $x,y\in A\cap B$, we get

$$v\left(A\right)-v\left(B\right)=\alpha\left(v'\left(A\right)-v'\left(B\right)\right).$$

Hence, for any $z \in X\setminus \{x, y\}$, we have that

$$v(z) = v(\{x, y, z\}) - v(\{x, y\}) = \alpha (v'(\{x, y, z\}) - v'(\{x, y\})) = \alpha v'(z).$$

We can apply the same logic with $\{x, w\}$ taking the role of $\{x, y\}$ to prove $v(y) = \alpha v'(y)$. We can also have $\{w, y\}$ take the role of $\{x, y\}$ to prove $v(x) = \alpha v'(x)$. Therefore, we have shown $v(z) = \alpha v'(z)$ for all $z \in X$. Combining $v(X) = \alpha v'(X)$ with the definition of α, it follows that $\sigma = \alpha\sigma'$.

To prove $\gamma = \gamma'$, we can use Equation (3) to get

$$\exp\left(\gamma \frac{v(x) - v(y)}{\sigma + v(X)}\right) = \exp\left(\gamma' \frac{v'(x) - v'(y)}{\sigma' + v'(X)}\right),$$

since both sides equal $\rho(x, A) / \rho(y, A)$. Since (x, y) is distinguishable, we know that neither side of the equation is equal to 1, so that $v(x) - v(y) \neq 0$. Using $(v, \sigma) = \alpha (v', \sigma')$ and taking natural logs of both sides gives

$$\gamma \frac{v(x) - v(y)}{\sigma + v(X)} = \gamma' \frac{v(x) - v(y)}{\sigma + v(X)}.$$

And the desired result follows using $v(x) - v(y) \neq 0$. □

5 Context Effects

In this section, we use our equivalence result to study how the divisive normalization model handles context effects. We use a standard notion of a context effect as a departure from the Independence of Irrelevant Alternatives (IIA) property developed by Luce (1959). Following the logic of our equivalence result, we find the precise analogs of IIA violations in our information-processing and divisve normalization models. In the normalization model, IIA violations are equivalent to differences across choice sets in the divisive factor. In the information-processing model, IIA violations are equivalent to changes in the marginal cost of reducing stochasticity across different sets. We also discuss the limitations on context effects implied by the divisive normalization model, which suggest a new testable prediction.

The IIA property requires that

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

whenever $x, y \in A \cap B$. In words, this says the relative choice probability between two alternatives is independent of the other alternatives in the set. For expositional purposes, we only consider probability ratios where $\rho(x, A)/\rho(y, A) \geq 1$. This allows us to interpret larger ratios as being further from the equal-probability case and will simplify the statements of results. This simplification is without loss of generality since we can just invert any ratio that is smaller than 1.

Proposition 2. *Suppose ρ has normalization representation (v, σ, γ). Consider $A, B \in \mathcal{A}$ and $x, y \in A \cap B$ such that $\rho(x, B)/\rho(y, B) \geq 1$. Then the following are equivalent:*

1. $\frac{\rho(x,A)}{\rho(y,A)} > \frac{\rho(x,B)}{\rho(y,B)}$,
2. $\sigma + v(A) < \sigma + v(B)$,
3. *ρ has an information-processing representation $\left(v, \{C_A\}_{A\in\mathcal{A}}\right)$ where $\delta(A) < \delta(B)$.*

Proof. The proof of Theorem 1 establishes that we can use the same parameters (v, σ, γ) in the normalization representation as in the information-processing representation and associated MCC. This, along with Theorem 1, immediately establishes the equivalence of Parts 2 and 3.

For the equivalence of Parts 1 and 2, let $x, y \in A \cap B$ such that $v(x) > v(y)$. From Equation (3), this implies $\rho(x, C) > \rho(y, C)$ whenever $x, y \in C$. By the definition of R_{xy}, we then know that

$$R_{xy}(A) < R_{xy}(B) \iff \frac{\rho(x, A)}{\rho(y, A)} > \frac{\rho(x, B)}{\rho(y, B)}.$$

By Equation (2), this can be written as,

$$\frac{\sigma + v(A)}{\sigma + v(X)} < \frac{\sigma + v(B)}{\sigma + v(X)} \iff \frac{\rho(x, A)}{\rho(y, A)} > \frac{\rho(x, B)}{\rho(y, B)},$$

which gives us the desired result. □

The equivalence between Parts 1 and 2 in Proposition 2 establishes that a larger divisive factor is equivalent to IIA violations that move the probability ratios closer to equal probability. This demonstrates that the normalization model captures context effects through the divisive factor. Previous papers have noted this relationship between the divisive factor and IIA violations in more limited contexts. For example, Louie, Khaw, and Glimcher (2013) studied this feature of the normalization model in three-item choice sets, while our result works for a choice set of any size. The equivalence between Parts 1 and 3 in Proposition 2 also shows that IIA violations correspond to changes in $\delta(A)$ in the information-processing model. Recall that $\delta(A)$ equals the marginal cost of reducing stochasticity on set A. This suggests a reason why the brain might have evolved to use a context-dependent computation — namely, to account for differing costs of stochasticity reduction in different sets.

We can use Proposition 2 to better understand how and why the divisive normalization model accounts for previously studied context effects. For example, the well-known Compromise and Attraction Effects create IIA violations by adding a third alternative to a two-alternative choice set (Huber, Payne, and Puto, 1982; Simonson, 1989). Louie, Khaw, and Glimcher (2013) found IIA violations when they replaced the worst alternative in a three-alternative choice set with a slightly improved option. Specifically, they found this increased choice stochasticity between the two unchanged alternatives, in the sense of pushing the probability ratio closer to 1. Proposition 2 shows exactly how divisive normalization can accommodate these IIA violations. It does so because adding an alternative or raising the value of an alternative both change the divisive factor that drives context effects. Proposition 2 also suggests why these context effects occur, namely, because of changes in the marginal cost of reducing stochasticity across choice sets. This interpretation lines up particularly nicely with the result in Louie, Khaw, and Glimcher (2013), since, under Proposition 2, raising the value of the worst alternative increases the marginal cost of reducing stochasticity, which naturally leads to more stochastic choices.

It is also important to note limitations on the types of context effects which divisive normalization can accommodate. For example, the Attraction Effect is often associated with (stochastic) preferences reversals, where an alternative x is chosen more often than y in one choice set but not in another. However, the divisive normalization model can never achieve these reversals, which is an immediate implication of Axiom 1. Instead, the values of normalization model will create a ordinal ranking of the alternative, where higher alternatives are always chosen more often. This also implies that the divisive normalization model obeys Weak Stochastic Transitivity (Block and Marschak, 1960), which requires that if $\rho(x, \{x, y\}) \geq \frac{1}{2}$ and $\rho(y, \{y, z\}) \geq \frac{1}{2}$, then $\rho(x, \{x, z\}) \geq \frac{1}{2}$.

Another restriction on context effects is that the direction of the IIA violation must be consistent across all pairs when moving across choice sets. In other words, if one pair of items violates IIA by being further from the equal-probability case in set A versus set B, then the same must be true of all pairs that appear in both sets. We state this formally as:

Corollary 3. *Suppose ρ has divisive normalization representation (v, σ). Let $x, y, x', y' \in A \cap B$ such that $v(x) > v(y)$ and $v(x') > v(y')$. Then*

$$\frac{\rho(x, A)}{\rho(y, A)} > \frac{\rho(x, B)}{\rho(y, B)} \Rightarrow \frac{\rho(x', A)}{\rho(y', A)} > \frac{\rho(x', B)}{\rho(y', B)}.$$

Corollary 3 follows immediately from Proposition 2, because the equivalence between

Parts 1 and 2 imply that whether

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

holds for any particular pair can be determined by an inequality that depends on the sets as a whole.

Another way to interpret Corollary 3 is in terms of the relative stochasticity of the choice sets. The choice between x and y is more stochastic when the probability ratio between x and y is closer to 1. Therefore, Corollary 3 says any two sets (that share at least two items) can be unambiguously ranked by how stochastic the choices are on them. This provides a novel testable restriction on the divisive normalization model.

6 Concluding Remarks

In this paper, we studied three different models that each presented a different perspective on choice behavior. The divisive normalization model says *how* the brain makes choices, namely, via the neurobiologically-motivated normalization computation. The information-processing formulation provides some insight into *why* that computation is used, namely, because it optimally balances the benefits and costs of reducing stochasticity. The axiomatic characterization pinpoints exactly *what* behavior arises by providing a set of testable behavioral predictions. Our main result proves an equivalence between these three models, uniting them into a single theory that can simultaneously address the "how," the "why," and the "what" of choice behavior.

We also explore how the parameters of the divisive normalization model can be identified from behavior, and what that tells us about the link between observable choice and observable neural quantities. We prove that, in the divisive normalization model, there is a one-to-one relationship between the neurally measurable subjective value function and the behavior it generates. This creates a theoretical foundation for work that links neural and behavioral data, and indicates that inference about neural variables can be made from choice behavior alone.

Lastly, we use our equivalence result to study how the divisive normalization model handles context effects. The divisive normalization model allows for context effects through changes in the divisive factor. When the divisive factor is equal across two choice sets, the choices on those sets will be context independent in the sense of obeying the Independence of Irrelevant Alternatives (IIA). We use our equivalence result to provide behavioral and information-processing analogs to the changes in divisive factor that drive context effects.

We then apply these analogs to shed new light on existing empirical work, and to provide a novel testable prediction on the type of context effects allowed in the divisive normalization model.

We conclude by commenting on one of the more unusual aspects our paper, relative to the economics literature — namely our inclusion of a neurobiologically-motivated functional form in a choice model. The inclusion of this aspect is motivated, in part, by the argument due to Simon (1955, p. 99) that a theory of decision making should be consistent "with the access to information and the computational capacities that are actually possessed by the organism." At the time of Simon's writing, the development of such a theory was hindered by a lack of empirical knowledge about precisely such information and computational capacities — a fact which Simon himself noted (Simon 1955, p. 100).

In the decades since then, advances in neuroscience have taught us a lot about the actual decision processes of various organisms, humans included. By capitalizing on these advances, we have been able to build a theory of decision-making consistent with how the human brain actually makes choices, and, in this way, to advance Simon's argument. With this, we hope to have taken a step towards reconciling traditional approaches to decision-making with the fact that all behavior making must, ultimately, have a physical implementation.

Appendix

A Proof of Theorem 1

We will show that all three parts of Theorem 1 are equivalent to Equation (3). Additionally, while not strictly required for Theorem 1, our proof will show that the same set of parameters (v, σ, γ) are used in Equation (3), the divisive normalization model, and the information-processing model. In other words, our proof will also show the following three statements are equivalent:

1. (v, σ, γ) satisfies Equation (3),
2. (v, σ, γ) is a divisive normalization representation of ρ,
3. (v, σ, γ) is an information information-processing representation of ρ that obeys the MCC.

A.1 Equivalence with Divisive Normalization

Proving the equivalence of the divisive normalization model and Equation (3) follows the lines of well-known arguments (Luce and Suppes, 1965; McFadden, 1978). To begin, suppose ρ has a divisive normalization representation (v, σ, γ), so that

$$\rho\left(x, A\right) = \Pr\left(x = \arg\max \gamma \frac{v\left(y\right)}{\sigma + v\left(A\right)} + \varepsilon_y\right),$$

where ε_y's are i.i.d. and Gumbel with location 0 and scale 1. Let $g(t) = \exp\left(-t - \exp\left(-t\right)\right)$ and $G(t) = \exp\left(-\exp\left(-t\right)\right)$ be the pdf and cdf of a Gumbel $(0, 1)$ random variable. We then have

$$\rho\left(x, A\right) = \int_{-\infty}^{+\infty} \left\{ \prod_{y \in A \setminus \{x\}} G\left(\gamma \frac{v\left(x\right) - v\left(y\right)}{\sigma + v\left(A\right)} + t\right) \right\} g\left(t\right) dt =$$
$$\int_{-\infty}^{+\infty} \left\{ \prod_{y \in A \setminus \{x\}} \exp\left(-\exp\left(-\gamma \frac{v\left(x\right) - v\left(y\right)}{\sigma + v\left(A\right)} - t\right)\right) \right\} \exp\left(-t - \exp\left(-t\right)\right) dt,$$

which we can rearrange to give

$$\rho\left(x,A\right)=\int_{-\infty}^{+\infty}\left\{\exp\left(-\exp\left(-t\right)\left(1+\sum_{y\in A\setminus\{x\}}\exp\left(-\gamma\frac{v\left(x\right)-v\left(y\right)}{\sigma+v\left(A\right)}\right)\right)\right)\right\}\exp\left(-t\right)dt,$$

which can be integrated to obtain

$$\rho\left(x,A\right)=\frac{1}{\left(1+\sum_{y\in A\setminus\{x\}}\exp\left(-\gamma\frac{v(x)-v(y)}{\sigma+v(A)}\right)\right)}\times$$
$$\exp\left(-\exp\left(-t\right)\left(1+\sum_{y\in A\setminus\{x\}}\exp\left(-\gamma\frac{v\left(x\right)-v\left(y\right)}{\sigma+v\left(A\right)}\right)\right)\right)\Bigg|_{t=-\infty}^{t=+\infty}.$$

Evaluating at the limits yields

$$\rho\left(x,A\right)=\frac{1}{1+\sum_{y\in A\setminus\{x\}}\exp\left(-\gamma\frac{v(x)-v(y)}{\sigma+v(A)}\right)}\left(1-0\right),$$

which can be rearranged to give

$$\rho\left(x,A\right)=\frac{\exp\left(\gamma\frac{v(x)}{\sigma+v(A)}\right)}{\sum_{y\in A}\exp\left(\gamma\frac{v(y)}{\sigma+v(A)}\right)},$$

as desired. This argument can be run backwards to prove the reverse implication.

A.2 Equivalence with Information Processing Model

Fix $v:X\to\mathbb{R}_{++}$, $\gamma>0$, and $\sigma>0$. Let $\{C_A\}_{A\in\mathcal{A}}$ be any family of strictly increasing, continuously differential, convex functions that map $\mathbb{R}\to\mathbb{R}$ and obey the MCC using (v,σ,γ). Note that such a family always exists. For example,

$$C_A\left(c\right)=\frac{\sigma+v\left(A\right)}{\gamma}c.$$

Recall, we defined ρ to have information-processing representation $\left(v,\{C_A\}_{A\in\mathcal{A}}\right)$ if, for each $A\in\mathcal{A}$, $\rho\left(\cdot,A\right)$ is a solution to the following maximization problem:

$$\max_{p\in\Delta A}\sum_{x\in A}p\left(x\right)v\left(x\right)-C_A\left(\Delta H\left(p\right)\right). \tag{5}$$

To prove the equivalence, it suffices to show that, for each $A\in\mathcal{A}$, the *unique* solution to this maximization problem is given by the measure defined by Equation (3) using (v, σ, γ). Fix $A \in \mathcal{A}$,and define $p^* \in \Delta A$ to be that measure. That is for each $x \in A$:

$$p^*(x) = \frac{\exp\left(\gamma \frac{v(x)}{\sigma + v(A)}\right)}{\sum_{y \in A} \exp\left(\gamma \frac{v(y)}{\sigma + v(A)}\right)}.$$

Next, note that any function $p : A \to [0, 1]$ can be viewed as a point in $\mathbb{R}^{|A|}$. Under this interpretation, ΔA forms a compact subset of $\mathbb{R}^{|A|}$ defined by affine constraints. Also note that, by standard properties of entropy, $H(\cdot)$ is strictly concave, which means $\Delta H(\cdot)$ is strictly convex. Using that $C_A(\cdot)$ is convex and strictly increasing, it follows that $C_A(\Delta H(\cdot))$ is strictly convex, and hence $-C_A(\Delta H(\cdot))$ is strictly concave. Therefore, the objective function of the maximization problem in (5) is strictly concave since the only other term is linear. Affine constraints and strictly concave objective function mean that the Karush-Kuhn-Tucker conditions are both necessary and sufficient for a feasible point to be a solution. Those conditions say

$$v(x) - C'_A(\Delta H(p))(\ln(p(x)) + 1) + \lambda + \mu_x = 0, \tag{6}$$

for some λ and μ_x, with the complimentary slackness condition that $p(x) \neq 0 \Rightarrow \mu_x = 0$. We now show that p^* satisfies those conditions. Since $p^*(x) > 0$ for all x, we set $\mu_x = 0$. Define

$$\lambda := -\frac{\sigma + v(A)}{\gamma}\left(\ln\left(\sum_{y \in A} \exp\left(\gamma \frac{v(y)}{\sigma + v(A)}\right)\right) - 1\right).$$

We need to show that

$$v(x) - C'_A(\Delta H(p^*))(\ln(p^*(x)) + 1) + \lambda = 0.$$

By definition, $\delta(A, v, \sigma, \gamma) = \Delta H(p^*)$. Hence we can apply the MCC to transform the above equation into

$$v(x) - \frac{\sigma + v(A)}{\gamma}(\ln(p^*(x)) + 1) + \lambda = 0.$$

Using our definition of p^* this becomes

$$v(x) - \frac{\sigma + v(A)}{\gamma}\left(\gamma \frac{v(x)}{\sigma + v(A)} - \ln\left(\sum_{y \in A} \exp\left(\gamma \frac{v(y)}{\sigma + v(A)}\right)\right) + 1\right) + \lambda = 0,$$

which simplifies to

$$\frac{\sigma + v(A)}{\gamma}\left(\ln\left(\sum_{y\in A}\exp\left(\gamma\frac{v(y)}{\sigma + v(A)}\right)\right) - 1\right) + \lambda = 0,$$

which holds by definition of λ.

We have now proved that p^* is a solution to the maximization problem. Next suppose q^* also solves the maximization problem. Since ΔA is closed under convex combinations, we can define a feasible $p \in \mathcal{P}$ by

$$p(x) := \frac{1}{2}q^*(x) + \frac{1}{2}p^*(x),$$

for each $x \in A$. Since the objective function is strictly concave, if $p^* \neq q^*$, then p would strictly improve on the optimal payoff, which is not possible. Hence p^* must be the unique maximizer, as desired.

A.3 Equivalence with Axioms

Equation (3) Implies the Axioms

Suppose ρ obeys Equation (3) using (v, σ, γ). We will show ρ obeys all six axioms. Axiom 1 follows immediately from the fact that, under Equation (3), $\rho(x, A) \geq \rho(y, A) \iff v(x) \geq v(y)$ since $\gamma > 0$ and $\sigma + v(A) > 0$ for all $A \in \mathcal{A}$. To show the necessity of the rest of the axioms, we use the following lemma.

Lemma 1. *If ρ obeys Equation (3) using (v, σ, γ), then*

$$R_{xy}(A) = \frac{\sigma + v(A)}{\sigma + v(X)}.$$

whenever (x, y) is distinguishable in A.

Proof. By definition,

$$R_{xy}(A) = \left(\ln\frac{\rho(x, A)}{\rho(y, A)}\right)^{-1}\left(\ln\frac{\rho(x, X)}{\rho(y, X)}\right).$$

Applying Equation (3) to the right-hand side gives

$$R_{xy}(A) = \left(\gamma\frac{v(x) - v(y)}{\sigma + v(A)}\right)^{-1}\left(\gamma\frac{v(x) - v(y)}{\sigma + v(X)}\right),$$

and the desired conclusion follows. □

Axiom 2 follows immediately from Lemma 1. Also, Lemma 1 allows us to rewrite the equation in Axiom 3 as

$$\frac{\sigma+v\left(A\right)}{\sigma+v\left(X\right)}-\frac{\sigma+v\left(A\backslash\left\{z\right\}\right)}{\sigma+v\left(X\right)}=\frac{\sigma+v\left(B\right)}{\sigma+v\left(X\right)}-\frac{\sigma+v\left(B\backslash\left\{z\right\}\right)}{\sigma+v\left(X\right)},$$

which holds since both sides are equal to $v\left(z\right)/\left(\sigma+v\left(X\right)\right)$.

Using Lemma 1, the equation in Axiom 4 is equivalent to

$$\frac{v\left(x\right)-v\left(y\right)}{\sigma+v\left(X\right)}\left(\frac{v\left(x'\right)-v\left(y'\right)}{\sigma+v\left(X\right)}\right)^{-1}=\ln\frac{\rho\left(x,X\right)}{\rho\left(y,A\right)}\left(\ln\frac{\rho\left(x',X\right)}{\rho\left(y',X\right)}\right)^{-1},$$

which can be verified by applying Equation (3) to the right-hand side.

Now suppose that $A\in\mathcal{A}$ contains a distinguishable pair (x,y) and $z,z'\in A\backslash\left\{x,y\right\}$ such that $\rho\left(z,X\right)>\rho\left(z',X\right)$. By Equation (3), $v\left(z\right)>v\left(z'\right)$. Using this plus $v,\sigma>0$, it follows that

$$\frac{\sigma+v\left(A\right)}{\sigma+v\left(X\right)}>\frac{\sigma+v\left(A\backslash\left\{z'\right\}\right)}{\sigma+v\left(X\right)}>\frac{\sigma+v\left(A\backslash\left\{z\right\}\right)}{\sigma+v\left(X\right)},$$

which, via Lemma 1, proves Axiom 5.

Finally, let the sets A,B contain a distinguishable pair. Then Axiom 6 is equivalent to

$$\frac{\sigma+v\left(A\right)}{\sigma+v\left(X\right)}+\frac{\sigma+v\left(B\right)}{\sigma+v\left(X\right)}>\frac{\sigma+v\left(A\cup B\right)}{\sigma+v\left(X\right)}.$$

Since the denominators are all positive, this is equivalent to

$$\sigma+v\left(A\cap B\right)>0$$

which holds because $\sigma,v>0$.

The Axioms Imply Equation (3)

Now assume Axioms 1-6 hold. We will find (v,σ,γ) such that Equation (3) holds. By Axiom 1, if (x,y) is distinguishable in A, then (x,y) is distinguishable in all sets that contain this pair. So, we will simply say (x,y) is distinguishable to indicate that $\rho\left(x,A\right)\neq\rho\left(y,A\right)$ whenever $x,y\in A$. By Axiom 2, for any $A\in\mathcal{A}$, we can set $R\left(A\right)=R_{xy}\left(A\right)$ for all distinguishable (x,y) in A. If A does not contain any distinguishable pairs, set $R\left(A\right)=1$. Also, set $R\left(\emptyset\right)=0$.

Lemma 2. *There exists a distinguishable pair* $(x^\star, y^\star)$ *such that* $X\backslash\{x^\star, y^\star\}$ *contains a distinguishable pair.*

Proof. By assumption, $\rho(\cdot, X)$ contains at least three distinct choice probabilities. Let $\{x, y, z\}$ denote the three items generating the distinct probabilities. Since $|X| \geq 4$, there exists $w \in X\backslash\{x, y, z\}$. It must be that either $\rho(w, X) \neq \rho(x, X)$ or $\rho(w, X) \neq \rho(y, X)$. In the first case, set $(x^\star, y^\star) = (y, z)$, and. in the second case, set $(x^\star, y^\star) = (x, z)$. Either way we get the desired result. □

From here on, let $(x^\star, y^\star)$ be a distinguishable pair such that $X\backslash\{x^\star, y^\star\}$ contains a distinguishable pair. For any $x \in X$, note that either $X\backslash\{x\} \supseteq X\backslash\{x^\star, y^\star\}$ or $\{x^\star, y^\star\} \subseteq X\backslash\{x\}$. Hence, $X\backslash\{x\}$ always contains a distinguishable pair.

Now, for any $x \in X$, define

$$v(x) := R(X) - R(X\backslash\{x\}).$$

Also define

$$\sigma := R(\{x^\star, y^\star\}) - v(x^\star) - v(y^\star).$$

Lemma 3. *If* $A \in \mathcal{A}$ *contains distinguishable pair* (x, y)*, then*

$$R(A) = R(\{x, y\}) + v(A\backslash\{x, y\}).$$

Proof. Let $z \in A\backslash\{x, y\}$. By Axiom 3,

$$R(X) - R(X\backslash\{z\}) = R(A) - R(A\backslash\{z\}).$$

Combining the above with the definition of v yields

$$R(A) = R(A\backslash\{z\}) + v(z).$$

For any $z' \in A\backslash\{z, y, x\}$, the same logic yields

$$R(A) = R(A\backslash\{z, z'\}) + v(z') + v(z).$$

We repeatedly apply these steps to get the desired result. □

Lemma 4. *For any* $A \in \mathcal{A}$ *that contains a distinguishable pair*

$$R(A) = \sigma + v(A).$$

Proof. Using Lemma 3, it suffices to show that, for any distinguishable pair (x, y)

$$R(\{x, y\}) = \sigma + v(x) + v(y).$$

First, we can apply Lemma 3 and the definition of σ to get:

$$R(\{x^\star, y^\star, x\}) = R(\{x^\star, y^\star\}) + v(x) = \sigma + v(x^\star) + v(y^\star) + v(x).$$

Since $\rho(x^\star, X) \neq \rho(y^\star, X)$, it must be that either $(x, x^\star)$ is distinguishable or $(x, y^\star)$ is distinguishable. We will treat the first case, since the proof for the other case is similar. By Lemma 3, we have

$$R(\{x^\star, y^\star, x\}) = R(\{x^\star, x\}) + v(y^\star).$$

Combining the previous two equations gives

$$R(\{x^\star, x\}) = \sigma + v(x^\star) + v(x).$$

Since $(x^\star, x)$ and (x, y) are distinguishable pairs, we can use Lemma 3 to get

$$R(\{x^\star, x, y\}) = R(\{x^\star, x\}) + v(y) = R(\{x, y\}) + v(x^\star).$$

Combining this with the previous display equation yields

$$R(\{x, y\}) = \sigma + v(\{x\}) + v(\{y\}),$$

as desired. □

Since $\rho(\cdot, X)$ contains at least three distinct choice probabilities, there exists $z^\star \in X$ such that all pairs in $\{x^\star, y^\star, z^\star\}$ are distinguishable. Define

$$\gamma := \frac{\ln \frac{\rho(x^\star, X)}{\rho(y^\star, X)}}{R(\{x^\star, z^\star\}) - R(\{y^\star, z^\star\})}.$$

That γ is well-defined (i.e., that the denominator does not equal 0) is implicit in Axiom 4, where we apply the axiom to the case that $(x, y, z) = (x', y', z') = (x^\star, y^\star, z^\star)$. That $(x^\star, y^\star)$ is distinguishable ensures that $\gamma \neq 0$.

Now fix any $A \in \mathcal{A}$, and we claim that

$$\frac{\rho(x, A)}{\rho(y, A)} = \exp\left(\gamma \frac{v(x) - v(y)}{\sigma + v(A)}\right), \tag{7}$$

for all $x, y \in A$. Chose any pair $x, y \in A$. Since $\rho(\cdot, X)$ contains at least three distinct choice probabilities, there must exist $z \in X$ such that (x, z) and (y, z) form distinguishable pairs. Applying Axiom 4, we then get

$$\frac{R(\{x, z\}) - R(\{y, z\})}{R(\{x^\star, z^\star\}) - R(\{y^\star, z^\star\})} = \ln \frac{\rho(x, X)}{\rho(y, X)} \left(\ln \frac{\rho(x^\star, X)}{\rho(y^\star, X)} \right)^{-1}.$$

Applying the definition of γ and Lemma 4, we have

$$\gamma(v(x) - v(y)) = \ln \frac{\rho(x, X)}{\rho(y, X)}.$$

If (x, y) is not distinguishable, then the above equation implies $v(x) = v(y)$. Therefore, Equation (7) holds since both sides are equal to 1. If (x, y) is distinguishable, then we can use our definition of $R(A)$ to get

$$\gamma \frac{v(x) - v(y)}{R(A)} = \ln \frac{\rho(x, A)}{\rho(y, A)}.$$

Since we are assuming (x, y) is distinguishable, we can apply Lemma 4 and take the exponent of both sides to get

$$\frac{\rho(x, A)}{\rho(y, A)} = \exp\left(\gamma \frac{v(x) - v(y)}{\sigma + v(A)} \right).$$

Hence we have proved Equation (7). Using the fact that choice probabilities must sum to 1, we can get that for each $A \in \mathcal{A}$ and $x \in A$:

$$\rho(x, A) = \frac{\exp\left(\gamma \frac{v(x)}{\sigma + v(A)}\right)}{\sum_{y \in A} \exp\left(\gamma \frac{v(y)}{\sigma + v(A)}\right)}.$$

All that remains is to show that v, σ, γ are strictly positive. Recall that $X \backslash \{x\}$ contains a distinguishable pair for all $x \in A$. Hence, by Axiom 5,

$$v(x) = R(X) - R(X \backslash \{x\}) > 0,$$

for all $x \in X$. Next, note that, using Lemma 4, we get

$$R(\{x^\star, z^\star\}) - R(\{y^\star, z^\star\}) = R(X \backslash \{y^\star\}) - R(X \backslash \{x^\star\}).$$

Therefore, by Axiom 5, $R(\{x^\star, z^\star\}) - R(\{y^\star, z^\star\})$ and $\ln(\rho(x^\star, X) / \rho(y^\star, X))$ have the same sign. Since γ is defined as the ratio of these two terms, and using thefact that we already showed γ is well-defined and non-zero, we get that $\gamma > 0$. Finally, recall that, by construction, $\{x^\star, y^\star\}$ and $X\backslash \{x^\star, y^\star\}$ both contain a distinguishable pair. Hence we can apply Lemma 4 to get

$$R\left(\{x^\star, y^\star\}\right) + R\left(X\backslash \{x^\star, y^\star\}\right) - R\left(X\right) = \sigma.$$

By Axiom 6, the left-hand side of that equation is strictly positive, and it follows that $\sigma > 0$.

References

Agranov, M. and P. Ortoleva (2017). Stochastic Choice and Preferences for Randomization. *Journal of Political Economy 125*(January).

Anderson, S., A. de Palma, and J.-F. Thisse (1992). *Discrete Choice Theory of Product Differentiation*. Cambridge: MIT Press.

Attneave, F. (1954). Some Informational Aspects of Visual Perception. *Psychological review 61*(3), 183–193.

Barlow, H. (1961). Possible Principles Underlying the Transformation of Sensory Messages. In W. A. Rosenblith (Ed.), *Sensory Communication*, pp. 217–234. Cambridge: M.I.T. Press.

Block, H. D. and J. Marschak (1960). Random Orderings and Stochastic Theories of Responses. In I. Olkin (Ed.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pp. 97–132. Stanford: Stanford University Press.

Caplin, A. and M. Dean (2015). Revealed Preference, Rational Inattention, and Costly Information Acquisition. *American Economic Review 105*(7), 2183–2203.

Carandini, M. and D. Heeger (2012). Normalization as a Canonical Neural Computation. *Nature Reviews Neuroscience* (November), 1–12.

David, H. and H. Nagaraja (2003). *Order Statistics* (3rd ed.). New York: Wiley.

Echenique, F. and K. Saito (2015). General Luce Model. (452), 1–27.

Echenique, F., K. Saito, and G. Tserenjigmid (2014). The Perception-Adjusted Luce Model. (1959), 1–38.

Fehr, E. and A. Rangel (2011). Neuroeconomic Foundations of Economic Choice-Recent Advances. *Journal of Economic Perspectives 25*(4), 3–30.

Fudenberg, D., R. Iijima, and T. Strzalecki (2015). Stochastic Choice and Revealed Perturbed Utility. *Econometrica 83*(6), 2371–2409.

Glimcher, P. (2011). *Foundations of Neuroeconomic Analysis*. OUP.

Glimcher, P. W. and A. A. Tymula (2018). Expected Subjective Value Theory (ESVT): A Representation of Decision under Risk and Certainty.

Gul, F., P. Natenzon, and W. Pesendorfer (2014). Random Choice as Behavioral Optimization. *Econometrica 82*(5), 1873–1912.

Huber, J., J. W. Payne, and C. Puto (1982). Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis. *Journal of Consumer Research 9*(1), 90–98.

Itthipuripat, S., K. Cha, N. Rangsipat, and J. T. Serences (2015). Value-Based Attentional Capture Influences Context-Dependent Decision-Making. *Journal of Neurophysiology 114*(1), 560–569.

Khaw, M. W., P. W. Glimcher, and K. Louie (2017). Normalized Value Coding Explains Dynamic Adaptation in the Human Valuation Process. *Proceedings of the National Academy of Sciences 114*(48), 201715293.

Knutson, B., C. M. Adams, G. W. Fong, and D. Hommer (2001). Anticipation of Increasing Monetary Reward Selectively Recruits Nucleus Accumbens. *J Neurosci 21*(16), RC159.

Landauer, R. (1961). Irreversibility and Heat Generation in the Computing Process. *IBM Journal of Research and Development 5*(July), 183–191.

Landry, P. and R. Webb (2017). Pairwise Normalization: A Neuroeconomic Theory of Multi-Attribute Choice. *Ssrn.*

Loewenstein, G., S. Rick, and J. Cohen (2008). Neuroeconomics. *Annual review of psychology 59*, 647–672.

Louie, K., P. W. Glimcher, and R. Webb (2015). Adaptive Neural Coding: From Biological to Behavioral Decision-Making. *Current Opinion in Behavioral Sciences 5*, 91–99.

Louie, K., L. E. Grattan, and P. W. Glimcher (2011). Reward Value-Based Gain Control: Divisive Normalization in Parietal Cortex. *Journal of Neuroscience 31*(29), 10627–10639.

Louie, K., M. W. Khaw, and P. W. Glimcher (2013). Normalization is a General Neural Mechanism for Context-Dependent Decision Making. *Proceedings of the National Academy of Sciences of the United States of America 110*(15), 6139–44.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.

Luce, R. D. and P. Suppes (1965). Preference, Utility, and Subjective Probability. In R. D. Luce, R. Bush, and E. Galanter (Eds.), *Handbook of Mathematical Psychology*, pp. 249–410. New York: Wiley.

Machina, M. J. (1989). Dynamic Consistency and Non-Expected Utility Models of Choice Under Uncertainty. *Journal of Economic Literature 27*(4), 1622–1668.

Mandl, F. (1988). *Statistical Physics* (2nd ed.). John Wiley & Sons.

Marley, A. A. J., T. N. Flynn, and J. J. Louviere (2008). Probabilistic Models of Set-Dependent and Attribute-Level Best-Worst Choice. *Journal of Mathematical Psychology 52*(5), 281–296.

Matějka, F. and A. McKay (2014). Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model. *American Economic Review 105*(1), 1–55.

Mattsson, L.-G. and J. W. Weibull (2002). Probabilistic Choice and Procedurally Bounded Rationality. *Games and Economic Behavior 41*(1), 61–78.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior.

McFadden, D. (1978). Modelling the Choice of Residential Location. In S. Karlqvist, L. Lundqvist, F. Snickars, and J. Weibull (Eds.), *Spatial Interaction Theory and Planning Models*, Volume 673, pp. 75–96. Amsterdam: North-Holland.

Platt, M. L. and P. W. Glimcher (1999). Neural Correlates of Decision Variables in Parietal Cortex. *Nature 400*(6741), 233–238.

Ravid, D. (2015). Focus, Then Compare. *Working Paper*, 1–40.

Rieskamp, J., J. R. Busemeyer, and B. A. Mellers (2006). Extending the Bounds of Rationality: Evidence and Theories of Preferential Choice. *Journal of Economic Literature 44*(3), 631–661.

Schwartz, O. and E. P. Simoncelli (2001). Natural Signal Statistics and Sensory Gain Control. *Nature neuroscience 4*(8), 819–825.

Shannon, C. E. (1948). A Mathematial Theory of Communication. *Bell System Technical Journal 27*(3), 379–423.

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *Quarterly Journal of Economics 69*(1), 99–118.

Simonson, I. (1989). Choice Based on Reasons: The Case of Attraction and Compromise Effects. *Journal of Consumer Research 16*(2), 158–174.

Sims, C. (1998). Stickiness. *Carnegie-Rochester Conference Series on Public Policy 49*, 317–356.

Sims, C. A. (2003). Implications of Rational Inattention. *Journal of Monetary Economics 50*(3), 665–690.

Swait, J. and A. A. J. Marley (2013). Probabilistic Choice (Models) as a Result of Balancing Multiple Goals. *Journal of Mathematical Psychology 57*(1-2), 1–14.

Thurstone, L. (1927). A Law of Comparative Judgment. *Psychological Review 34*(4), 273–286.

Tserenjigmid, G. (2016). The Order-Dependent Luce Model. pp. 1–20.

Webb, R., P. Glimcher, and I. Levy (2013). Neural Random Utility and Measured Value. *Available at SSRN . . .* , 1–36.

Webb, R., P. W. Glimcher, and K. Louie (2014). Rationalizing Context-Dependent Preferences: Divisive Normalization and Neurobiological Constraints on Choice. *Working Paper*, 1–56.