WEIGHTED LINEAR DISCRETE CHOICE

CHRISTOPHER P. CHAMBERS[†], YUSUFCAN MASATLIOGLU[§], AND COLLIN RAYMOND[*]

ABSTRACT. We introduce a new model of stochastic choice. The model modifies Luce by adding one additional parameter which reflects salience or other economic frictions involved in making a choice. The model is consistent with many classical approaches including random utility as well as preference maximization as in [Machina, 1985]. We characterize our model behaviorally and investigate its comparative statics properties. Finally, we demonstrate the applicability of our model in equilibrium settings where firms can choose price, quality, and advertising. The model generates intuitive closed form solutions in these situations, produces results consistent with data which other models cannot accommodate, and allows for preference parameter identification.

1. INTRODUCTION

The primary goal of this paper is to introduce a *simple and tractable* yet *flexible* two-parameter model of stochastic choice, which we call the weighted linear (WL) model of discrete choice. The two parameters have a simple interpretation: one can be seen as reflecting the underlying quality or utility of an item, while the second reflects the ease of choosing an item, thinking about it, or how salient it is (i.e. the frictions involved in choice). Although it is only a one-parameter extension of the widely used Luce model, the WL approach can accommodate behavior that the most widely used approaches to discrete choice fail to allow for, as [Berry and Pakes, 2007] note, "though [typical discrete choice] models can do quite a good job in approximating some aspects of demand, they also have some counter-intuitive implications..." (see also

Date: February-May 2021. We thank Joe Mazur, Yesim Orhun, Andre Veiga and Chenyu Yang for helpful comments. We are especially grateful to Christopher Turansick for excellent research assistance.
[†] Department of Economics, Georgetown University, ICC 580 37th and O Streets NW, Washington DC 20057. E-mail: Christopher.Chambers@georgetown.edu.
[§] University of Maryland, 3147E Tydings Hall, 7343 Preinkert Dr., College Park, MD 20742. E-mail: yusufcan@umd.edu.
[*] Krannert School of Management, Purdue University, 403 W State St, West Lafayette, IN 47907, collinbraymond@gmail.com.

[Benkard et al., 2001]).[1] For example, the WL model can generate flexible cross-price substitution effects, as well as intuitive patterns of choice probabilities from large choice sets. Despite this added explanatory power, the WL model also re mains easy to use: under standard assumptions the parameters can be estimated from market shares via a set of simple linear equations. The additional parameter of the WL model allows us to analyze the behavior of firms who can compete via mechanisms which manipulate the salience of outcomes, such as advertising, and we can derive closed form solutions for oligopolistic competition.

In the WL model the probability of choosing an item is the sum of a base component plus a comparative component. The base component simply reflects the relative ease of choosing that particular item compared to other items in the choice set and reflects Luce's choice rule [Luce, 1959]. The comparative component, in line with Fechnernian stochastic choice [Fechner, 1860], reflects the difference in utility between the item and other items in the choice set, scaled by the ease of choosing the item. As we will demonstrate, our model represents a significant improvement on the explanatory power of the Luce model.

Despite the simplicity of the model, the choice probabilities it generates are consistent with several different potential microfoundations. For example, our model can be derived from a very simple optimization procedure by decision-makers: individuals try to maximize their expected utility, subject to a quadratic cost of increasing choice probabilities. In addition, this model has the advantage of a simple behavioral characterization. We show that identification and testing of the model are straightforward in terms of observable choice probabilities, and the behavioral foundations correspond in an intuitive way to natural generalizations of Luce's model, demonstrating how it can flexibly accommodate patterns of choice that other models exclude.

Section 2 introduces the WL model. Outcomes are ranked by two orderings: a utility ordering u, which captures "how good" an item is, and a weighting ordering, m, which captures the ease of being able to choose an item (representing psychological factors, such as salience, attention or noticeability, or physical costs of choice, such as

[1] As is well known, the Luce model is behaviorally equivalent to a logit model of choice.

search frictions). The probability of choosing x from S is equal to

$$\rho(x|S) = \underbrace{\frac{m(x)}{\sum_{y \in S} m(y)}}_{\text{Base Probability}} + \underbrace{m(x)[u(x) - \bar{u}_m(S)]}_{\text{Comparative Probability}}$$

where $\bar{u}_m(S)$ is the weighted average utility of outcomes in S with respect to m. The probability of choosing x is the sum of two components. The first, the base probability, reflects how easy it is to think of x compared to the rest of the choice set. The second component, the comparative probability, reflects how good x is compared to a weighted average of other items in the choice set, where the weights depend on the ease of being able to choose the item scaled by the ease of choosing x.

Our approach is the stochastic choice equivalent of a linear demand system, widely used in economics, where demand for a product reflects: (i) a base component independent of the utility of the item and (ii) a component that reflects the difference in utility between the item and the average item scaled by some number that represents the friction in the market (as the number gets bigger, the best items eventually attract the entire market). The approach of linear demand systems has been used extensively in applied settings. Early references include [Shubik and Levitan, 1980], [Spence, 1976, Dixit and Stiglitz, 1977], and [Singh and Vives, 1984].

After introducing our model, we discuss several special cases. Our model naturally nests the logit model, as well as the "simple" version of linear demand systems, where all firms are symmetric in their ease of choice.

In Section 3 we turn to providing formal microfoundations for WL stochastic choice. We show that despite its simple form, the model can be microfounded in several ways. We begin by providing foundations for our model as the solution to a simple utility maximization problem subject to a cost of choosing items too frequently. In particular, the decision maker, as in other recent models (see [Machina, 1985, Clark, 1990, Fudenberg et al., 2015, Feldman and Rehbeck, 2018, Cerreia-Vioglio et al., 2019]) chooses the probability with which each option is realized. Her goal is to maximize the expected utility of a lottery, less a cost of the chosen lottery. Our model is characterized by a cost which is quadratic in individual probabilities, but the scale of the cost can depend on the alternative.

We present an alternative formulation of our maximization problem, where for each outcome the decision maker has an optimal probability (a bliss point) assigned to that outcome. The decision-maker pays a quadratic cost for deviating from this optimal probability where the marginal cost of deviations depends on the outcome in a systematic way.

Section 4 demonstrates that our model, when all choice probabilities are strictly positive, has very simple behavioral foundations. It demonstrates that three simple axioms characterize the WL model. Structurally, the behavioral content of the model is characterized by a novel type of acyclicity condition. We show that identification can be achieved by observing choices on relatively few choice sets.[2]

Section 5 elaborates further on WL stochastic choice, comparing it to other recent models. A main insight here is that in the context of binary choice, we can construct a "utility measure" that uses both u and c, where c is the cost function described above. Thus, we can link our model to to the ideas of Fechnerian choice, though our model applies both to binary as well as non-binary choice sets. Here, we also show we are a subset of the moderate stochastic choice models recently characterized by [He and Natenzon, 2019]. In the context of binary choice, we can also clearly see the link between the WL approach developed in this paper and the classic [Hotelling, 1929] model. We discuss other properties of our model (such as the comonotonicity of probabilities with respect to choice sets), as well as the relation to other standard models of stochastic choice. We also establish that our model is a subset of the random utility model (hereafter RUM), characterized by [Falmagne, 1978]. Moreover, it shares some important features with the additive perturbed utility model of [Fudenberg et al., 2015] (and so more generally, with models of deliberate randomization). We show that, although distinct from the additive perturbed utility approach, our model has a non-trivial intersection with it, which nests both the logit model, the simplest formulation of linear demand, as well as the simple linear demand approach described above.

Section 6 illustrates the applicability of our model in five ways. We first discuss the comparative statics embodied in our model. Second, we think about how the WL model can accommodate various forms of context dependence in choice, demonstrating that it can allow for various choice anomalies.

[2]We defer the characterization of our model when choice probabilities can be zero until Appendix B.

Third, we show how the WL model easily extends to simple setting that allows for strategic interactions among firms. We demonstrate that the WL model naturally captures consumers' response to competition among firms in multiple dimensions: price, advertising and quality. Moreover, despite this richness, the model is tractable enough to lead to simple closed form solutions for firms' strategies and payoffs. These solutions capture useful intuitions about firm behavior, including the relationship between advertising, markups, and the number of firms in the market, as well as the size of the market and the number of firms (i.e. market fragmentation).

We then turn to discussing how our approach also provides for additional explanatory power above and beyond that provided by the standard discrete choice approaches often used in the literature. For example, unlike logit, but like many of its generalizations, our model can allow for much richer patterns of cross price substitution. Even more surprising, unlike almost all widely used discrete choice models, our model can allow, even in the limit as the number of firms goes to infinity, for some firms to hold dominant, non-negligible market shares, while other firms may be driven out of the market entirely.

Last, we discuss how our model can be easily identified in choice settings with a fixed choice set, but with observable attributes of outcomes. Although standard discrete choice results do not apply to our model, we can show that the parameters can be identified via simple linear equations.

While many other discrete choice models have been developed that both relax some of the restrictions of the logit choice approach, and are identifiable, we believe that our model also has the benefit of being extremely tractable and leading to closed form solutions in simple applied models. Thus, it allows for a more careful combining of theory and empirics.

Finally, Section 7 concludes. The Appendices include proofs and discuss other relevant ideas, such as behavioral properties of the analogue of our model which allows for zero probabilities.

2. Model

We initially describe our model in an abstract environment, and then in later sections apply it to particular settings. Let X be a finite set of alternatives. Let $\mathcal{X}$ be the set of all probability measures in X. That is, $\rho(\cdot|X) \in \mathcal{X}$ implies $\rho(x|X) \geq 0$ and $\sum_{x \in X} \rho(x|X) = 1$. Let $\mathcal{D}$ denote the set of non-empty subsets of X. For every $\emptyset \neq S \subseteq X$, denote by $\mathcal{S}$ the elements in $\mathcal{X}$ which naturally induce probability measures on S, i.e., $\rho(.|S) \in \mathcal{S}$ means $\rho(x|S) \in \mathcal{X}$ and $\rho(x|S) = 0$ whenever $x \notin S$. We also denote the sum of choice probabilities in $T \subset S$ as $\rho(T|S)$. Similarly, for any real function f on X, $f(S)$ denotes the sum of $f(x)$ for all $x \in S$. We will denote binary choices as $\rho(x, y)$ instead of $\rho(x|\{x, y\})$.

We now introduce our two-parameter model of stochastic choice. In this model, each alternative is represented by two values: u and m.

Definition 1. A stochastic choice ρ is consistent with a *weighted linear* (WL) model on $\mathcal{D}$ if there exist functions $u : X \to \mathbb{R}$ and $m : X \to \mathbb{R}_{++}$ such that for all $S \in \mathcal{D}$,

$$\rho(x|S) = \frac{m(x)}{m(S)} + m(x)[u(x) - \bar{u}_m(S)]$$

where $\bar{u}_m(S) \equiv \frac{\sum_{x \in S} u(x)m(x)}{m(S)}$ is the weighted utility with respect to m in S. We also say (u, m) *represents* ρ, or (u, m) is a *WL representation* of ρ.[3]

We refer to elements of X as *items*. The model suggests that each item has a base probability of attracting consumers $m(x)$. This can be thought of as the "presence" of the product in the marketplace. Of course, not all products might be available at any given time. So, conditional on a particular choice set S, we study the conditional probability of product x getting noticed, or x's presence in the market (given the set of available products).

The base probabilities are not the only determinants of choice probabilities in our model. The "utility" of each product also plays an important role. Products gain or lose consumers in proportion to the difference between their utility and the weighted average utility in the market. The deviation from $\bar{u}_m(S)$ captures the market

[3]Note that not all pairs (u, m) constitute a stochastic choice, since in principle the implied choice could be negative. A joint condition on u and m, $m(X)(\bar{u}_m(X) - \min_{z \in X} u(z)) < 1$, eliminates such cases. In general, (u, m) might induce zero probability choices. We will investigate this in detail Section B.

influence on probabilistic demand. In this formulation, the choice probability of a product increases linearly in its own utility. If a product offers a higher utility than the market average, it enjoys additional choice probability. Otherwise, being less than the market average reduces choice probabilities. Similarly, the probability of choice is linearly decreasing in the utility of other products. Our formulation is a specific case of a larger class of demand functions where choice probabilities (i.e. market shares) are linear in the net utility of a product, which are often used in applications. Early examples are [Shubik and Levitan, 1980], [Spence, 1976, Dixit and Stiglitz, 1977], and [Singh and Vives, 1984], while [Choné and Linnemer, 2020] provides a survey.

We can also interpret this representation via the lens of individual random choice. We can interpret m as the probability that an individual "thinks" about x. We could also think of it as the salience of x. Again, there is a two-stage process for deciding demand. The quantity, $m(x)$, is the probability that the individual thinks about x in both steps. In the first step, an individual's base demand for any given product is allocated to how much the individual thinks about x (i.e. how salient it is). In the second step, the individual forms their impression of the average quality again using the salience of products.

The model exhibits the expected comparative statics with respect to u and m. We show in Section 6.1 that the probability of choosing x from any set is increasing in both $u(x)$ and $m(x)$. In other words, each item benefits from having higher utility and higher salience.

Two Special Cases. Our formulation aims to nest two important special cases. The first one is when the utility function is constant but m's differ across alternatives. If $u(x) = u(y)$ for all x, y, the second term in the representation disappears and the choice probabilities are solely driven form m:

$$\rho(x|S) = \frac{m(x)}{m(S)}$$

Clearly, this is the classical model of [Luce, 1959]. Thus, if utility is constant, we obtain Luce choice probabilities, and, conversely, any Luce individual also conforms to our model simply by inverting the preceding algebra. It is routine to show that ρ is a WL where $u(x) = u(y)$ all $x, y \in X$ if and only if it has a Luce representation.

The other "extreme" case is where m's remain constant, $m(x) = \bar{m}$ for all x. In this case, the first term is simply $\frac{1}{|S|}$, each alternative attracts attention uniformly. Then the weighted average becomes the ordinary average, $\bar{u}_m(S) = \bar{u}(S)$. Then we have

$$\rho(x|S) = \frac{1}{|S|} + \bar{m}[u(x) - \bar{u}(S)]$$

This is equivalent to the basic linear demand system that feature prominently in many models of monopolistic competition (see below for a more extensive discussion of this). In order to see this more clearly, without loss we can $u(x) = \bar{u} - p(x)$. If we call $p(x)$ as the price of x, then we can define demands as

$$\rho(x|S) = \frac{1}{|S|} + \bar{m}[\bar{p}(S) - p(x)]$$

Although this equation abstracts away from the possibility of demand which is zero, a few points bear mentioning. The equation is intended to measure the market share for each firm x. Each firm gets a base share $\frac{1}{|S|}$, and the residual arises from the deviation of their price from the average market price. The level of this deviation is influenced by the parameter $\bar{m}$, which we suggest can be interpreted as a cardinal measure of market friction. This means that, by lowering price, each firm can steal some of the market share, but that amount depends on the friction in the market. As $\bar{m}$ marginally increases, all firms with prices lower than the average receive a positive gain in market share. On the other hand, if $\bar{m}$ gets very small, consumers tend to purchase equally across all firms, ignoring price.

Uniqueness. Our model enjoys very strong uniqueness properties. If (u, m) represents ρ, then $(au + b, \frac{1}{a}m)$ also represents ρ for $a > 0$ and b. We also show that if (u, m) and (u', m') represent the same choice data, they are equivalent up to the same class of transformations. Utility is unique up to an affine transformation, whereas salient function is only unique up to a scale transformation. The scale parameter of utility is the inverse of the scale parameter of salience. The uniqueness result suggests that our interpretation of the model's parameters is warranted, as we tend to think of utility as being defined only up to affine transformation.

Theorem 1 (Uniqueness)**.** *Let (u, m) be a WL representation of ρ. Then (u', m') is a WL representation of ρ if and only if $u' = au + b$ and $m' = \frac{1}{a}m$.*

3. Micro Foundations and Connections

We now turn to providing foundations for why the choice probabilities in the WL model might emerge. We provide two separate micro-foundations for our model, both based on underlying optimization procedures. They embed a preference for randomization but via distinct conceptual lenses.[4] The first approach demonstrates that our model can be seen as the outcome of a consumer trying to maximize utility subject to a (probability) cost. It supposes individuals want to maximize expected utility, but face a cost of assigning too much weight to any one particular product. Thus we can think of these consumers as essentially standard, subject to an item-specific cost of controlling trembles (in this way it is reminiscence of the APU model of [Fudenberg et al., 2015], a relationship we explore below).

The second approach shows that our model can be seen as the outcome of a consumer trying to optimize probabilities subject to a (probability) constraint. In particular, the individual is *not* trying to maximize expected utility. Instead, each item has an optimal probability that the individual would like to choose it with. However, the sum of the probabilities across all items must, obviously, sum to 1. Thus, the individual may have to deviate from the optimal probability for any given item. When deviating they face an item specific marginal cost.

3.1. **Deliberate Randomization through Expected Utility Maximization.** We now turn to exploring how choice probabilities emerge as the outcome of a consumer maximizing expected utility subject to a cost. A simple change of variable will be useful here: namely, let us write $c(x) \equiv \frac{1}{m(x)}$. The function c will then operate as a "cost."

[4]A distinct way of providing microfoundations for random choice is via a representative consumer approach. A representative agent chooses not probabilities, but rather market shares, which are equivalent to probabilities. They maximize their utility by allocating their spending across different goods. [Anderson et al., 1988] link the traditional Luce model of random choice to a representative agent approach (and show that the representative agent utility function depends on the entropy of the choice process). Not surprisingly, given that WL's "demand" function is linear in utilities, one can, derive these demands from a representative consumer trying to maximize a quadratic quasilinear utility function subject to a budget constraint. This accords with the results in [Choné and Linnemer, 2020], who discuss in detail foundations for the general class of linear demand functions. Please contact authors for details.

Motivated by the "stochastic choice as optimization" paradigm of [Machina, 1985] and [Cerreia-Vioglio et al., 2019], we suppose that for any set S, probabilities are chosen so that the decision maker maximizes expected utility less quadratic cost:

$$\mathcal{P}(S) = \underset{\rho(.|S)\in\Delta(S)}{\text{argmax}} \sum_{x\in S} \Big(\rho(x|S)u(x) - \frac{c(x)}{2}\rho(x|S)^2\Big) \tag{1}$$

Our key assumption is that the cost is always positive: $c(x) > 0$. We interpret u as a utility function, the benefit of choosing option x, while c represents the marginal cost of increasing the probability that x is chosen.

For the moment we will focus on situations where the solution to (1) features only positive probabilities. As noted by [McFadden, 1974], a zero probability is empirically indistinguishable from a positive but small probability (see further discussion in [Fudenberg et al., 2015]). Thus, one could think that assuming positive probability could be considered "innocuous". That said, in Section B we extend our analysis to allow for 0 probabilities as part of the solution to (1).

Given that all items are chosen with positive probability, the first order conditions to equation (1) are

$$u(x) - c(x)\rho(x|S) = \lambda_S \text{ for all } x \in S$$

where λ_S is the Lagrange multiplier of the constraint that the probabilities add up to 1. Summing across the elements of S,

$$\sum_{y\in S}\frac{u(y)}{c(y)} - 1 = \lambda_S \sum_{y\in S}\frac{1}{c(y)}.$$

Then we have

$$\lambda_S = \frac{\sum_{y\in S}\frac{u(y)}{c(y)} - 1}{\sum_{y\in S}\frac{1}{c(y)}} =: \Lambda(S).$$

Plugging back into FOC, then we get

$$\rho(x|S) = \frac{u(x) - \Lambda(S)}{c(x)} \text{ for all } x \in S. \tag{2}$$

Notice that we can write Equation (2) as

$$\rho(x|S) = \frac{\frac{1}{c(x)}}{\sum_{y\in S}\frac{1}{c(y)}} + \frac{1}{c(x)}[u(x) - \bar{u}_{\frac{1}{c}}(S)]$$

We just illustrated above that ρ is a WL with representation $(u, \frac{1}{c})$ if and only if ρ is the solution to (1) with functions u and c. In other words, Equation (2) provides a closed form solution for the maximization problem defined in (1).

Further, refer back to equation (1). This equation demonstrates that $\rho(\cdot|S)$ is the gradient, with respect to u, of the convex function

$$f(u|S) \equiv \max_{\rho(.|S)\in\Delta(S)} \sum_{x\in S} \Big(\rho(x|S)u(x) - \frac{c(x)}{2}\rho(x|S)^2\Big), \tag{3}$$

where we are treating c as fixed. From this, the symmetry and positive semidefiniteness of cross-term effects follows quite easily. We can explicitly represent $f(u|S) = \frac{1}{2}\sum_{x\in S} \frac{u(x)^2-\Lambda(S)^2}{c(x)}$, with the understanding that $\Lambda(S)$ depends on u.[5] Writing $\rho_u(\cdot|S)$ to emphasize the dependence on u, we also see that $(u|_S - v|_S)\cdot(\rho_u(\cdot|S) - \rho_v(\cdot|S)) \geq 0$, generalizing the monotonicity in u described above.

Figure 3.1 illustrates the model graphically for $X = \{x_1, x_2, x_3\}$. The point inside the triangle represents the choice probabilities from X, and is denoted by $\rho(x_1, x_2, x_3)$. Each green elliptical curve represents an indifference curve in the simplex. The preference is increasing towards $\rho(x_1, x_2, x_3)$. The DM maximizes his utility at this point when he is free to choose from the entire simplex. When x_3 is not available, the DM maximizes his utility subject to being on the $x_1 - x_2$ edge. The choice from $\{x_1, x_2\}$ is determined by the highest level curve having intersection with this line, denoted by $\rho(x_1, x_2)$.

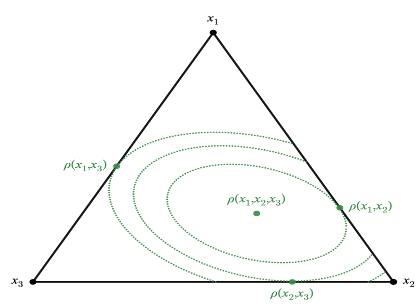

FIGURE 1. Graphical Illustrations

[5]The function $f(u|S)$ is the Fenchel conjugate of $\sum_{x\in S}\frac{c(x)}{2}\rho(x)^2 + \delta_{\Delta(S)}(\rho)$, where $\delta_{\Delta(S)}$ is the convex analytic indicator function of $\Delta(S)$. In particular, this representation is related to the Williams-Daly-Zachary theorem, as in [McFadden, 1981]. See also [Fosgerau et al., 2020].

3.2. **Deliberate Randomization through an Ideal point.** An alternative specification of the model follows. We imagine that the parameters of the model include the *ideal* or *bliss point*, which is the "best" point (in the space of choice probabilities) for the decision maker when choosing from X. The decision maker evaluates probabilities via their "distance" from the ideal point, where the notion of distance is a weighted Euclidean one.

For every S, probabilities are chosen so that the DM minimizes a cost weighted deviation from a given ideal point:

$$\min_{\rho(.|S)\in\mathcal{S}} \sum_{x\in S} \frac{(\rho(x|S) - w(x))^2}{m(x)} \text{ subject to} \sum_{x\in S} \rho(x|S) = 1$$

Here, w represents the ideal choice frequencies for the DM. If DM can attain her ideal point, i.e., $S = X$, the DM chooses the ideal point, i.e., $w(X) = 1$. When some alternatives are not available, the DM cannot achieve her ideal frequencies. So there is a need for re-optimization. Deviation from the ideal point is based on Euclidean distance, where each alternative is weighted by the corresponding inverse cost: m represents inverse of the product specific cost. Again, supposing all choice probabilities are positive, the closed form solution of the minimization problem is:

$$\rho(x|S) = w(x) + (1 - w(S))\frac{m(x)}{m(S)}. \tag{4}$$

It is routine to show that the (w, m)-representation has similar uniqueness properties as the original model.[6] By Equation (4), we have $w(x) > 0$ since $\rho(x|X) > 0$ for all x. We can thus impose the normalization that $\sum_X w(x) = 1$. Then the value $w(x)$ can be interpreted as the "base probability" that alternative x enjoys, since for every S, $\rho(x|S) \geq w(x)$. Clearly, $\rho(x|X)$ is equal to $w(x)$ since $w(X) = 1$. In other words, when choosing from X, each alternative is chosen according to its base probability.

For any set $S \neq X$, after assigning base probabilities, there will be excess probability that must be distributed among items since $w(X \setminus S) = 1 - w(S) > 0$. The

[6]Let (w, m) and (w', m') be two representations. Since $\rho(x|X) = w(x)$, we must have $w(x) = w'(x)$. Given that, we can show that $m(x) = am'(x)$.

weighted ratio of m's dictates how this excess base probability is distributed across different products. The higher the inverse cost, the higher the additional weight ascribed to an alternative. Thus, $\rho(x|S)$ is the sum of base probability of x and the weighted portion of the excess base probability according to m.[7] Notice that in the Luce model, the w parameter defines both the unconstrained "optimum" choice probabilities over X and describes how excess probability will be assigned in a set different from X. In our model, while the w parameter still defines the unconstrained "optimum" probability, m decides the assignment of excess probability independent of w.

Table 1 summarizes the equivalent representations with different interpretations. Since the relationship between primitives of these models are straightforward, we interchangeably use these representations throughout the paper. It turns out that all of these different representations have their benefits in terms of establishing different results. We also provide illustrations of each representation later.

Table 1. Summary of Three Equivalent Representations

Model	Representation	Parameters	Uniqueness
(u, m)	$\frac{m(x)}{m(S)} + m(x)[u(x) - \bar{u}_m(S)]$	u : utility m : salience	$(au + b, \frac{m}{a})$
(u, c)	$\frac{u(x) - \Lambda(S)}{c(x)}$	u : utility c : cost	$(au + b, ac)$
(w, m)	$w(x) + (1 - w(S))\frac{m(x)}{m(S)}$	w : ideal point m : inverse cost	$(w, am)^*$

$^*w$ are uniquely identified after $w(X) = 1$ normalization.

4. Characterization

We now turn to providing behavioral foundations for our model. The first axiom is the often invoked notion of positivity. Positivity says every alternative is chosen with positive probability.[8]

Axiom 1. [Positivity] $\rho(x|S) > 0$ for every $x \in S$ and $S \in \mathcal{D}$.

[7]Note that by defining $u(x) = \frac{w(x)}{m(x)}$, it can easily be shown that ρ has a WL representation.

[8]This assumption cannot be rejected by any finite data set. In addition, it is usually made for estimation purposes. We relax this assumption in Appendix.

The next axiom is a well-known property in stochastic choice literature. It states that when the competition gets fiercer among alternatives, the choice probabilities decreases strictly.

Axiom 2. [Strict Regularity] $\rho(y|S) < \rho(y|S \setminus x)$ for every $x \in S$ and $S \in \mathcal{D}$.

To provide the third axiom, we first define an auxiliary function

$$r_{S,T}(x,y) := \frac{\rho(x|S) - \rho(x|T)}{\rho(y|S) - \rho(y|T)}$$

provided that $x, y \in S \cap T$ and $\rho(y|S) \neq \rho(y|T)$. The quantity $r_{S,T}(x,y)$ measures the relative probability change of x and y from S to T. Given S and T, we only define this function for (x,y) such that $\rho(y|S) \neq \rho(y|T)$. This ratio resembles the ratio that appears in the Luce IIA axiom. The difference is that this is a ratio of relative levels rather than the absolute levels as in Luce's IIA.

This function has several interesting properties: i) $r_{S,T}(x,y) = r_{T,S}(x,y)$ for all S and T, ii) $r_{S,T}(x,y)r_{S,T}(y,x) = 1$, and iii) $r_{S,T}(x,x) = 1$. In the Luce model, this function is constant for any pair (x,y). That is, the ratio is independent of choice of decision problems: $r_{S_1,T_1}(x,y) = r_{S_2,T_2}(x,y)$. Indeed, in the Luce model, both relative and absolute ratios are constant. In contrast, while the absolute ratio might not be constant, the relative ratio is constant in our model. However, it is not strong enough to be sufficient. The next axiom is a stronger version of this idea. The axiom states that this ratio is not only constant but also enjoys a simple transitivity property.

$$r_{S_1,T_1}(x,z) = r_{S_2,T_2}(x,y)r_{S_3,T_3}(y,z) \tag{5}$$

Notice that the function r might not be well-defined for some sets due to dividing by zero. To account for this possibility, we postulate a different form of the same idea. To state our next axiom, we define $d(x|S,T) := \rho(x|S) - \rho(x|T)$, where $S \neq T$ and $x \in S \cap T$. The quantity $d(x|S,T)$ is simply the change in the probability of choosing x as the choice set T changes to S.[9]

Axiom 3. [Relative-IIA] For any x, y, z and $S_i, T_i \in \mathcal{D}$,

$$d(x|S_1,T_1)d(y|S_2,T_2)d(z|S_3,T_3) = d(z|S_1,T_1)d(x|S_2,T_2)d(y|S_3,T_3)$$

[9]Note that d is only defined for (x,S,T) where $S \neq T$ and $x \in S \cap T$. We abuse notation and denote $d(x|\{x,y\},X)$ by $d(x,y)$.

whenever the expressions are well-defined.

Observe that if all differences are non-zero, the axiom is equivalent to the transitivity condition in (5). This axiom states that the multiplication of differences in probabilities of $x \to y \to z$ is the same that of $z \to x \to y$.

In terms of interpretation, the property states that this product of probability differences depends only on the collection of elements with respect to which differences are taken. It does not, however, depend on how these elements are assigned to the given budgets.

It is easy to show that if Axiom 3 holds in the whole domain, then the axiom implies that an analogous condition holds for products of differences of length n, for any $n > 3$. A weaker condition (independence of cycle of 2) is not strong enough to deliver the characterization. That is, $d(x|S_1, T_1)d(y|S_2, T_2) = d(y|S_1, T_1)d(x|S_2, T_2)$. This is the property we discuss above $(r_{S_1,T_1}(x, y) = r_{S_2,T_2}(x, y))$. At the same time, as can be seen by taking $y = z$, Axiom 3 implies this property.

We now state our main theorem.

Theorem 2. *Suppose $\mathcal{D}$ contains all menus with size 2 and 3. Then a stochastic choice function ρ has a WL representation on $\mathcal{D}$ if and only if it satisfies Axiom 1-3.*

The idea of the proof for sufficiency is as follows. We utilize (u, c) representation. We first define the cost of each alternative by using $r_{S,T}$ where S and T are menus with size 2 and 3. Then, instead of directly constructing the utility function, we define the "shadow values" for an optimization problem, for each set in the domain. This step helps us to define the utility function. We then show that the data can be represented by the quadratic model.

Theorem 2 provides two simple tests for our model. While Axiom 2 is innocuous, Axiom 3 is based on a principle similar in spirit to Luce's IIA. In our axiom, the ratio of *relative* levels are important rather than the absolute levels as in Luce's IIA.

Finally, we would like to illustrate that we can make out of sample predictions (outside $\mathcal{D}$) given that our representation is unique.[10] This illustration will help the

[10]The parameters of the models are derived from the domain $\mathcal{D}$.

reader to understand the proof of Theorem 2 better. For example, consider four alternatives $X \equiv \{x, y, z, t\}$, and suppose that $\mathcal{D}$ consists of all sets containing at most three elements. Imagine the choice probabilities from binary and ternary sets satisfy:

(1) Choices from pairs are equiprobable: for all $a, b \in X$, $\rho(a, b) = 0.50$,
(2) For any triple, if y, z, t are members of the triple, then they are chosen with equal probabilities,
(3) x is chosen with probability 0.30 from any triple.

These are our choice data on $\mathcal{D}$. Since $X \notin \mathcal{D}$, choices from the entire set X are not observed. Our goal is to predict choice behavior from X.

When a and b are both chosen with positive probability from two sets, and the probability that they are chosen differs in both sets, then it becomes easy to identify the ratio of their salience parameters: $\frac{m(a)}{m(b)} = r_{S,T}(a, b)$. So, we may conclude directly that $m(y) = m(z) = m(t) = 3/4m(x)$. We may normalize up to scale, so let us suppose that $m(x) = 4$, and that $m(y) = m(z) = m(t) = 3$.

Once m is identified up to scale, we can use the equality

$$u(a) - u(b) = \frac{\rho(a|S)}{m(a)} - \frac{\rho(b|S)}{m(b)}$$

to identify u. Since we may normalize u up to translation, this allows us to choose $u(x) = 0$. In so doing, it becomes apparent that $u(y) = u(z) = u(t) = 1/24$.

With these identifications in hand, we may directly conclude that $\rho(x|\{x, y, z, t\}) = 5/26$, whereas $\rho(y|\{x, y, z, t\}) = \rho(z|\{x, y, z, t\}) = \rho(t|\{x, y, z, t\}) = 7/26$, thus affording an out of sample prediction.

Even outside of this particular situation, our approach allows for very transparent identification of the two parameters. The key function introduced above, $r_{S,T}(x, y)$, identifies m (or c) up to a scale factor: $r_{S,T}(x, y) = \frac{m(x)}{m(y)} = \frac{c(y)}{c(x)}$. Given our identified m's, we can then identify the ranking of u's by defining $u(x) - u(y) = \frac{1}{m(x)}\rho(x|S) - \frac{1}{m(y)}\rho(y|S)$.

5. Discussions

We now turn to discussing how our approach links up to other popular and well known models of random choice. In particular, we consider the behavior of our model in binary environments, relate it to both the RUM and APU models of stochastic choice, and discuss its consistency with various notions of stochastic transitivity.

5.1. **Binary Choice.**

Since [Thurstone, 1927], a key focus of research on random choice has been binary comparisons— Thurstone's model is only built to explain binary choices. In order to understand the WL model in binary choice settings we will utilize the (u, c) representation.

If both alternatives are chosen with strictly positive probability, a simple calculation demonstrates that in the WL model:

$$\rho(x, y) = \frac{u(x) - u(y) + c(y)}{c(x) + c(y)}$$

The comparative statics discussed previously are very apparent in this setting: when the utility of x increases, the probability of x being chosen increases. It is also clear that as $u(x)$ becomes larger than $u(y)$ by an amount equal to $c(x)$, the DM will choose x with probability 1. Hence, y is chosen with positive probability if and only if $u(x) - u(y) < c(x)$. On the other hand, the DM might still choose x in the presence of y even when y yields higher utility than x. Indeed, as long as $u(x) - u(y) + c(y) > 0$ x will be chosen with positive probability.

As previously observed, lowering the cost of an alternative also increases its choice probability. If $c(x)$ is equal to the utility difference, then x must be chosen with probability 1. On the other hand, if x is chosen with positive probability, then increasing the cost of x never results in choosing x with zero probability.

If $u(x) = u(y)$ then the choice probabilities are dictated by costs. In this case, no matter how small or how large the cost is, the alternative is chosen with positive probability.

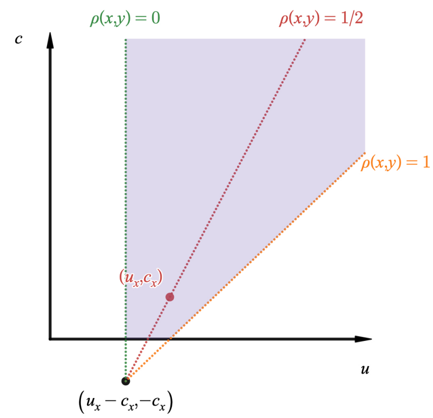

Figure 2. Utility-Cost Diagram

Figure 2, called the utility-cost diagram, illustrates these points. Fix an alternative, say x, with given utility and cost (u_x, c_x) in the utility-cost diagram. The shaded area represents the pairs (u_y, c_y) representing alternatives y for which each of x and y are chosen with positive probability in the binary comparison $\{x, y\}$ $(\rho(x, y) \in (0, 1))$. Each dotted line illustrates the locus of utility-cost pairs chosen with a given probability against x. For example, the red dotted line represents all the alternatives where x is chosen with 0.5 probability in a binary comparison. Interestingly, all such dotted lines intersect at the point $(u_x - c_x, -c_x)$.

The binary world also allows us to transparently link our model to the classic model of Hotelling competition. Recall that we can interpret c as the difficulty of choosing an item — it embeds a kind of friction in the market. Consider a simple Hotelling model where there are two products, x and y, each located at the end-points of a line with length 1. There is a unit mass of consumers uniformly distributed along the line.

Imagine each consumer attaches a utility of $2u(x) - c(x)$ to alternative x. Further, a consumer located at β has to travel to reach either of the alternatives. Assume that travel costs are linear, and that the travel cost per unit of distance is $c(x) + c(y)$. Suppose alternative x is located at 0, and y at 1. Then a consumer located at $\beta \in [0, 1]$ chooses x when

$$2u(x) - c(x) - \beta(c(x) + c(y)) \geq 2u(y) - c(y) - (1 - \beta)(c(x) + c(y)).$$

By setting equality, we get a cutoff β^*, which solves $\beta^* \equiv \frac{u(x)-u(y)+c(y)}{c(x)+c(y)}$. All consumers with $\beta \leq \beta^*$ consume x, and those with $\beta \geq \beta^*$ consume y, thus the proportion choosing x exactly coincides with our model, under parameters u and c.[11]

5.2. **Stochastic Transitivity.** We previously showed that our model has a particularly simple representation when applied to binary choice sets. In the literature, stochastic binary choices are often taken as an indication of a strength of deterministic preference among alternatives. More formally, the *weak stochastic preference relation* P_ρ on X is defined as follows: for any $x, y \in X$, $x \; P_\rho \; y$ if $\rho(x,y) \geq 1/2$.

P_ρ has been interpreted as "underlying" preferences; it is an ordinal ranking. P_ρ is complete by construction but need not be transitive in general. It will be transitive if and only if it satisfies the axiom of *weak stochastic transitivity* (WST).

In our model, we can provide a real-valued representation for P_ρ, thus demonstrating its transitivity. To do so, apply the following transformation $U(x) = 2u(x) - c(x)$, and obtain

$$\rho(x,y) = \frac{1}{2} + \frac{1}{2}\frac{U(x)-U(y)}{c(x)+c(y)} \tag{6}$$

Given Equation (6), we can write

$$x \; P_\rho \; y \text{ if and only if } U(x) \geq U(y)$$

Therefore, in line with the literature, we interpret U as an "implied" utility function of the quadratic model. In Figure 2, the red dotted line (representing $\rho(x,y) = 1/2$) represents the indifference curves of U. Note that P_ρ is transitive in our model. Nevertheless, the final choice is sensitive to the total cost too. Higher the total cost, the DM is less sensitive for the same U difference.

A stronger notion of transitivity for the binary model is *strong stochastic transitivity* (SST). SST states

$$\rho(x,y), \rho(y,z) \geq 1/2 \rightarrow \rho(x,z) \geq \max\{\rho(x,y), \rho(y,z)\}$$

[11]For more than two products equivalent constructions exist, where each product is a vertex on a complete graph, and each competes in a Hotelling fashion over a mass of consumers that are assigned to each edge (please contact authors for details).

In general, our model does not satisfy SST. For example, take $u(x) = 2 = u(z)$ and $u(y) = 1$, $c(x) = 2$, $c(y) = 1$, and $c(z) = 3$. Then $\rho(x, y) = 2/3$, $\rho(y, z) = 1/2$, but $\rho(x, z) = 3/5$.

On the other hand, our model satisfies *moderate stochastic transitivity* (MST):[12]

$$\rho(x, y), \rho(y, z) \geq 1/2 \to \rho(x, z) \geq \min\{\rho(x, y), \rho(y, z)\}$$

A recent paper by [He and Natenzon, 2019] show that any model of binary choices satisfying MST can be represented by the following functional form:

$$\rho(x, y) = F\Big(\frac{u(x) - u(y)}{d(x, y)}\Big)$$

where F is an increasing function and d is a metric on X. Equation (6) establishes that our model has such a representation, where $d(x, y) = c(x) + c(y)$ for all $x \neq y$, $d(x, x) = 0$ and

$$F(t) = \begin{cases} 1 & t > 1 \\ 0.5 + 0.5t & t \in [-1, 1] \\ 0 & t < -1 \end{cases}$$

5.3. **Ranking of Sets and Alternatives.** Our model induces an ordinal ranking of the menus, represented by $\Lambda(S) := \bar{u}_m(S) - \frac{1}{m(S)}$. Note that

$$\rho(x|S) - \rho(x|T) = m(x)[\Lambda(T) - \Lambda(S)]$$

Since for all $x \in X$, $m(x) > 0$, our model implies that $\rho(x|S) \geq \rho(x|T)$ if and only if $\rho(y|S) \geq \rho(y|T)$. Hence, we say T is *weakly preferred to* S, denoted by $T \rhd S$, if for any $x \in S \cap T$, $\rho(x|S) \geq \rho(x|T)$. Clearly, T is weakly preferred to S if and only if $\Lambda(S) \leq \Lambda(T)$. Notice that $\rhd$ captures the indirect utility.

Like us, [Fudenberg et al., 2015] discuss how to rank choice sets, given choice probabilities. They also consider a ranking (again derived from choice probabilities) over alternatives. They say x is *ranked higher than* y if $\rho(x|S) > \rho(y|S)$ for all S. However, our choice probabilities do not induce such a ranking of alternatives. Following, we provide an example demonstrating that there can be two alternatives such that one of them is chosen more often than the other in one choice problem; yet the ranking can be

[12]The proof of this claim is given in Appendix D.

reversed in another choice problem, ($\rho(x|S) > \rho(y|S) > 0$ and $0 < \rho(x|T) < \rho(y|T)$). Of course, our model derives *two* rankings from choice probabilities: u and m.

Example 1. Consider three alternatives x, y_ϵ, and z such that $u(x) = u(z) = 1, u(y_\epsilon) = 2$ and $m(x) = 1$, $m(y) = \frac{1}{3+\epsilon}$, and $m(z) = \frac{1}{3}$. In a binary comparison x is chosen over y_ϵ more often, that is, $\rho(x, y_\epsilon) = \frac{2+\epsilon}{4+\epsilon} > \frac{1}{2} > \frac{2}{4+\epsilon} = \rho(y_\epsilon, x)$ for any $\epsilon > 0$. But introducing z will make y_ϵ chosen with a higher probability than x, i.e., $0 < \rho(x, \{x, y_\epsilon, z\}) < \rho(y_\epsilon, \{x, y_\epsilon, z\})$ for any $1 > \epsilon > 0$. This ranking depends on the third alternative. For example, when another alternative t with $u(t) = 2$ and $m(t) = \frac{1}{3}$ is offered along with x and $y_{0.5}$, we get $0 < \rho(y_{0.5}, \{x, y_{0.5}, t\}) < \rho(x, \{x, y_{0.5}, t\})$. Hence, the ranking between x and $y_{0.5}$ depends on the third alternative.[13]

5.4. **Random Utility Model.** The most well known generalization of the Luce model is the random utility model (RUM). Our model also belongs to RUM. We construct a simple RUM representation for our model. Let $\mathcal{R}$ be the set of all possible linear orders (rankings) on X and π be a probability distribution over rankings. $\pi(\succ)$ represents the probability of $\succ$ being realized as the preference. Given a set of available alternatives A, the probability of an alternative x being chosen is determined by the probability of a ranking for which x is at the top of A. Let $\mathcal{R}(a, A)$ be the set of rankings of X which rank a at the top of A, that is, $\mathcal{R}(a, A) := \{\succ \in \mathcal{R} \ : \ a \succ b \text{ for all } b \in A \setminus a\}$. The RUM stochastic choice associated with π is ρ_π defined by:

$$\rho_\pi(a, A) = \sum_{\succ \in \mathcal{R}(a,A)} \pi(\succ)$$

Our model belongs to the random utility model. Thus, although it is more general than Luce, it still fits within the paradigm of random utility.

Proposition 1. *If ρ has a WL representation then it is RUM.*

We now construct a RUM representation for a given choice function ρ. To do this, we use the (w, m)-representation of ρ. First, consider a ranking of alternatives, enumerated $x_1 \succ x_2 \succ \cdots \succ x_n$. Then our RUM representation assigns the following

[13]While Example 1 illustrates that there is no global ranking of alternatives in general, as discussed below our model induces a complete and transitive ranking for binary sets.

weight the corresponding ranking:

$$\pi(\succ) = w(x_1)\frac{m(x_2)}{m(X \setminus x_1)}\frac{m(x_3)}{m(X \setminus \{x_1, x_2\})}\cdots\frac{m(x_n)}{m(x_n)}.$$

This proof mimics the construction of [Falmagne, 1978]. It is then standard to show that $\rho = \rho_\pi$.[14]

5.5. **Additive Perturbed Utility:** Our discussion of micro-founding our approach by explicitly modeling it as emerging from an optimization process brings to mind the Additive Perturbed Utility (APU) model ([Fudenberg et al., 2015]). APU is one for which

$$\rho(x|S) \equiv \arg\max_{p\in\Delta(S)} \sum_{x\in S} \left[u(x)p(x) - k(p(x))\right],$$

where k is some strictly convex and smooth function.

Although this formulation is similar to our approach, the distinction is twofold:

(1) The cost of probability is independent of the alternative in question (k depends only on the probability of x but not x itself)
(2) The function k need not be quadratic.

In their working paper, [Fudenberg et al., 2014] weaken the first condition, and consider the more general model:

$$\rho(x|S) \equiv \arg\max_{p\in\Delta(S)} \sum_{x\in S} \left[u(x)p(x) - k(p(x), x)\right]$$

which nests our approach. They fully characterize this using an acyclicity condition.

[Fudenberg et al., 2015] show that the APU approach generates two orderings. The first ordering is over sets, analogous to our ordering $\rhd$: it corresponds to the ordering induced by λ over choice sets. The APU approach also generates a partial ordering over outcomes, which is determined by the relative probabilities that items are chosen in a particular choice set. In particular, it means that if x is chosen with higher probability than y in S where both are available, then in all T where both are available, x needs to be chosen with higher probability than y. As we have shown, our model lacks such an ordering — it is entirely possible that $\rho(x|S) > \rho(y|S) > 0$ and $\rho(y|T) > \rho(x|T) > 0$.

[14] This representation is not unique in general. See Turansick [2021] for details.

However, our model does have a non-trivial intersection with the APU class. Due to the discrete nature of the domain, one can find many situations where choices are represented both with our model, as well as the APU model. However, we want to highlight here three models that, regardless of the domain, are consistent with both models. One obvious point of intersection is the Luce model. The second obvious point is when c is constant, and so the cost function is the same for all outcomes. This is exactly the situation where we recover simple linear demands. The third type is where u and c are affine transformations of one another.

6. Applications

We now turn to demonstrating the applicability of our proposed functional form. We use our model of random choice in a very standard "product" setting. We aim to show that our model is

(1) tractable in this setting, allowing for closed form solutions and analysis,
(2) easily accommodates behavior that is considered intuitive, but which the standard logit approach cannot allow, and
(3) has transparent identification in oft-used empirical settings.

We present five applications. The first formalizes the comparative statics generated by the WL model: how choice probabilities change with utility and salience. The second instead considers context dependent choice — we consider how consumer choice probabilities change as the surrounding choice set changes. Third, rather than simply assuming that the product characteristics are exogenous, we turn to endogenizing the choice of u and c. In particular, we consider a situation where firms can endogenously affect both u and c. We suppose firms can compete with each other by charging a price or quality (which change u), and can also engage in actions which change the salience of their product, for example via advertising (changing c). Thus, in a setting featuring firm competition, our model allows for a richer set of behaviors than many other models of random choice, which rely on firms manipulating a single utility ranking. We show that our model generates closed form solutions that have intuitive properties.

Fourth, with a basic understanding of what kind of behavior firms will engage in, we turn to demonstrating that our model can accommodate several empirical patterns of choice that the logit model, and any discrete choice model with unbounded and continuous support for utility shocks (*i.e.* essentially all the discrete choice models in the applied literature), cannot. Thus, our model can reasonably approximate some kinds of market outcomes.

Our previous results on identification have relied on variation in choice sets. This raises questions of how to identify the parameters in our model where the choice set is fixed (the standard situation in many empirical industrial organizational papers), but where product attributes are observable. Although our model lacks certain typical sufficient conditions often used for identification in discrete choice settings, the fact that it has explicit micro-foundations via optimization allows for extremely tractable identification. In essence, given observable attributes, a set of linear equations serve to identify preference parameters. This sidesteps the issues of non-linearity embedded in many estimation techniques.

6.1. **Comparative Statics.** In order to increase our understanding of the model, we now formally discuss its comparative statics. First, we establish that increasing either $u(x)$ or $m(x)$ will result in an increase in the corresponding choice probability. Observe that there is a direct effect, as well as an indirect effect on $\rho(x|S)$ when we change either parameter. The indirect effect emerges via the change in the market average.

To see the indirect effect of $u(x)$, the derivative of $\bar{u}_m(S)$ with respect to $u(x)$ is $\frac{m(x)}{m(S)} \in (0, 1)$. The total effect is then

$$\frac{\partial \rho(x|S)}{\partial u(x)} = m(x)\Big[1 - \frac{m(x)}{m(S)}\Big].$$

As $m(x) > 0$, increasing the utility of x increases the choice probability. This relationship is linear for a given S: the marginal change according to utility is fixed for a given product. The effect of changing u is “unbounded”: for any choice set, reducing $u(x)$ by enough will result in x being chosen with probability 0. Similarly, increasing $u(x)$ enough will eventually lead to all other items being chosen with 0 probability.

We now investigate how the probability that x is chosen from S changes as the utility of another alternative y changes. As expected, this relationship is linear but inversely related to the effects described above. Moreover, we have symmetry as

$$\frac{\partial \rho(x|S)}{\partial u(y)} = \frac{\partial \rho(y|S)}{\partial u(x)} = -\frac{m(x)m(y)}{m(S)},$$

analogous to the Slutsky substitution patterns. Positive semidefiniteness follows as well, because this matrix of substitution patterns is obviously diagonally dominant, with positive diagonal elements.[15]

In contrast, if $m(x)$ increases, there are again both direct and indirect effects on choice probabilities, with the overall effect being positive. In particular, the derivative of $\bar{u}_m(S)$ with respect to $m(x)$ is $\frac{u(x)m(S)-\sum_z u(z)m(z)}{m(S)^2}$. Thus, the derivative of $\rho(x|S)$ with respect to $m(x)$ is

$$\frac{\partial \rho(x|S)}{\partial m(x)} \quad = \quad \frac{\rho(x|S)}{m(x)}\left[1 - \frac{m(x)}{m(S)}\right]$$

One can interpret this as follows: if a product is already chosen with positive probability, then as $m(x) > 0$, increasing the salience of x will increase the choice probability of x. Unlike with u, the the effect of m on choice probabilities is non-linear. If $u(x) \geq \bar{u}_m(S)$, then the probability of x being chosen will go to 1. If not, then the probability of x being chosen converges to an upper bound strictly below 1.

As $m(x)$ decreases, the probability of x being chosen falls to 0. Here, we do not have symmetric responses to cross-salience effects. In particular

$$\frac{\partial \rho(x|S)}{\partial m(y)} \quad = \quad -m(x)\frac{u(y)m(S) - \sum_z u(z)m(z)}{m(S)^2}$$

This should be relatively intuitive. The effect of changing z's salience should depend on the quality of z, $u(y)$, since higher utility items benefit more from increased salience.

6.2. **Context Effects.** In this subsection, we explore how changing the choice set S changes the probability with which outcomes are chosen.

The traditional way context effects have been studied in marketing, economics and finance ([Huber et al., 1982]) is by considering how relative choice probabilities

[15]The fact that $\rho(x|S)$, viewed as a function of u, is actually the subgradient of a specified convex function, as discussed elsewhere, explains this.

between two options change when a third option is added. Here, we both address our model's predictions when moving from binary to ternary choice sets, as well as more general changes in choice sets. A key intuition, implicit in our axioms, but which we highlight explicitly here, is that adding additional alternatives to a choice set always helps (in terms of relative choice probabilities) high utility outcomes.

For example, assume that there are two distinct products receiving equal market share in a binary choice set. That is, $x \neq y$ and $\rho(x, y) = 0.5$. Then the product with higher utility will get more market share compared the other alternative when a third alternative is introduced. That is, $\rho(x|\{x, y, z\}) \geq \rho(y|\{x, y, z\})$ whenever $u(x) \geq u(y)$.

Remark 1. *Including a third alternative benefits the higher utility alternative given that they are chosen equally in binary comparison.*

This result holds not only when moving from binary to ternary choice sets, but also in larger choice sets. Adding additional options always benefits the relative choice probabilities of higher utility options.

Proposition 2. *Suppose $u(x) \geq u(y)$. Then $\rho(x|S) \geq \rho(y|S)$ implies $\rho(x|S \cup T) \geq \rho(y|S \cup T)$. Moreover, it is possible that $\rho(y|S') > \rho(x|S')$ yet $\rho(x|S' \cup T) \geq \rho(y|S' \cup T)$.*

A distinct consideration is what happens when choices are replicated. A replicate only differs from the the original product in terms of seemingly unimportant attributes, such as in the famous "red bus-blue bus" example. Again we assume that there are two distinct products receiving equal market share in a binary comparison. In our model, if we replicate x, then the total market share that x and its replica capture from the ternary choice set is larger than the market share x captures in the binary choice set. Let x' be a replica of x, $u(x) = u(x')$ and $m(x) = m(x')$. Then $\rho(\{x\}|\{x, y\}) < \rho(\{x, x'\}|\{x, x', y\})$.

Remark 2. *Replication helps to increase the total market share.*

Our previous remark also allows us to make an even stronger claim. If $u(x) \geq u(y)$ then adding x' not only shifts market share away from y, but can in fact increase the relative market share of x compared to y — $\rho(x|\{x, y, x'\}) > \rho(y|\{x, y, x'\})$. This is the "similarity effect" for lower utility products. In our model, introducing a replica of a lower utility product reduces the average utility in the market. This is beneficial for the higher utility product due to a higher deviation from the weighted average.

Remark 3. *Replication of lower utility product creates the similarity effect.*

In larger choice sets, if we introduce a large number of replicas of x, this eventually drives out of the market all outcomes with a utility level lower than x.

Proposition 3. *For all x, if $\rho(x|S) > 0$, and denoting T_n as S with n replicas of x, then as n goes to infinity, for all $y \in S$, $\rho(y|T_n) > 0$ if and only if $u(y) > u(x)$.*

6.3. **Bertrand-Advertising Equilibrium.** Up until now, the paper has focused on how consumers respond to a particular choice set. In this subsection, we ask, given consumers' behavior, how do firms strategically interact in order to maximize profits, given the opportunity to manipulate the choice set, via choice of c and u.

The flexibility embodied in our model allows us to easily capture novel considerations outside of the typical logit approach, as well as many other one-parameter random utility models. In particular, our model can be seen as a random choice model with microfoundations that allows for a natural way of thinking about advertising. The WL model easily allows us to explore what happens when firms can compete on salience. In line with our interpretation of c above, one can think of reducing c as reducing the mental friction required to purchase the item, through increasing its salience, say via advertising. As [Bagwell, 2007], points out, advertising has been seen by economists as having three approaches: persuasive, informative, and complementary. Our model is closest in spirit with the informative approach to advertising, where advertising helps raise "awareness" of a product.

Thus, changes in c can be seen as controlling the extent of "competition" in the market. By allowing firms to choose c we can endogenize the degree of competition in the market. This intuition can be more clearly seen by the fact that $\frac{1}{c}$ and u are complements in terms of choice probabilities: the cross-second order condition is $(1 - \frac{\frac{1}{c}}{\sum_j \frac{1}{c_j}})^2 > 0$.[16] Thus, reducing c increases the impact of adjusting the utility of an item — increased salience increases the returns (in terms of choice probabilities) to improving an item. Similarly, firms with higher quality, i.e. a higher u, find adjusting c more beneficial.

[16] This is distinct from the complementarity in terms of utility discussed in [Becker and Murphy, 1993].

In order to demonstrate our results, we focus on a simple, stylized setting. There are n firms indexed by $\{1, 2, \cdots, n\} = N$. Throughout we assume that all firms have the same (constant) marginal cost of production of k.

We will work with the u, c (utility, cost) representation described previously. We assume that the utility of product i is actually the difference between two things: a quality u_i, and a price p_i, i.e. $u_i - p_i$. The friction in choosing i (i.e. the inverse of salience) is c_i.

We can consider what happens as firms can control more and more of the variables. First, we can consider what happens when the firm only chooses p_i, and c_i and u_i are exogenous. In other words the firms only compete on price. We can then allow the firms to compete on advertising as well, and make, in addition, c_i a choice variable. Last, we can also allow the firms to choose quality, and thus have u_i as a choice variable.

We will think of c_i as being proportional to the amount of advertising — we represent advertising as $\frac{1}{c}$ — so that higher c_i corresponds to less advertising. If firms can chooses advertising level c_i we will assume for the sake of simplicity that the cost of advertising level c_i is $\gamma(c_i) = \frac{g}{c_i^2}$, where g is the marginal cost of increasing advertising. Similarly, we suppose, if the firms can choose quality that firms pay some fixed cost, such as R&D costs, to develop a particular quality, where the cost for developing quality u_i is hu_i^2. Thus h is the marginal cost for quality improvement. We also allow the total size of the market (i.e. the number of consumers), to vary, and denote the size by M. We will assume that that revenues depend on M, but that advertising costs (and R&D costs) are independent of the size of the market.

Given these assumptions, the objective function for any given firm i is

$$M(p_i - k)\left(\frac{u_i - p_i - \Lambda(N)}{c_i}\right) - \frac{g}{c_i^2} - hu_i^2$$

where $\Lambda(N) = \frac{\sum_{j \in N} \frac{u_j - p_j}{c_j} - 1}{\sum_{j \in N} \frac{1}{c_j}}$. Such a formulation, when rewritten, is equivalent to

$$M(p_i - k)\left(\frac{\frac{1}{c_i}}{\sum_j \frac{1}{c_j}} \frac{1}{c_i}\left(u_i - p_i - \overline{(u_j - p_j)}_c\right)\right) - \frac{g}{c_i^2} - hu_i^2$$

where $\overline{(u_j - p_j)}_c$ is the c weighted average of $u_j - p_j$. Thus, demand is linear in prices.[17]

For the moment we will focus on environments where we can find symmetric equilibria, and so will suppose that exgenous variables are the same across all firms, and that all firms take the same strategies. We initially suppose that n firms exist in the market, but will later consider what happens when there is entry and exit. We also suppose k is small enough so that an equilibrium with positive firms profits exists for a fixed n.[18]

To solve for the symmetric equilibrium in the most general case (the simpler cases are easily derived in a similar fashion), first we take the first order conditions for p_i, c_i and u_i. The first order condition with respect to p_i is (where $\Lambda(N)$ is the Lagrangian multiplier associated with the available choice set):

$$\frac{u_i - p_i - \Lambda(N)}{c_i} + (p_i - k)\left(\frac{-1 - \Lambda_p(N)}{c_i}\right) = 0$$

The first order condition with respect to c_i is

$$M(p_i - k)\left(\frac{-\Lambda_c(N)c_i - (u_i - p_i - \Lambda(N))}{c_i^2}\right) + \frac{2g}{c_i^3} = 0$$

The first order condition with respect to u_i is

$$M(p_i - k)\left(\frac{1}{c_i} - \frac{\Lambda_u(N)}{c_i}\right) - 2hu_i = 0$$

Since the firms are symmetric, we assume $p_i = p$, $c_i = c$ and $u_i = u$. Then we have $\Lambda(N) = u - p - \frac{c}{n}$ and $\Lambda_p(N) = -\frac{1}{n}$, $\Lambda_c(N) = -\frac{1}{n^2}$, and $\Lambda_u(N) = \frac{1}{n}$

[17] This approach does not nest the standard Bertrand price competition model with a logit demand function. This is because with logit demand, price would affect the utility in the logit model, i.e. $\frac{1}{c}$. One can easily show that we obtain the standard logit results if we allow for price, rather than affecting u, to affect $\frac{1}{c}$.

[18] Therefore, in a symmetric equilibrium all products will be purchased with positive probability.

Substituting them back into the first order conditions, we get

$$\frac{1}{n}=(p-k)(\frac{n-1}{nc}),\; M(p-k)(\frac{\frac{c}{n}-\frac{c}{n^2}}{c^2})=\frac{2g}{c^3}\text{ and }2hu=M(p-k)\frac{n-1}{cn}$$

which yields

$$p=k+\frac{c}{n-1},\; c=\frac{(2gn^2)^{1/3}}{M^{\frac{1}{3}}}\text{ and }u=\frac{M}{2hn}$$

The next proposition summarizes these findings, making appropriate substitutions.

Proposition 4. *With n firms the symmetric equilibrium has the following properties:*

- *If $u_i=u$ and $c_i=c$ are exogenous and symmetric across firms, and p_i is a choice variable, then $p_i=k+\frac{c}{n-1}$.*
- *If $u_i=u$ is exogenous and symmetric across firms and p_i and c_i are choice variables, then $p_i=k+\frac{c_i}{n-1}$ and $c_i=\frac{(2gn^2)^{\frac{1}{3}}}{M^{\frac{1}{3}}}$.*
- *If p_i, c_i and u_i are choice variables then $p_i=k+\frac{c_i}{n-1}$ and $c_i=\frac{(2gn^2)^{\frac{1}{3}}}{M^{\frac{1}{3}}}$ and $u_i=\frac{M}{2hn}$.*

When the number of firms is exogenous, our model implies that markups converge to 0 as the number of firms grow — in contrast to the equilibrium in the logit model, where the limit price as the number of firms becomes large is $\frac{k+u}{2}>k$. Moreover, both price and advertising are independent of the cost of R&D, while quality is independent of the cost of advertising.

In this symmetric equilibrium, under each of the three scenarios,[19] firm's profits are

$$\frac{Mc}{(n-1)n}$$

$$M^{\frac{2}{3}}\frac{g^{\frac{1}{3}}(n+1)}{2^{\frac{2}{3}}(n-1)n^{\frac{4}{3}}}$$

and

$$M^{\frac{2}{3}}\frac{g^{\frac{1}{3}}(1+n)}{2^{\frac{2}{3}}(n-1)n^{\frac{4}{3}}}-\frac{M^2}{4hn^2}$$

respectively. If there is a fixed cost F of entry, and firms must earn 0 profits in equilibrium, then the equilibrium number of firms is the largest integer n such that

[19]We assume if a variable is exogenous it imposes no costs.

profits are (weakly) larger than F. We can use this to understand how the number of firms change as exogenous parameters change.

For example, consider the situation where only u_i is exogenous, so that firms choose prices and advertising. As the cost of advertising grows (i.e. g increases), we see less advertising, and a larger number of firms. The increase in c (i.e. reduced advertising) as g increases happens via a direct effect (as c directly increases with g), and an indirect effect, where c increases via the increase in n. There is similarly a direct (via g) and indirect (via n) effect on price. Thus, price, and so markups, are increasing in the equilibrium number of firms n (this is the opposite of the partial equilibrium reasoning, where an exogenous increase in n leads to lower markups). With firm entry and exit, we would expect low advertising to be correlated with high markups, and with a larger number of firms.

Intuitively high advertising costs means firms cannot use advertising as effectively to gain market share, and so firms compete less with each other on price. This leads to higher markups, as well as more firms in the market overall. In other words, with free entry markets with many firms are those in which it is hard to advertise, driving up markups.

Of course, we are not the first to think about the relationship between prices and advertising; there is an extensive literature discussing this issue. Our result, a correlation between low prices and high levels of advertising, is reminiscent of results that emerge in completely different contexts in [Robert and Stahl, 1993] and [Bagwell and Ramey, 1994] (who find that advertising is greater when prices are lower), and which have been extensively investigated in the literature (see [Bagwell, 2007] for a relatively recent survey).

If quality can also be chosen by firms, then the comparative statics can be somewhat different. It could be that if adjusting quality is cheap enough, then increases in g cause firms to advertise less, but in order to compete with other firms, they instead raise the quality of their products. The increase cost of quality drives some firms out of the market, and so we end up with a smaller number of firms.

With a fixed cost of entry, more and more firms enter as M grows if u_i is exogenous. This immediately implies that as the market grows large, the number of firms grows large, and so we observe fragmentation. A seminal contribution, [Sutton, 1991], summarizes and extends a literature that focuses on market fragmentation: i.e. what happens to the number of firms as the size of the market increases (in particular, does the number of firms converge to infinity). The key insight in [Sutton, 1991] is that whether or not the market becomes fragmented depends on the structure of fixed costs in market.

Observe that our model may not necessarily generate market fragmentation. If firms get to choose quality as well as advertising and price, then firm profits are not always increasing in M. This immediately implies that the number of firms does not necessarily increase without bound as n goes to infinity.[20]

We view this flexibility as a feature, not a bug, of our approach. As shown in the empirical work in [Sutton, 1991] as well as an extensive follow-up literature (see [Sutton, 2007] for a survey), the relationship between market size and market fragmentation varies extensively by industry specific factors, including the ability to adjust quality and advertising. The flexibility of our approach allows us to capture these different outcomes.

Our model can be extended in a number of ways. For example, everything discussed up until now supposes a symmetric environment and equilibrium, where all firms face the same exogenous parameters. The WL model can be extended to incorporate asymmetric equilibrium, where firms have different exogenous levels of quality, advertising costs, or marginal costs of production.

In a distinct vein, we can also study the extent to which the degree of advertising is efficient (as in a large literature beginning with [Butters, 1977]). The micro-foundations for the (u, c) representation of WL model implies that advertising directly creates consumer surplus (a mechanism developed in other papers, e.g., by [Grossman and Shapiro, 1984] in a horizontally differentiated market). Thus, our model can directly link the amount of advertising in the market to consumer surplus via changes both in price, the mental costs associated with choice.

[20] This result depends on the fact that price, quality and advertising are all choice variables. If only one or two of the three are choice variables then we observe fragmentation. In a similar vein one can easily show that with a Luce model of price and quality competition market fragmentation must occur.

6.4. **Accommodating Empirical IO Patterns.** Despite the tractability of the logit model, which has made it a workhouse model in demand estimation, it has many limitations that have been pointed out over the years. Many alternative approaches have been developed to try and better accommodate empirical patterns. However, even within the class of typically used discrete choice models, certain choice patterns can not be rationalized. We discuss how our model can accommodate some of these patterns of choice, providing four examples. The examples are organized so that later examples address larger classes of models than earlier examples, beginning with the logit model in the first example and ending with discrete choice models with unbounded and continuous error distributions in the last. We begin by discussing how the cross-price substitution patterns generated by our model are more flexible than those generated by the logit approach. Second, we highlight the fact (mentioned briefly previously) that in the WL model markups go to 0 in large markets, unlike what happens in logit, GEV, and random coefficients logit models. Third, unlike most discrete choice models in the literature, which assume that choice errors are continuous and have unbounded support, we can naturally allow for zero demand for some goods. Last, again like most discrete choice models in the literature, some products (i.e. dominant products) can have a market share which is strictly bounded away from 0, even as the number of products goes to infinity.

(1) **Cross-Price Substitution Patterns**: It is well known that in the logit model that the cross price elasticity of i for k do not depend on i, a condition that is at odds with both intuition and reality.

In contrast, our demand functions allow for the substitution patterns to vary across items.[21] In order to demonstrate this point, we consider the model in the previous subsection. The cross price elasticity of i for k is

$$p_k \frac{1}{c_k(u_i - p_i - \Lambda(N)) \sum_j \frac{1}{c_j}}$$

Thus, in contrast with the predictions of the logit model, our cross-price elasticities depend on both i and k. This also distinguishes it from models such as nested logit, where within each "category" the same restrictive cross price elasticity pattern still applies. Moreover, given the parameters, the cross-price elasticities in the WL model are easily computable.

[21] Other approaches that do this already exist in the literature — e.g., random coefficients approaches are also designed to get around these restrictive cross price substitution patterns.

(2) **Firm Markups and Substitutes:** In logit models, even though the share of the market that each firm has typically goes to 0 as the number of firms grows large, it is not the case that firms lose all their market power. In fact, as is well known, in logit specification, firms maintain market power even with an infinite number of firms. Consumers are unwilling to simply substitute away from a particular product. This can be seen from the fact that firm markups, even with a large number of competitors, are strictly larger than 0. As [Benkard et al., 2001] point out, this issues applies more generally to GEV and random coefficients logit models.[22]

In contrast, in the WL approach, with Bertrand price competition alone, we can accommodate this pattern. As noted in the previous subsection, in a symmetric price setting equilibrium price satisfies the equation $p = k + \frac{c}{n-1}$, which goes to 0 with the number of firms.

(3) **Zero Demand**: The logit approach to random choice has the well-known issue that it cannot allow for zero choice probabilities. In particular, so long as the net utility of any item is positive, it must be chosen with positive probability.[23] [Benkard et al., 2001] show that under very mild assumptions (always satisfied in applications of discrete choice models), this issue extends to any model where the conditional error distributions have unbounded upper support and a continuous upper tail. This immediately implies that any item which is chosen with positive probability in one choice set must be chosen with positive probability in all choice sets.

Both of these implications are often violated in reality. Our model can accommodate both zero-probabilities and for some items to be chosen with zero probabilities in some choice sets but not in others. As discussed previously, adding additional competitors can drive demand for i from a strictly positive level to 0. In other words, our model allows for increased competition to drive some existing firms out of the market, even if they generate positive net utility.[24]

[22] Although they also mention that probit models can avoid these implications.

[23] Many specifications of the logit model impose the requirement that net utilities (i.e. gross utility less price) are always positive (e.g., if the probability of choosing x is $\frac{e^{v_i - p_i}}{\sum_j e^{v_j - p_j}}$), but this result still holds even for logit models where net utility can be negative (e.g., if the probability of choosing x is $\frac{v_i - p_i}{\sum_{j:v_j - p_j > 0} u_j - p_j}$) so long as this number is greater than 0, and 0 otherwise.

[24] However, our approach still puts some structure on zero probability choice — if firm i has 0 demand facing a set of competitors S, then adding additional competitors can never increase its demand above 0.

To see this, consider the Bertrand price competition model in previous section. Observe that even if $u_i - p_i > 0$, if it is small enough we can construct choice sets T and S where $T \subset S$ and where $u_i - p_i - \Lambda(S) < 0 < u_i - p_i - \Lambda(T)$. Thus x will be chosen with positive probability in T but with 0 probability in S, despite it generating strictly positive net utility, conditional on purchase. At the same time, there could also be a $y \in T$ such that y is chosen with positive probability in both T and S.

(4) **Negligible Shares in Large Markets**: Although, as just discussed, many models imply that demand cannot be 0 for any item that generates positive net utility, it simultaneously implies that in large markets (i.e. markets with large numbers of products), demand for any one item must converge to 0. [Benkard et al., 2001] discuss how this is true of (focusing on the case of where there is an outside good present) in the models where, in addition to mild technical conditions, the support of the error terms is unbounded above and continuous.

In contrast, the WL model, while allowing for 0 probabilities, can also allow for non-negligible markets shares even as the number of products grows infinitely large. Moreover, this can happen even when each of the additional products attract strictly positive probability.

Thus, our model can accommodate situations where the market is composed of a (fixed) set of "dominant" products and a set of "inferior" products. As the number of inferior products in the market grows, each additional inferior product is chosen with positive probability, but this probability goes to 0 with the number of inferior products. In contrast, the dominant products continue to be chosen with a probability bounded strictly away from 0.

The simplest example to see this is where we have one dominant item x. Recall the probability that x is chosen is:

$$\frac{u_i}{c_i} - \frac{1}{c_i} \frac{\sum_j \frac{u_j}{c_j} - 1}{\sum_j \frac{1}{c_j}}$$

Notice that in order to ensure this stays bounded away from 0, even as the number of products goes to infinity, a sufficient condition is to make sure that $u_i - \frac{\sum \frac{u_j}{c_j} - 1}{\sum \frac{1}{c_j}}$ is bounded away from 0.

To see that this can happen, suppose we initially consider choice set $A = \{x, y\}$ where $\rho(x|A) > \rho(y|A) > 0$. Moreover, suppose that $u_x > u_y$ (so that y is an inferior item). Now, we increase the choice set by adding replicas of

y. Because y was chosen with positive probability from A, we know that the replicas will also all be chosen with positive probability. Thus, denoting S_n as x along with n replicas of y, $\lambda(S_n)$ converges to u_j. In the limit $\rho(i|S_n)$ will go to $\frac{u_i}{c_i} - \frac{u_j}{c_i}$ which is positive and bounded away from 0.

6.5. **Identification with Attributes.** One major reason for the popularity of discrete choice models, of which logit is a prime example, is that they have found wide application in a variety of setting where they are used to estimate consumer preferences. Thus, a prime consideration for models of random choice are the extent to which they are readily identifiable from typical data in the industrial organization literature, e.g., market shares and product characteristics.

Here, we turn to discussing identification of our model in a standard empirical setting — with a fixed choice set, but observable product characteristics.[25] Our previous identification results rely on observing multiple choice sets, and seeing how market shares (*i.e.* choice probabilities) change with the set. In contrast, here we will focus on having a single choice set, and ask how choice probabilities can change with observable characteristics.[26]

Of course, as the previous results in this paper should make clear, given a single set S and choice probabilities $\rho(i|S)$ for all $x \in S$ we can always construct a WL model that rationalizes the data. In other words, with no additional information our model is not falsified by observing any single choice set. Similarly, we cannot uniquely identify u and c with such data.

However, if we assume (as is typical) that both u and c are functions of observable attributes, then identification proceeds in a clear manner. In particular, suppose that there is a set of observable attributes, of cardinality m, with a_i denoting the vector of attributes for product i. a_i includes not only things that affect product quality, but also things like price, advertising, etc.

Typically utility is assumed to be a linear function of attributes. We maintain the same assumption here and extend it to c. Thus, we assume that there exists a vector

[25][Allen and Rehbeck, 2019] proposes an axiomatic theory in such a framework.

[26]An additional consideration in typical applications is that there may be endogenous, unobserved, characteristics to products. Given that there is extensive discussion of this issue in the literature, we abstract away from it, and suppose that all attributes are observable to the researcher.

β such that $u_i = \beta a_i$ for each i. Similarly there exists a vector α such that $c_i = \alpha a_i$ for each i. Thus, we assume that attributes affect utility in the same way for all products, and similarly for costs. The only difference between products is the value that each attribute takes on.

Then under relatively mild conditions our model is identified. Moreover, it can be identified using standard linear equations.

Proposition 5. *Suppose that $u_i = \beta a_i$ and $c_i = \alpha a_i$ where a_i is a $m \times 1$ vector. Suppose that we have at least $2m$ linearly independent observations of $(\rho(i)a_i - \rho(j)a_j, a_i - a_j)$ for $i, j \in S$. Then β and α are identified from choices in S up to positive scalar multiplication.*

For identification of the preference parameter vectors β and α, corresponding to the weights c and u put on each attribute, one simply finds the solutions to the system of equations: $\beta[\rho(i)a_i - \rho(j)a_j] = \alpha[a_i - a_j]$ (where there is one equation for each pair of outcomes). The equations are linear in the attributes, making this a computationally simple problem.

Notice that we obtain one equation for each pair of outcomes. Thus, fixing a number of attributes m, as long as there are (i) enough products and (ii) enough linearly independent combinations of attributes across products, the model is identified. In particular, if the choice set has $|S|$ products we will have $\frac{|S|(|S|-1)}{2}$ pairwise comparisons. If all pairs of comparison are linearly independent then all we need for identification is that $\frac{|S|(|S|-1)}{2} \geq 2m$.

Recall that u, c is unique up to transformations of the type $\kappa u + \gamma, \kappa c$ where $\kappa > 0$ and for any γ. Observe that because we assume that $u_i = \alpha a_i$, we impose that $\gamma = 0$. Thus, the representation is unique up to transformations of the form $\kappa u, \kappa c$. Notice that this occurs if and only if α, β are unique up to transformation $\kappa\alpha, \kappa\beta$.

In order to highlight our approach, we will consider a stylized example. Suppose we have products with two attributes, and there are four products (the minimum needed for identification). Moreover we observe choices from the grand choice set (again, necessary for identification), and $\rho(1|\{1,2,3,4\}) = 0.273$, $\rho(2|\{1,2,3,4\}) = 0.197$, $\rho(3|\{1,2,3,4\}) = 0.121$, and $\rho(4|\{1,2,3,4\}) = 0.409$. Moreover, suppose $a_1 = (4,2)$; $a_2 = (4,1)$; $a_3 = (2,8)$; $a_4 = (4,8)$. Then we have 6 pairwise comparisons. One can show that each of the six entries in the $(\rho(i)a_i - \rho(j)a_j, a_i - a_j)$ are linearly independent of all the others. Thus, our conditions are met. Using any subset of 4 of the pairwise comparisons delivers the result that, using α_1 as the "numeraire" coefficient, $\beta_1 = 3\alpha_1$, $\beta_2 = \alpha_1$ and $\alpha_2 = 2\alpha_1$. In this world, we can do out of sample predictions, which include introducing a new product or improving existing product. For example, if product 1 is improved in terms of attribute 1 (i.e., $a_{11} \geq 6$), the product 2 will be driven out of the market.

Our approach is distinct from that typically used for discrete choice models, as outlined in [Berry, 1994]. In fact, [Armstrong and Vickers, 2015], building on the work of [Jaffe and Weyl, 2010] and [Jaffe and Kominers, 2012], show that although linear demands (which our model is an example of) can be consistent with a model of discrete choice, they fail some standard assumptions about the distribution of the utility shock, assumptions which are used to ensure identification in standard discrete choice approaches (i.e. continuity and full support assumptions on the error term).

7. Conclusion

This paper has introduced a new model of stochastic choice: the weighted linear model of discrete choice. The choice probabilities of any given product depend on two dimensions, the utility of an item, and the "weight" or salience of an item. The second dimension can be interpreted as describing how difficult it is to choose an item. The choice probabilities can be seen as coming from a simple model of optimal choice. We demonstrate that our model sits at the intersection of the classic models of random utility and models of deliberate stochastic choice.

The weighted linear model represents a generalization of the classic model of Luce, and, as such, provides more explanatory power. It is closely related to well-known models in which demand is linear in the utility differences between products. Moreover, it also overcomes many problematic implications of other random/discrete choice approaches used in the literature. At the same time, it is quite tractable. The model lends itself naturally to describing consumers who experience some frictions in being able to choose the best item (whether they be physical or attentional frictions). When used in models of strategic firm interactions, it can generate intuitive closed form solutions that shed light on advertising, quality choice, and the number of firms serving a market. Moreover, we show our model's parameters are easily identified from either choices from across choice sets or choices from within choice sets when the parameters are linear functions of product characteristics (as is often assumed).

Our hope is that the flexibility of our model, its intuitive approach to choice as depending on both utility and a friction, and its tractability can help economists better understand market interactions between firms and consumers. In particular, the fact that our model allows for intuitive empirical patterns, such as flexible patterns of cross-price substitution patterns, or the existence of dominant market shares in large markets, can lead to new insights in many markets where these kinds of behavior need to be captured. We also think that empirical work geared towards understanding which kind of product attributes affect utility versus choice costs (e.g., does advertising increase the perceived utility of an item versus changing the cost of choosing it) could help shed useful insights into the structure of consumer choice.

References

Allen, R. and J. Rehbeck (2019): "Revealed stochastic choice with attributes," Working paper.

Anderson, S. P., A. De Palma, and J.-F. Thisse (1988): "A representative consumer theory of the logit model," *International Economic Review*, 461–466.

Armstrong, M. and J. Vickers (2015): "Which demand systems can be generated by discrete choice?" *Journal of Economic Theory*, 158, 293–307.

Bagwell, K. (2007): "The economic analysis of advertising," in *Handbook of Industrial Organization*, ed. by M. Armstrong and R. H. Porter, Elsevier, vol. 3, 1701–1844.

Bagwell, K. and G. Ramey (1994): "Coordination economies, advertising, and search behavior in retail markets," *The American Economic Review*, 84, 498–517.

Becker, G. S. and K. M. Murphy (1993): "A simple theory of advertising as a good or bad," *The Quarterly Journal of Economics*, 108, 941–964.

Benkard, C. L., P. Bajari, et al. (2001): "Discrete Choice Models as Structural Models of Demand: Some Economic Implications of Common Approaches," Tech. rep.

Berry, S. and A. Pakes (2007): "The pure characteristics demand model," *International Economic Review*, 48, 1193–1225.

Berry, S. T. (1994): "Estimating discrete-choice models of product differentiation," *The RAND Journal of Economics*, 242–262.

Butters, G. R. (1977): "Equilibrium Distributions of Sales and Advertising Prices," *The Review of Economic Studies*, 44, 465–491.

Cerreia-Vioglio, S., D. Dillenberger, P. Ortoleva, and G. Riella (2019): "Deliberately stochastic," *American Economic Review*, 109, 2425–45.

Choné, P. and L. Linnemer (2020): "Linear demand systems for differentiated goods: Overview and user's guide," *International Journal of Industrial Organization*, 102663.

Clark, S. A. (1990): "A concept of stochastic transitivity for the random utility model," *Journal of Mathematical Psychology*, 34, 95–108.

Dixit, A. K. and J. E. Stiglitz (1977): "Monopolistic competition and optimum product diversity," *The American Economic Review*, 67, 297–308.

Falmagne, J.-C. (1978): "A representation theorem for finite random scale systems," *Journal of Mathematical Psychology*, 18, 52–72.

Fechner, G. T. (1860): *Elemente der psychophysik*, vol. 2, Breitkopf u. Härtel.

Feldman, P. and J. Rehbeck (2018): "Revealing a preference for mixtures: An experimental study of risk," Working paper.

Fosgerau, M., E. Melo, A. De Palma, and M. Shum (2020): "Discrete choice and rational inattention: A general equivalence result," *International Economic Review*, 61, 1569–1589.

Fudenberg, D., R. Iijima, and T. Strzalecki (2014): "Stochastic Choice and Revealed Perturbed Utility," Working paper.

——— (2015): "Stochastic Choice and Revealed Perturbed Utility," *Econometrica*, 83, 2371–2409.

Grossman, G. M. and C. Shapiro (1984): "Informative advertising with differentiated products," *The Review of Economic Studies*, 51, 63–81.

He, J. and P. Natenzon (2019): "Moderate expected utility," *Available at SSRN 3243657*.

Hotelling, H. (1929): "Stability in competition," *The Economic Journal*, 39, 41–57.

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, 90–98.

Jaffe, S. and S. D. Kominers (2012): "Discrete choice cannot generate demand that is additively separable in own price," *Economics Letters*, 116, 129–132.

Jaffe, S. and E. G. Weyl (2010): "Linear demand systems are inconsistent with discrete choice," *The BE Journal of Theoretical Economics*.

Luce, R. D. (1959): *Individual choice behavior*, Wiley, New York.

Machina, M. J. (1985): "Stochastic choice functions generated from deterministic preferences over lotteries," *The Economic Journal*, 95, 575–594.

Mangasarian, O. L. (1994): *Nonlinear programming*, SIAM.

McFadden, D. (1974): "Conditional logit analysis of qualitative choice behavior," in *Frontiers in Econometrics*, ed. by P. Zarembka, Academic Press, 105–142.

——— (1981): "Econometric models of probabilistic choice," in *Structural analysis of discrete data with econometric applications*, ed. by C. Manski and D. McFadden, MIT Press, vol. 198272.

Robert, J. and D. O. Stahl (1993): "Informative price advertising in a sequential search model," *Econometrica*, 61, 657–686.

Shubik, M. and R. Levitan (1980): "Market structure and innovation," *New York: John Wileyand Sons*.

Singh, N. and X. Vives (1984): "Price and quantity competition in a differentiated duopoly," *The Rand journal of economics*, 546–554.

Spence, M. (1976): "Product selection, fixed costs, and monopolistic competition," *The Review of Economic Studies*, 43, 217–235.

Sutton, J. (1991): *Sunk costs and market structure: Price competition, advertising, and the evolution of concentration*, MIT press.

——— (2007): "Market structure: theory and evidence," *Handbook of industrial organization*, 3, 2301–2368.

Thurstone, L. L. (1927): "A law of comparative judgment." *Psychological review*, 34, 273.

Turansick, C. (2021): "Identification in the random utility model," Working paper.

Appendix A. Proofs

A.1. **Proof of Theorem 1.**

Proof. Observe that for any S,

$$\underset{p\in\Delta(S)}{\text{argmax}}\sum_{x\in S}\Big(\rho(x)(au(x)+b)-\frac{ac(x)}{2}\rho(x)^2\Big)=\underset{p\in\Delta(S)}{\text{argmax}}\,a[\sum_{x\in S}\Big(\rho(x)u(x)-\frac{c(x)}{2}\rho(x)^2\Big)]+b,$$

the result follows as strictly monotone transformations of objective functions preserve maxima.

To show the second claim, we use the (w,m) representation instead of (u,c). Take two different representations of ρ: (u,c) and (u',c'). Given any representation (u,c), let $a_{(u,c)}=\sum_{y\in X}\frac{1}{c(y)}$ and $b_{(u,c)}=\sum_{y\in X}\frac{u(y)}{c(y)}$. Given (u,c) represents ρ, $(a_{(u,c)}u+b_{(u,c)},a_{(u,c)}c)$ also represents ρ, where $w_{(u,c)}(x)=\frac{a_{(u,c)}u(x)+b_{(u,c)}}{a_{(u,c)}c(x)}$ and $m_{(u,c)}(x)=\frac{1}{a_{(u,c)}c(x)}$ and $w_{(u,c)}(X)=\sum_{y\in X}\frac{a_{(u,c)}u(y)+b_{(u,c)}}{a_{(u,c)}c(y)}=1$ and $m_{(u,c)}(X)=\sum_{y\in X}\frac{1}{a_{(u,c)}c(y)}=1$. Hence $a_{(u,c)}c(x)>1$ for all x. After these normalizations, we have two corresponding representations (w,m) and (w',m') such that $w(X)=m(X)=w'(X)=m'(X)=1$. Since they both represent the same choice behaviour, we must have $w(x)=w'(x)$ and $m(x)=m'(x)$ for all x. This implies that $c(x)=\frac{a_{(u',c')}}{a_{(u,c)}}c'(x)$ and $u(x)=\frac{a_{(u,c)}}{a_{(u,c)}}u'(x)+\frac{b_{(u',c')}-b_{(u,c)}}{a_{(u,c)}}$. By letting $a=\frac{a_{(u',c')}}{a_{(u,c)}}$ and $b=\frac{b_{(u',c')}-b_{(u,c)}}{a_{(u,c)}}$, we get the desired result. □

A.2. **Proof of Theorem 2.**

Proof. Let $\mathcal{D}$ be the domain of stochastic choice functions containing all menus with size 2 and 3. Necessity of the axioms is straightforward.

For sufficiency, select any $y^*\in X$ and define binary sets $B_z:=\{z,y^*\}$ and ternary sets $T_{zz'}:=\{z,z',y^*\}$, which belong to our domain. We use these sets to define

$$m(x):=\frac{d(x|B_x,T_{xz})}{d(y^*|B_x,T_{xz})}=\frac{d(x|\{x,y^*\},\{x,y^*,z\})}{d(y^*|\{x,y^*\},\{x,y^*,z\})}$$

for some z different from x and y^*. First note that Axiom 2 implies that both the denominator and the numerator are strictly positive. Hence, the ratio is well-defined and positive. Axiom 3 implies that $d(x|S,T)d(y|S,T')=d(y|S,T)d(x|S,T')$ for all T and T' such that $x,y\in S\cap T\cap T'$, hence the choice of z does not alter the ratio. Therefore, these observations guarantee that m is well-defined and strictly positive for any x. Notice that $m(y^*)=1$. Since m is non-zero, we can define $c(x):=\frac{1}{m(x)}$.

Claim 1. *For all $S \neq T \in \mathcal{D}$ and $x, y \in S \cap T$, $c(x)d(x|S,T) = c(y)d(y|S,T)$.*

Proof. Consider distinct $S, T \in \mathcal{D}$ and $x, y \in S \cap T$ distinct from y^*. By Axiom 3, $d(x|S,T)d(y|B_y, T_{yz})d(y^*|B_x, T_{xy}) = d(y|S,T)d(y^*|B_y, T_{yz})d(x|B_x, T_{xy})$. Hence, by Axiom 2, we get the desired result: $c(x)d(x|S,T) = c(y)d(y|S,T)$. □

We now recursively define λ for every element in $\mathcal{D}$ by the following formula

$$\Lambda(S) := c(x)d(x|T,S) + \Lambda(T)$$

Given S, to define $\Lambda(S)$, we first must find a set T such that $S \cap T \neq \emptyset$. Hence, in the first step, we define λ for all S with y^* as a member. Denote this set by $\mathcal{A}_0 := \{S \in \mathcal{D} \mid y^* \in S \neq \emptyset\}$. Then $S \in \mathcal{A}_0$, $\Lambda(S)$ is defined as $1 - \rho(y^*|S)$. In the second step, we define λ for the rest of the subsets, denoted by $\mathcal{A}_1 := \{S \in \mathcal{D} \mid y^* \notin S \neq \emptyset\}$. Take $S \in \mathcal{A}_1$ and find $T \in \mathcal{A}_0$ such that $S \cap T \neq \emptyset$, and define $\Lambda(S) := c(x)d(x|T,S) + \Lambda(T)$ for some x in $S \cap T$.

Claim 2. *λ is well-defined and $\Lambda(S) + c(x)\rho(x|S) = \Lambda(T) + c(x)\rho(x|T)$ for all $x \in S \cap T$.*

Proof. We need to show that λ is well-defined (i.e., $\Lambda(S)$ is independent of the choice of x and T). By Claim 1, $\Lambda(S)$ is independent of the choice of x for a given T. Now we establish independence of T. Take two overlapping sets S and T. If $S, T \in \mathcal{A}_0$, by definition we get $\Lambda(S) = c(y^*)d(y^*|T,S) + \Lambda(T)$ (note that $c(y^*) = 1$). By Claim 1,

$$\Lambda(S) = c(x)d(x|T,S) + \Lambda(T) \text{ for all } x \in S \cap T \tag{7}$$

This establishes that the equation holds on $\mathcal{A}_0$. Now take $S \in \mathcal{A}_1$ and assume that $T, T' \in \mathcal{A}_0$ such that $x \in S \cap T, x' \in S \cap T'$. Then

$$\begin{aligned}
\Lambda(S) &= c(x)d(x|T,S) + \Lambda(T) \\
&= c(x)d(x|T,S) + c(x)d(x|\{x,x',y^*\},T) + \lambda(\{x,x',y^*\}) \text{ by Eq (7)} \\
&= c(x)d(x|\{x,x',y^*\},S) + \lambda(\{x,x',y^*\}) \\
&= c(x')d(x'|\{x,x',y^*\},S) + \lambda(\{x,x',y^*\}) \text{ by Claim 1} \\
&= c(x')d(x'|T',S) + c(x')d(x'|\{x,x',y^*\},T') + \lambda(\{x,x',y^*\}) \\
&= c(x')d(x'|T',S) + \lambda(T') \text{ by Eq (7)}
\end{aligned}$$

This establishes that λ is well-defined on $\mathcal{A}_0 \cup \mathcal{A}_1$. Finally, we must show that $\Lambda(S) = c(x)d(x|T,S) + \Lambda(T)$ holds for $S, T \in \mathcal{A}_1$. Let x be an alternative in $S \cap T$. Since $B_x \in \mathcal{A}_0$, we have $\Lambda(S) = c(x)d(x|B_x,S) + \lambda(B_x)$ and $\Lambda(T) = c(x)d(x|B_x,T) + \lambda(B_x)$. Subtracting these two equations yields $\Lambda(S) - \Lambda(T) = c(x)d(x|B_x,S) - c(x)d(x|B_x,T) = c(x)d(x|T,S)$. □

Now define $u(x) := c(x)\rho(x|S) + \Lambda(S)$. u is well-defined by Claim 2. Note that $\Lambda(S) < u(x)$ for all $x \in S \in \mathcal{D}$ since $c(x)\rho(x|S) > 0$. The (u, c) representation directly follows from this definition. Since (u, c) representation is equivalent to a WL representation with $(u, \frac{1}{c})$, we have the desired form. □

A.3. **Proof of Proposition 1.**

Proof. Recall that

$$\rho(x|S) = w(x) + (1 - w(S))\frac{m(x)}{m(S)}.$$

Let us write the BM polynomial for this model: $\sum_{S':S\subseteq S'}(-1)^{|S'\setminus S|}\rho(x|S)$. Non-negativity is obvious for $S = X$. Assume $S \neq X$, then the BM polynomial for (x, S) is

$$\sum_{S':S\subseteq S'}(-1)^{|S'\setminus S|}\frac{w(X\setminus S')}{m(S')}m(x).$$

Now let us rewrite this, indexing by elements not in S. Then we get

$$\sum_{z\notin S} w(z)\left(\sum_{S\subseteq S'\subseteq X\setminus\{z\}}\frac{m(x)}{m(S')}(-1)^{|S'\setminus S|}\right)$$

Observe the term in parentheses for a given z:

$$\sum_{S\subseteq S'\subseteq X\setminus\{z\}}\frac{m(x)}{m(S')}(-1)^{|S'\setminus S|}$$

Observe that this expression is exactly the BM-polynomial for (x, S), given a Luce rule defined on $X\setminus\{z\}$ with weights m. Therefore, this expression is non-negative given that Luce has RUM representation. Since $w(z) > 0$ for all $z \in X$, we obtain

$$\sum_{S':S\subseteq S'}\frac{w(X\setminus S')}{m(S')}m(x)(-1)^{|S'\setminus S|} \geq 0$$

Thus the quadratic rule is RUM. □

A.4. **Proof of Proposition 2.**

Proof. We prove this claim by contradiction. Assume $\rho(x|S) \geq \rho(y|S) > 0$ and $\rho(x|S \cup T) < \rho(y|S \cup T)$. Then we have

$$\frac{u(x) - \Lambda(S)}{u(y) - \Lambda(S)} \geq \frac{m(y)}{m(x)} > \frac{u(x) - \Lambda(S \cup T)}{u(y) - \Lambda(S \cup T)}$$

Then

$$u(y)(\Lambda(S \cup T) - \Lambda(S)) > u(x)(\Lambda(S \cup T) - \Lambda(S))$$

Since Λ is increasing, then $\Lambda(S \cup T) - \Lambda(S) > 0$, which implies that $u(y) > u(x)$, a contradiction. □

A.5. **Proof of Proposition 3.**

Proof. Since $\rho(x|S) > 0$, we have $u(x) > \Lambda(S)$. Since $\Lambda(T_n)$ approaches to $u(x)$ from below, as n goes to infinity, for all $y \in S$, $\rho(y|T_n) = \frac{u(y)-\Lambda(T_n)}{m(y)} > 0$ if and only if $u(y) > u(x)$. □

A.6. **Proof of Proposition 5.**

Proof. We know that within a choice set it must be the case that: $\rho(i)c_i - u_i = -\Lambda(S) = \rho(j)c_j - u_j$ or $\rho(i)c_i - u_i = \rho(j)c_j - u_j$. This means $\rho(i)\alpha a_i - \beta a_i = \rho(j)\alpha a_j - \beta a_j$ or $\alpha[\rho(i)a_i - \rho(j)a_j] = \beta[a_i - a_j]$. Denote $PA(i,j) = \rho(i)a_i - \rho(j)a_j$ and $A(i,j)$ as $a_i - a_j$. Suppose for all pairs $PA(i,j)$ and $A(i,j)$ are linearly independent. With at least $2m$ pairs the model is then identified. □

Appendix B. Allowing Zero Probabilities

Our characterization provided by Theorem 2 is based on the positivity assumption. Since our general model allows for zero probability choice, we show how to extend our characterization to allow for zero probability choice.

Our goal is to establish a method whereby, given u, c, and $S \subseteq X$, we can actually recover alternatives chosen with positive probability. This is not as simple as in the full-support case. The reason is that Λ must be defined over positively chosen alternatives. Before we illustrate how we can do this, we prove a useful lemma.

Lemma 1. *$\Lambda(S) \gtreqqless u(x)$ if and only if $\Lambda(S \setminus x) \gtreqqless \Lambda(S)$.*

Proof.

$$\begin{aligned}
\frac{\sum\limits_{y\in S}\frac{u(y)}{c(y)}-1}{\sum\limits_{y\in S}\frac{1}{c(y)}} &\gtreqless u(x)\\
\sum_{y\in S}\frac{u(y)}{c(y)}-1 &\gtreqless u(x)\sum_{y\in S}\frac{1}{c(y)}\\
\frac{1}{c(x)}-\frac{1}{c(x)}\sum_{y\in S}\frac{u(y)}{c(y)} &\lesseqgtr -\frac{u(x)}{c(x)}\sum_{y\in S}\frac{1}{c(y)}\\
(\sum_{y\in S}\frac{u(y)}{c(y)}-1)(\sum_{y\in S}\frac{1}{c(y)})+\frac{1}{c(x)}-\frac{1}{c(x)}\sum_{y\in S}\frac{u(y)}{c(y)} &\lesseqgtr -\frac{u(x)}{c(x)}\sum_{y\in S}\frac{1}{c(y)}+(\sum_{y\in S}\frac{u(y)}{c(y)}-1)(\sum_{y\in S}\frac{1}{c(y)})\\
(\sum_{y\in S}\frac{u(y)}{c(y)}-1)(\sum_{y\in S}\frac{1}{c(y)}-\frac{1}{c(x)}) &\lesseqgtr (\sum_{y\in S}\frac{u(y)}{c(y)}-\frac{u(x)}{c(x)}-1)(\sum_{y\in S}\frac{1}{c(y)})\\
(\sum_{y\in S}\frac{u(y)}{c(y)}-1)(\sum_{y\in S\setminus x}\frac{1}{c(y)}) &\lesseqgtr (\sum_{y\in S\setminus x}\frac{u(y)}{c(y)}-1)(\sum_{y\in S}\frac{1}{c(y)})\\
\Lambda(S) &\lesseqgtr \Lambda(S\setminus x)
\end{aligned}$$

□

Observe first that Λ here can be explicitly calculated by summing across the alternatives which are chosen with positive probabilities. We first denote the set of alternatives chosen with positive probability in S by $\text{supp}_\rho(S)$. That is, $\text{supp}_\rho(S) := \{x \in S|\ \rho(x|S) > 0\}$. Summing across elements of $\text{supp}_\rho(S)$, we must have $\Lambda(\text{supp}_\rho(S)) = \frac{\sum_{y\in\text{supp}_\rho(S)}\frac{u(y)}{c(y)}-1}{\sum_{y\in\text{supp}_\rho(S)}\frac{1}{c(y)}}$.

That said, the challenge is that we are not given the alternatives which are chosen with positive probabilities, but, in general, are only given the parameters u and c. What we have claimed is that if we know the positive probability elements, we can give an explicit analytic solution to the stochastic choice function. With this in mind, we now proceed to describe a simple algorithm which outputs the support for a given set of parameters u and c and budget S. The following result describes the procedure. Intuitively, what we do is for any given set S, we find a subset Q such that if elements of Q are the only ones chosen with positive probability, then given the implied $\Lambda(Q)$, for all $y \in Q$, $u(y) \geq \Lambda(Q)$, and for all $z \in S\backslash Q$ $u(z) \leq \Lambda(Q)$.

Proposition 6. *For all u, $c > 0$, and $S \subseteq X$, there is a unique $\varnothing \neq Q \subseteq S$ for which $Q = \{x \in S|\ \Lambda(Q) < u(x)\}$. Furthermore, by setting $S_1 := \arg\max_{x\in S} u(x)$ and* *defining recursively* $S_{k+1} := \{x \in S_k | \ \Lambda(S_k) < u(x)\}$*, there is a finite* K^* *for which* $Q = S_{K^*}$*, so that for all* $k \geq K^*$*,* $Q = S_k$*.*

Proof. Define $M_S \equiv \arg\max_{x \in S} u(x)$. Observe that for any $\varnothing \neq T \subseteq S$, and any $x \in M_S$, we have $\Lambda(T) < u(x)$. Initialize $S_1 := S$, and observe $M_S \subseteq S_1$. Given $M_S \subseteq S_k$, define $S_{k+1} := \{x \in S_k | \ \Lambda(S_k) < u(x)\}$. First, observe that $M_S \subseteq S_{k+1}$. Second, observe that if $x \in S_k \backslash S_{k+1}$, it follows that $u(x) \leq \Lambda(S_k)$, from which we obtain (using repeated applications of Lemma 1) that $\Lambda(S_k) \leq \Lambda(S_{k+1})$. Define $Q(S) := \bigcap_k S_k$; clearly $M_S \subseteq Q(S)$. We claim that $x \in Q(S)$ if and only if $u(x) > \Lambda(Q(S))$. By finiteness, there is K^* for which $Q(S) = S_{K^*} = S_k$ for all $k \geq K^*$. Suppose that $u(x) > \Lambda(Q(S))$; then $u(x) > \Lambda(S_{K^*})$ and so $x \in S_{K^*+1} = Q(S)$. Conversely, suppose that $x \in Q(S)$; then $x \in S_{K^*+1}$, so that $u(x) > \Lambda(S_{K^*}) = \Lambda(Q(S))$.

Next, we claim that Q is unique. Assume by means of contradiction that there exist two distinct subsets of S, say T_1 and T_2, such that

$$T_1 = \{x \in S | \ \Lambda(T_1) < u(x)\} \text{ and } T_2 = \{x \in S | \ \Lambda(T_2) < u(x)\}.$$

Without any loss of generality, assume $x \in T_1 \setminus T_2$. This implies that $\Lambda(T_1) < u(x) \leq \Lambda(T_2)$. Hence, T_2 is a proper subset of T_1. Since $u(z) > \Lambda(T_1)$ for all $z \in T_1 \setminus T_2$, by repeated applications of Lemma 1, $\Lambda(T_2) \leq \Lambda(T_1)$, a contradiction. □

Accordingly, let us define $Q(S)$ to be the unique $Q \subseteq S$ for which $Q = \{x \in S : \Lambda(Q) < u(x)\}$. As demonstrated, $Q(S)$ can be explicitly constructed from the primitives u, c and S via an iterative algorithm. Obviously, this algorithm must terminate in at most $|S| - 1$ steps. We are now ready to define our new model allowing alternatives chosen with zero probability.

Definition 2. A stochastic choice ρ function is a *weak quadratic* stochastic choice (WQSC) if there exist a utility function $u : X \to \mathbb{R}$ and a cost function $c : X \to \mathbb{R}_{++}$ such that for all $S \subset X$

$$\rho(x|S) = \begin{cases} \frac{u(x) - \Lambda(Q(S))}{c(x)} & x \in Q(S) \\ 0 & x \notin Q(S) \end{cases}$$

where $Q(A)$ is the unique subset of A satisfying $Q(A) = \{x \in A | \ \Lambda(Q(A)) < u(x)\}$ and $\Lambda(A) = \frac{\sum_{y \in A} \frac{u(y)}{c(y)} - 1}{\sum_{y \in A} \frac{1}{c(y)}}$.

Above we illustrate that ρ is a WQSC with representation (u, c) if and only if ρ is the solution to Equation (1) with functions (u, c). Hence $WQSC$ captures the full range of solutions to Equation (1).

We now provide two characterizations when zero probabilities are allowed. These characterizations differ in terms of how much they relax positivity. In the first one, we assume that positivity holds for menus with size 2 and 3. This will capture the example given at the end of Section 4. In Appendix C, we provide another characterization which entirely drops the positivity requirement. This characterization is based on linear programming duality. We provide both characterizations because, while the second is more general, the axioms used in the first characterization have a simpler behavioral interpretation, and as such, are useful for gaining intuition for the model.

For the first characterization, we modify our original axioms. The next axiom requires that the positivity holds for pairs and triples. In addition, the axiom also states that the strict regularity holds for these sets. Hence, the next axiom is a weakening of both Axiom 1 and Axiom 2.

Axiom 1*. For all binary and ternary set S, $\rho(x|S) > 0$ and $\rho(x|S) < \rho(x|S \setminus y)$.

The next axiom requires that the choice probabilities are not affected by removing alternative chosen with zero probability. If there is an alternative that is chosen with zero probability, removing it should not change the choice probabilities of the remaining items. This axiom is novel.

Axiom Z. For all $S, S \setminus x \in \mathcal{D}$, if $\rho(x|S) = 0$ and $z \in S \setminus x$ then $\rho(z|S) = \rho(z|S \setminus x)$.

The next axiom is a version of Axiom 3. There are two differences. First, without positivity, we explictly assume that some choice probabilities are positive. Second, the implication of the axiom is weaker now. The equality of Axiom 3 is replaced by an inequality. Other than these difference the intuition of the axiom stays the same: the ratio of relative levels of choice are important rather than the absolute levels.

Axiom 3*. For any list of three quadruples $((x_1, x_2, S_1, T_1), (x_2, x_3, S_2, T_2), (x_3, x_4, S_3, T_3))$ such that $x_4 = x_1$, $x_i, x_{i+1} \in S_i \cap T_i$ and $\rho(x_1|S_1), \rho(x_2|T_1) > 0$ and $\rho(x_i|A_i), \rho(x_{i+1}|A_i) > 0$ for all $i \in \{2, 3\}$ and $A_i \in \{S_i, T_i\}$,

$$d(x_1|S_1, T_1)d(x_2|S_2, T_2)d(x_3|S_3, T_3) \leq d(x_2|S_1, T_1)d(x_3|S_2, T_2)d(x_1|S_3, T_3)$$

Our first characterization in is as follows.

Theorem 3. *Suppose $\mathcal{D}$ contains all menus with size 2 and 3. Then a stochastic choice function ρ has a weak quadratic representation (u, c) on $\mathcal{D}$ such that $\Lambda(S) < min_{x \in S} u(x)$ for all $|S| \leq 3$ if and only if it satisfies Axiom 1*, 3* and Axiom Z.*

This theorem also enjoys the same uniqueness results observed in Theorem 2.

Proof. Necessity of the axioms is straightforward. We now illustrate sufficiency.

Since the domain of stochastic choice functions contains all menus with size 2 and 3 and positivity holds for these sets, we can define c the same way as we did in the proof of Theorem 2. That is,

$$c(x) := \frac{d(y^*|B_x, T_{xz})}{d(x|B_x, T_{xz})} = \frac{d(y^*|\{x, y^*\}, \{x, y^*, z\})}{d(x|\{x, y^*\}, \{x, y^*, z\})}$$

for some z different from x and y^*. We showed that c is well-defined and strictly positive for any x. Notice that $c(y^*) = 1$.

Now define the support of choice data for each set S,

$$Q(S) := \{x \in S| \ \rho(x|S) > 0\}$$

By Axiom Z, if $x \notin Q(S)$ then $Q(S \setminus x) = Q(S)$. Hence, $|Q(S)| \geq 3$ for every set S with $|S| \geq 3$ by Axiom 1*.

Claim 3. *If $\rho(x|S) > 0$ and $\rho(y|T) > 0$ then*

$$c(x)[\rho(x|S) - \rho(x|T)] \leq c(y)[\rho(y|S) - \rho(y|T)]$$

Proof. Consider distinct $S, T \in \mathcal{D}$ and $x, y \in S \cap T$. Axiom 1* implies that all choice probabilities in binary and ternary sets are different from zero. Since $\rho(x|S)$ and $\rho(y|T)$ are positive, Axiom 3* yields

$$d(x|S, T)d(y|B_y, T_{yz})d(y^*|B_x, T_{xy}) \leq d(y|S, T)d(y^*|B_y, T_{yz})d(x|B_x, T_{xy})$$

This implies $c(x)d(x|S, T) \leq c(y)d(y|S, T)$. □

Claim 4. *If $x, y \in Q(S) \cap Q(T)$ then*

$$c(x)[\rho(x|S) - \rho(x|T)] = c(y)[\rho(y|S) - \rho(y|T)]$$

Proof. Applying Claim 3 twice yields $c(x)d(x|S, T) = c(y)d(y|S, T)$. □

We now recursively define λ for every element in $\mathcal{D}$ by the following formula

$$\Lambda(S) := c(x)d(x|T, S) + \Lambda(T)$$

Given S, to define $\Lambda(S)$, we first must find a set T such that $Q(S) \cap Q(T) \neq \emptyset$. Hence, in the first step, we define λ for all S with y^* as a member and $\rho(y^*|S) > 0$. Denote this set by $\mathcal{A}_0 := \{S \in \mathcal{D} \mid y^* \in Q(S)\}$. Then for all $S \in \mathcal{A}_0$, $\Lambda(S)$ is defined as $1 - \rho(y^*|S)$. In the second step, we define λ for the set of subsets, denoted by $\mathcal{A}_1 := \{S \in \mathcal{D} \mid y^* \notin Q(S)\}$. Take $S \in \mathcal{A}_1$ and for all $T \in \mathcal{A}_0$ such that $Q(S) \cap Q(T) \neq \emptyset$, and define $\Lambda(S) := c(x)d(x|T, S) + \Lambda(T)$ for some x in $Q(S) \cap Q(T)$. Existence of such T is trivial since $B_x \in \mathcal{A}_0$ whenever $x \in Q(S)$. Since $\mathcal{A}_0 \cup \mathcal{A}_1 = \mathcal{D}$, λ is defined for the entire choice problems.

Claim 5. *λ is well-defined and $\Lambda(S) + c(x)\rho(x|S) = \Lambda(T) + c(x)\rho(x|T)$ for all $x \in Q(S) \cap Q(T)$.*

Proof. We need to show that λ is well-defined (i.e., $\Lambda(S)$ is independent of the choice of x and T). By Claim 4, $\Lambda(S)$ is independent of the choice of x for a given T. Now we establish independence of T. Take two sets S, T such that $Q(S) \cap Q(T) \neq \emptyset$. If $S, T \in \mathcal{A}_0$, by definition we get $\Lambda(S) = c(y^*)d(y^*|T, S) + \Lambda(T)$ since $\Lambda(S) = 1 - \rho(y^*|S)$, $\Lambda(T) = 1 - \rho(y^*|T)$ and $c(y^*) = 1$. By Claim 4,

$$\Lambda(S) = c(x)d(x|T, S) + \Lambda(T) \text{ for all } x \in Q(S) \cap Q(T) \tag{8}$$

This establishes that the equation holds on $\mathcal{A}_0$. Now take $S \in \mathcal{A}_1$ and assume that $T, T' \in \mathcal{A}_0$ such that $x \in Q(S) \cap Q(T)$, $x' \in Q(S) \cap Q(T')$. Such alternatives exist since $|Q(S)| \geq \min\{3, |S|\}$ for every set S. Then

$$\begin{aligned}
\Lambda(S) &= c(x)d(x|T, S) + \Lambda(T) \\
&= c(x)d(x|T, S) + c(x)d(x|T_{xx'}, T) + \lambda(T_{xx'}) \text{ by Eq (8)} \\
&= c(x)d(x|T_{xx'}, S) + \lambda(T_{xx'}) \\
&= c(x')d(x'|T_{xx'}, S) + \lambda(T_{xx'}) \text{ by Claim 4} \\
&= c(x')d(x'|T', S) + c(x')d(x'|T_{xx'}, T') + \lambda(T_{xx'}) \\
&= c(x')d(x'|T', S) + \lambda(T') \text{ by Eq (8)}
\end{aligned}$$

This establishes that λ is well-defined on $\mathcal{A}_0 \cup \mathcal{A}_1$. Finally, we must show that $\Lambda(S) = c(x)d(x|T, S) + \Lambda(T)$ holds for $S, T \in \mathcal{A}_1$. Let x be an alternative in $Q(S) \cap Q(T)$. Since $B_x \in \mathcal{A}_0$, we have $\Lambda(S) = c(x)d(x|B_x, S) + \lambda(B_x)$ and $\Lambda(T) = c(x)d(x|B_x, T) + \lambda(B_x)$. Subtracting these two equations yields $\Lambda(S) - \Lambda(T) = c(x)d(x|B_x, S) - c(x)d(x|B_x, T) = c(x)d(x|T, S)$. ◻

By Axiom Z, $Q(S) = Q(Q(S))$. In other words, the choice probabilities in S and $Q(S)$ are the same by Axiom Z. This means that $d(x|S, Q(S)) = 0$ for all x in $Q(S)$. This then gives us that $\Lambda(S) = \lambda(Q(S))$ for all S. Now define $u(x) := c(x)\rho(x|S) + \Lambda(S)$ for some S such that $x \in Q(S)$. u is well-defined by Claim 5. Note that $\Lambda(S) = \lambda(Q(S)) < u(x)$ for all $x \in Q(S)$ since $c(x)\rho(x|S) > 0$. Hence, for $x \in Q(S)$, the representation holds for those alternatives. Now assume $x \notin Q(S)$. That is, $\rho(x|S) = 0$. We need to show that $u(x) \leq \lambda(Q(S))$. Take $y \in Q(S)$. Then by definition, we have $\Lambda(S) = c(y)d(y|\{x, y\}, S) + \lambda(\{x, y\})$ and $u(x) = c(x)\rho(x|\{x, y\}) + \lambda(\{x, y\})$. Then we have

$$\Lambda(S) - u(x) = c(y)d(y|\{x,y\},S) - c(x)\rho(x|\{x,y\})$$
$$\Lambda(S) - u(x) \geq c(x)d(x|\{x,y\},S) - c(x)\rho(x|\{x,y\}) \text{ by Claim 3}$$
$$\Lambda(S) - u(x) \geq -c(x)\rho(x|S) = 0$$

Since $\Lambda(S) = \lambda(Q(S))$, the representation holds.

Finally, for all S such that $|S| \leq 3$, by Axiom 1^*, we have $Q(S) = S$ and $min_{x \in S} u(x) > \Lambda(S)$.

□

Appendix C. A Characterization without Positivity

In Appendix A, Axiom 1^* still imposes a weak version of positivity. In this appendix we do not impose any positivity requirement. The intuition behind our result is that if we allow for arbitrary zero choice probabilities, the characterization holds if and only if there exists (i) a $c(x) > 0$, (ii) a $u(x)$, (iii) a $\mu(x|S) \geq 0$ only if $\rho(x|S) = 0$, and otherwise $\mu(x|S) = 0$, and (iv) $\gamma(S)$ so that for all (x, S):

$$c(x)\rho(x|S) - u(x) - \gamma(S) - \mu(x|S) = 0. \tag{9}$$

Here, μ is the Kuhn-Tucker multiplier on the non-negativity constraint for choice probabilities and γ is the constraint on probabilities summing to one (we have one constraint for each S).

Observe that, with knowledge of $\rho(x|S)$ for each x, S, equation (9) is a linear constraint. Here, γ is obviously $-\lambda$, but writing it in positive form makes the characterization slightly easier to state and helps us clearly distinguish the case allowing for zero probabilities compared to when probabilities are non-zero.

We will use the notation $\alpha(x|S)$ to refer to a multiplier on the constraint in equation (9). The unknowns are: (i) $c(x)$ for each x, $u(x)$ for each x, (ii) $\gamma(S)$ for each S, and (iii) $\mu(x|S)$ for each (x, S) with $x \in S$. Furthermore, we have the restrictions that (i) $\mu(x|S) = 0$ if $\rho(x|S) > 0$, and otherwise, $\mu(x|S) \geq 0$, and (ii) that $c(x) > 0$. These form a system of homogeneous linear inequalities.

The following is an application of Motzkin's Theorem of the Alternative.

Theorem 4. *The stochastic choice ρ has a quadratic representation with zero probabilities if and only if for any system of numbers $\alpha(x|S) \in \Re$ for which:*

- *For every* $x \in X$, $\sum_{S:x\in S} \alpha(x|S) = 0$ *(cycle condition across sets)*
- *For every* $A \subseteq X$, $\sum_{x:x\in S} \alpha(x|S) = 0$ *(cycle condition across alternatives)*
- *If* $\rho(x|S) = 0$, *then* $\alpha(x|S) \geq 0$
- *For every* $x \in X$, $\sum_{S:x\in S} \alpha(x,S)\rho(x|S) \leq 0$

it follows that for every $x \in X$, $\sum_{S:x\in S} \alpha(x|S)\rho(x|S) = 0$.

Proof. We apply Motzkin's Theorem of the Alternative (see [Mangasarian, 1994]). Let $\alpha(x|S)$ be the multiplier on the constraint specified by equation (9), let $\beta(x|S)$ be the multiplier on the constraint on $\mu(x|S) \geq 0$ when $\rho(x|S) = 0$, and let $\eta(x)$ be the multiplier on the constraint that $c(x) > 0$. Observe that there is no quadratic representation with zeroes if and only if there is $\alpha(x|S)$ for each x, S with $x \in S$, $\beta(x|S) \geq 0$ for each x, S where $x \in S$ and $\rho(x|S) = 0$, and finally $\eta(x) \geq 0$ for each x and where there exists x^* for which $\eta(x^*) > 0$, for which:

- For every $x \in X$, $\sum_{S:x\in S} \alpha(x|S) = 0$
- For every $S \subseteq X$, $\sum_{x:x\in S} \alpha(x|S) = 0$
- For every (x, S) with $x \in S$ and $\rho(x|S) = 0$, $-\alpha(x|S) + \beta(x|S) = 0$
- For every x, $\eta(x) + \sum_{S:x\in S} \alpha(x|S)\rho(x|S) = 0$.

By eliminating the multipliers η and β,[27] we get that the preceding is equivalent to the existence of $\alpha(x|S)$ for which

- For every $x \in X$, $\sum_{S:x\in S} \alpha(x|S) = 0$
- For every $S \subseteq X$, $\sum_{x:x\in S} \alpha(x|S) = 0$
- If $\rho(x|S) = 0$, then $\alpha(x|S) \geq 0$
- For every $x \in X$, $\sum_{S:x\in S} \alpha(x|S)\rho(x|S) \leq 0$
- There exists $x^* \in X$ for which $\sum_{S:x\in S} \alpha(x|S)\rho(x|S) < 0$.

Observe that the last of these properties is exactly what is ruled out by the conclusion of the statement in Theorem 4. Consequently, the satisfaction of this system must be equivalent to a violation of the statement listed in Theorem 4. □

To see how this result relates to Theorem 2, we will show how it implies that $\frac{d(x_1|S_1,T_1)}{d(x_2|S_1,T_1)}\frac{d(x_2|S_2,T_2)}{d(x_1|S_2,T_2)} = 1$; the related condition on multiplicative cycles of triples or cycles of larger length follows similarly.

To this end, first define an auxiliary function $d(x|S,T) = \rho(x|S) - \rho(x|T)$, where $S \neq T$ and $x \in S \cap T$. Let us assume that $d(x|S,T) \neq 0$ for all relevant sets. Then

[27] For example, we note that $\eta(x)+\sum_{S:x\in S} \alpha(x|S)\rho(x|S) = 0$ implies $\sum_{S:x\in S} \alpha(x|S)\rho(x|S) = -\eta(x) \leq 0$.

take $\alpha(x_1|S_1) = 1 = -\alpha(x_1|T_1)$ and $\alpha(x_1|S_2) = -\frac{d(x_1|S_1,T_1)}{d(x_1|S_2,T_2)} = -\alpha(x_1|T_2)$, and for each set E, $\alpha(x_2|E) = -\alpha(x_1|E)$. All remaining coefficients are zero.

Observe that the constraints listed in Theorem 4 are satisfied, and in particular that $\alpha(x_1|S_1)\rho(x_1|S_1) + \alpha(x_1|S_2)\rho(x_1|S_2) + \alpha(x_1|T_1)\rho(x_1|T_1) + \alpha(x_1|T_2)\rho(x_1|T_2) = 0$.

Now, we claim that $\alpha(x_2|S_1)\rho(x_2|S_1) + \alpha(x_2|S_2)\rho(x_2|S_2) + \alpha(x_2|T_1)\rho(x_2|T_1) + \alpha(x_2|T_2)\rho(x_2|T_2) = 0$. To this end, observe that Theorem 4 implies that $\alpha(x_2|S_1)\rho(x_2|S_1) + \alpha(x_2|S_2)\rho(x_2|S_2) + \alpha(x_2|T_1)\rho(x_2|T_1) + \alpha(x_2|T_2)\rho(x_2|T_2) \geq 0$. If we had $\alpha(x_2|S_1)\rho(x_2|S_1) + \alpha(x_2|S_2)\rho(x_2|S_2) + \alpha(x_2|T_1)\rho(x_2|T_1) + \alpha(x_2|T_2)\rho(x_2|T_2) > 0$, then by choosing the system with coefficients $-\alpha$ instead of α, we would obtain a contradiction.

In particular now observe that the fact that $\alpha(x_2|S_1)\rho(x_2|S_1) + \alpha(x_2|S_2)\rho(x_2|S_2) + \alpha(x_2|T_1)\rho(x_2|T_1) + \alpha(x_2|T_2)\rho(x_2|T_2) = 0$ implies:

$$-\rho(x_2|S_1) + \frac{d(x_1|S_1,T_1)}{d(x_1|S_2,T_2)}\rho(x_2|S_2) + \rho(x_2|T_1) - \frac{d(x_1|S_1,T_1)}{d(x_1|S_2,T_2)}\rho(x_2|T_2) = 0.$$

This implies $d(x_2|S_2,T_2)\frac{d(x_1|S_1,T_1)}{d(x_1|S_2,T_2)} = d(x_2|S_1,T_1)$. Conclude $\frac{d(x_1|S_1,T_1)}{d(x_2|S_1,T_1)}\frac{d(x_2|S_2,T_2)}{d(x_1|S_2,T_2)} = 1$. □

Appendix D. Moderate Stochastic Transitivity

Proposition 7. *The WL model satisfies moderate stochastic transitivity.*

Proof. Define $U(x) \equiv \frac{2w(x)-1}{m(x)}$.

First we rewrite $\rho(x,y)$. That is,

$$\begin{aligned}
\rho(x,y) &= \frac{m(x) + m(y)w(x) - m(x)w(y)}{m(x) + m(y)} \\
&= \frac{1}{2} + \frac{1}{2}\frac{2w(x)-1}{m(x)}m(y)\frac{1 - \frac{\frac{2w(y)-1}{m(y)}}{\frac{2w(x)-1}{m(x)}}}{1 + \frac{m(y)}{m(x)}} \\
&= \frac{1}{2} + \frac{1}{2}U(x)m(y)\frac{1 - \frac{U(y)}{U(x)}}{1 + \frac{m(y)}{m(x)}} \\
&= \frac{1}{2} + \frac{1}{2}\frac{U(x) - U(y)}{\frac{1}{m(x)} + \frac{1}{m(y)}}
\end{aligned}$$

Since $\rho(x,y), \rho(y,z) \geq \frac{1}{2}$, we have $U(x) \geq U(y) \geq U(z)$.

Case I: $\rho(x,y) \geq \rho(y,z)$. This implies that

$$\frac{1}{2} + \frac{1}{2}\frac{U(y)-U(z)}{\frac{1}{m(y)}+\frac{1}{m(z)}} \leq \frac{1}{2} + \frac{1}{2}\frac{U(x)-U(y)}{\frac{1}{m(x)}+\frac{1}{m(y)}}$$

$$m(z)\frac{U(y)-U(z)}{m(y)+m(z)} \leq m(x)\frac{U(x)-U(y)}{m(x)+m(y)}$$

$$\frac{m(z)(m(x)+m(y))}{m(x)(m(y)+m(z))}\Big[U(y)-U(z)\Big] \leq U(x)-U(y)$$

The left hand side of the inequality can be rewritten as follows:

$$\begin{aligned}
\frac{m(z)(m(x)+m(y))}{m(x)(m(y)+m(z))} &= \frac{m(z)(m(x)+m(y))+m(x)m(y)-m(x)m(y)}{m(x)(m(y)+m(z))} \\
&= \frac{m(y)(m(x)+m(z))}{m(x)(m(y)+m(z))} + \frac{m(x)(m(z)-m(y))}{m(x)(m(y)+m(z))} \\
&= \frac{m(y)(m(x)+m(z))}{m(x)(m(y)+m(z))} + \frac{m(z)-m(y)}{m(y)+m(z)}
\end{aligned}$$

Note that

$$-1 \leq \frac{m(z)-m(y)}{m(y)+m(z)} \leq 1$$

Hence, we have

$$\frac{m(y)(m(x)+m(z))}{m(x)(m(y)+m(z))} \leq \frac{m(z)(m(x)+m(y))}{m(x)(m(y)+m(z))} + 1 \tag{10}$$

By the assumption, we have

$$\frac{m(z)(m(x)+m(y))}{m(x)(m(y)+m(z))}\Big[U(y)-U(z)\Big] \leq U(x)-U(y)$$

$$\frac{m(z)(m(x)+m(y))}{m(x)(m(y)+m(z))}\Big[U(y)-U(z)\Big] \leq U(x)-U(z)+U(z)-U(y)$$

$$\left(\frac{m(z)(m(x)+m(y))}{m(x)(m(y)+m(z))}+1\right)\Big[U(y)-U(z)\Big] \leq U(x)-U(z)$$

By (10), we have

$$\frac{m(y)(m(x)+m(z))}{m(x)(m(y)+m(z))}\Big[U(y)-U(z)\Big] \leq U(x)-U(z)$$

Therefore, $\rho(x,z) \geq \rho(y,z)$.

Case II: $\rho(x,y) \leq \rho(y,z)$. This implies that

$$U(y) - U(z) \geq \Big[U(x) - U(y)\Big]\frac{m(x)(m(y)+m(z))}{m(z)(m(x)+m(y))}$$

$$U(x) - U(z) \geq \Big[U(x) - U(y)\Big]\left(1 + \frac{m(x)(m(y)+m(z))}{m(z)(m(x)+m(y))}\right)$$

$$\frac{m(x)(m(y)+m(z))}{m(z)(m(x)+m(y))} \quad = \quad \frac{m(y)(m(x)+m(z))}{m(z)(m(x)+m(y))} + \frac{m(x)-m(y)}{m(x)+m(y)}$$

Note that

$$-1 \leq \frac{m(x)-m(y)}{m(x)+m(y)} \leq 1$$

Hence, we have

$$\frac{m(y)(m(x)+m(z))}{m(z)(m(x)+m(y))} \leq \frac{m(x)(m(y)+m(z))}{m(z)(m(x)+m(y))} + 1 \tag{11}$$

By (11), we have

$$U(x) - U(z) \geq \Big[U(x) - U(y)\Big]\frac{m(y)(m(x)+m(z))}{m(z)(m(x)+m(y))}$$

Therefore, $\rho(x,z) \geq \rho(x,y)$. □