6-2017

Technical note—On the relation between several discrete choice models

Guiyun FENG
Singapore Management University, gyfeng@smu.edu.sg

Xiaobo LI

Zizhuo WANG

http://pubsonline.informs.org/journal/opre/

OPERATIONS RESEARCH
Vol. 65, No. 6, November–December 2017, pp. 1516–1525
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Technical Note—On the Relation Between Several Discrete Choice Models

Guiyun Feng,[a] Xiaobo Li,[a] Zizhuo Wang[a]

[a] Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, Minnesota 55455
Contact: fengx421@umn.edu (GF); lixx3195@umn.edu (XL); zwang@umn.edu (ZW)

Received: March 5, 2015
Revised: February 1, 2016; October 7, 2016
Accepted: December 17, 2016
Published Online in Articles in Advance: June 6, 2017

Subject Classifications: marketing: buyer behavior, choice models; utility/preference
Area of Review: Operations and Supply Chains

https://doi.org/10.1287/opre.2017.1602

Abstract. In this paper, we study the relationship between several well known classes of discrete choice models, i.e., the random utility model (RUM), the representative agent model (RAM), and the semiparametric choice model (SCM). Using a welfare-based model as an intermediate, we show that the RAM and the SCM are equivalent. Furthermore, we show that both models as well as the welfare-based model strictly subsume the RUM when there are three or more alternatives, while the four are equivalent when there are only two alternatives. Thus, this paper presents a complete picture of the relationship between these choice models.

Funding: The research of the third author is supported by the National Science Foundation [Grant CMMI-1462676].

Keywords: welfare function • random utility model • representative agent model • semiparametric choice model

1. Introduction

In this paper, we study the discrete choice models. Discrete choice models are used to model choices made by people among a finite set of alternatives. As examples, this includes examining which product to purchase for a consumer and which mode of transportation to take for a passenger. In the past few decades, discrete choice models have attracted great interest in the economics, marketing, operations research, and management science communities. Specifically, such models have been viewed as the behavioral foundation in many operational decision-making problems, such as transportation planning, assortment optimization, multiproduct pricing, etc.

In the past few decades, researchers have proposed a variety of discrete choice models. Among them, the most popular is the random utility model , in which a utility is assigned to each alternative. In the random utility model, the utility is composed of a deterministic part and a random part. Each individual then chooses the alternative with the highest utility, given the realization of the random part. Different choice models arise when different distributions for the random part are used. Some examples of the random utility model can be found in McFadden (1974, 1980) and Daganzo (1980). Another popular choice model is the representative agent model, in which a representative agent makes the choice on behalf of the population. In the representative agent model, there is again a utility associated with each alternative, and the representative agent maximizes a weighted utility of the choice (which is a vector of proportions for each alternative) plus a regularization term, which typically encourages diversification of the choice (Anderson et al. 1988). More recently, a class of semiparametric choice models has been proposed (Natarajan et al. 2009). This model is similar to the random utility model. However, instead of specifying a single distribution for the random utility, a set of distributions is considered. Then an extreme distribution in that set is chosen to determine the choice probabilities. There are other choice models based on the dynamics of choice decisions or other non-parametric ideas. We provide a more detailed review of these models in Section 2.

Although these models (the random utility model, the representative agent model and the semiparametric choice model) have all provided excellent theoretical and empirical explanations for how people make choices in practice, a central question remains: What is the exact relationship between these choice models? In this paper, we present a complete answer to this question. To do so, we view from another perspective of choice models and consider a *welfare-based approach*. This approach is based on the observation that many existing choice models take the form of mapping a utility vector to a probability vector and admit a welfare function of the utilities whose gradient gives the choice probability vector. By summarizing properties that are satisfied by welfare functions of existing choice models, we define the class of *welfare-based choice models*.

Using the welfare-based choice model as an intermediate model, we show that the representative agent model and the semiparametric model are the same. More precisely, under mild regularity assumptions,

given either a regularization function (which defines a representative agent model) or a distribution set (which defines a semiparametric model), one can construct the other to define exactly the same choice model. This is somewhat surprising because they seem to have very different origins. In addition, our proof of the equivalence of these three models is constructive, therefore, it gives explicit methods to convert one model to another, potentially alleviating the need to construct correspondence in a case by case manner as is done in the literature.

Furthermore, we study the relationship between the above three models and the random utility model. We show that when there are only two alternatives, the random utility model is equivalent to the above three models. We also demonstrate that this is not true in general if there are three or more alternatives, in which case the above three models strictly subsume the random utility model. This is an improvement on the current known result.

Notations. Throughout the paper, the following notations will be used. We use notation $\mathcal{R}$ to denote the set of real numbers, and $\bar{\mathcal{R}} = \mathcal{R} \cup \{-\infty, +\infty\}$ to denote the set of extended real numbers. We use $\mathbf{e}$ to denote a vector of all ones, $\mathbf{e}_i$ to denote a vector of zeros except 1 at the ith entry, and $\mathbf{0}$ to denote a vector of all zeros (the dimension of these vectors will be clear from the context). Also, we write $\mathbf{x} \geqslant \mathbf{y}$ to denote a component-wise relationship and Δ_{n-1} to denote the $(n-1)$-dimensional simplex, i.e., $\Delta_{n-1} = \{\mathbf{x} \mid \mathbf{e}^T\mathbf{x} = 1, \mathbf{x} \geqslant \mathbf{0}\}$. In our discussions, ordinary lowercase letters $x, y, \ldots$ denote scalars and boldfaced lowercase letters $\mathbf{x}, \mathbf{y}, \ldots$ denote vectors.

2. Review of Existing Discrete Choice Models

In this section, we review several prevailing classes of discrete choice models that are related to the discussion in this paper.

2.1. Random Utility Model

Perhaps the most popular class of discrete choice model is the random utility model (RUM), first proposed by Thurstone (1927) and later studied in a vast literature in economics (see Anderson et al. 1992 for a comprehensive review). In such a model, a random utility is assigned to each of the alternatives, and an individual will pick the alternative with the highest realized utility. Here, the randomness in utilities could be due to the lack of information on the alternatives for a particular individual or to the idiosyncrasies of preferences among a population. As the output, the RUM predicts a vector of choice probabilities among the alternatives, rather than a single deterministic choice. Mathematically, suppose there are n alternatives denoted by $\mathcal{N} = \{1, 2, \ldots, n\}$, then the RUM assumes that the utility of alternative i takes the following form:

$$u_i = \pi_i + \epsilon_i, \quad \forall\, i \in \mathcal{N}, \tag{1}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ is the deterministic part of the utility and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ is the random part. In the RUM, it is assumed that the joint distribution θ of $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ is known. Then the probability that alternative i will be chosen is (to ensure that the following equation is well defined, we assume θ is absolutely continuous, an assumption we make for all the RUMs we discuss later):

$$q_i(\boldsymbol{\pi}) = \mathbb{P}_{\boldsymbol{\epsilon}\sim\theta}\left(i = \arg\max_{k\in\mathcal{N}}(\pi_k + \epsilon_k)\right). \tag{2}$$

RUMs can be further classified by the distribution function of the random components. The most widely used RUM is the multinomial logit (MNL) model, first proposed by McFadden (1974). The MNL model is derived by assuming that $(\epsilon_1, \ldots, \epsilon_n)$ follow independent and identically distributed Gumbel distributions with scale parameter η. Given this assumption, the choice probability in (2) can be further written as follows:

$$q_i^{\text{mnl}}(\boldsymbol{\pi}) = \frac{\exp(\pi_i/\eta)}{\sum_{k\in\mathcal{N}}\exp(\pi_k/\eta)}.$$

The expected utility for individual under the MNL model is:

$$w^{\text{mnl}}(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\epsilon}\sim\theta}\left[\max_{i\in\mathcal{N}} \pi_i + \epsilon_i\right] = \eta \log\left(\sum_{i\in\mathcal{N}} \exp(\pi_i/\eta)\right).$$

The existence of a closed-form formula for the MNL model makes it a very popular choice model. See Ben-Akiva and Lerman (1985), Anderson et al. (1992), and Train (2009) for more discussions on the properties of the MNL model. In addition to the MNL, there are other choices for the random part in (1) that lead to alternative choice models. Some popular choices among them are the probit model (in which $\boldsymbol{\epsilon}$ is chosen to be a joint normal distribution, see, e.g., Daganzo 1980), the nested logit model (in which $\boldsymbol{\epsilon}$ is chosen to be correlated general extreme value distributions, see, e.g., McFadden 1980), and the exponomial choice model (in which $\boldsymbol{\epsilon}$ is chosen to be negative exponential distributions, see Alptekinoğlu and Semple 2016).

2.2. Representative Agent Model

Another popular way to model choices is to use a representative agent model (RAM). In such a model, a representative agent makes a choice among n alternatives on behalf of the entire population. In particular, this agent may choose any fractional amount of each alternative, or equivalently, his/her choice is a vector $\mathbf{x} = (x_1, \ldots, x_n)$ on Δ_{n-1}. To make his/her choice, the

agent takes into account the expected utility while preferring some degree of diversification. More precisely, the representative agent solves an optimization problem as follows:

$$\max_{\mathbf{x}\in\Delta_{n-1}}(\boldsymbol{\pi}^T\mathbf{x}-V(\mathbf{x})). \tag{3}$$

Here $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_n)$ is the deterministic utility of each alternative, which is similar to that in the RUM. $V(\mathbf{x})\colon \Delta_{n-1}\mapsto\mathcal{R}$ is a regularization term such that $-V(x)$ rewards diversification. We denote the optimal value of (3) by $w^r(\boldsymbol{\pi})$, which is the utility a representative agent can obtain if the deterministic utility vector is $\boldsymbol{\pi}$. In this paper, without loss of generality, we assume that $V(\mathbf{x})$ is convex and lower semicontinuous.[1] Moreover, if for any $\boldsymbol{\pi}$, there is a unique solution to (3), then we define

$$\mathbf{q}^r(\boldsymbol{\pi})=\arg\max\{\boldsymbol{\pi}^T\mathbf{x}-V(\mathbf{x})\,|\,\mathbf{x}\in\Delta_{n-1}\} \tag{4}$$

to be the choice probability vector given by the RAM.

A recognized close connection exists between the RUM and the RAM. In Anderson et al. (1988), the authors show that the choice probabilities from an MNL model with parameter η can be equally derived from a RAM with $V(\mathbf{x})=\eta\sum_{i=1}^n x_i\log x_i$. Equivalently, we can write

$$\mathbf{q}^{\text{mnl}}(\boldsymbol{\pi})=\arg\max\left\{\boldsymbol{\pi}^T\mathbf{x}-\eta\sum_{i=1}^n x_i\log x_i\,\middle|\,\mathbf{x}\in\Delta_{n-1}\right\}.$$

Hofbauer and Sandholm (2002) further extend the result to general RUMs. They show that for any RUM with continuously distributed random utility, there exists a RAM that gives the same choice probability. The precise statement of their result is as follows:

Proposition 1. *Let* $\mathbf{q}(\boldsymbol{\pi})\colon \mathcal{R}^n\mapsto\Delta_{n-1}$ *be the choice probability function defined in* (2) *where the random vector* $\boldsymbol{\epsilon}$ *admits a strictly positive density on* $\mathcal{R}^n$ *and the function* $\mathbf{q}(\boldsymbol{\pi})$ *is continuously differentiable. Then for all* $\boldsymbol{\pi}$ *there exists* $V(\cdot)$ *such that*:

$$\mathbf{q}(\boldsymbol{\pi})=\arg\max\{\boldsymbol{\pi}^T\mathbf{x}-V(\mathbf{x})\,|\,\mathbf{x}\in\Delta_{n-1}\}.$$

They also show that the reverse statement of Proposition 1 is not true:

Proposition 2 (Proposition 2.2 in Hofbauer and Sandholm 2002). *When* $n\geqslant 4$, *there does not exist a RUM that is equivalent to the RAM with* $V(\mathbf{x})=-\sum_{i=1}^n\log x_i$.

Based on the two propositions above, we know that the RAM strictly subsumes the RUM as a special case.

2.3. Semparametric Choice Model

Recently, a new class of choice models, called the semiparametric choice model (SCM), was proposed by Natarajan et al. (2009). Unlike the RUM where a certain distribution of the random utility $\boldsymbol{\epsilon}$ is specified, in the SCM, one considers a set of distributions Θ for $\boldsymbol{\epsilon}$. Given the deterministic utility vector $\boldsymbol{\pi}$, one defines the maximum expected utility function $w^s(\boldsymbol{\pi})$ as follows:

$$w^s(\boldsymbol{\pi})=\sup_{\theta\in\Theta}\mathbb{E}_{\boldsymbol{\epsilon}\sim\theta}\left[\max_{i\in\mathcal{N}}\pi_i+\epsilon_i\right]. \tag{5}$$

Note that in the RUM, the maximum expected utility function can be similarly defined, but only with a single distribution θ. Thus the SCM can be viewed as an extension of the RUM. Let $\theta^*(\boldsymbol{\pi})$ denote the distribution (or a limit of a sequence of distributions) that attains the optimal value in (5). The choice probability for alternative i under this model is given by (provided that it is well defined):

$$q_i^s(\boldsymbol{\pi})=\mathbb{P}_{\theta^*(\boldsymbol{\pi})}\left(i=\arg\max_{k\in\mathcal{N}}(\pi_k+\epsilon_k)\right). \tag{6}$$

Several special cases of SCMs have been studied recently. One such model, called the marginal distribution model (MDM), is proposed by Natarajan et al. (2009). In the MDM, the distribution set Θ contains all the distributions with certain marginal distributions. The following proposition proved in Natarajan et al. (2009) shows that the marginal distribution model can be equivalently represented by a RAM:

Proposition 3. *Suppose* $\Theta=\{\theta\,|\,\epsilon_i\sim F_i(\cdot),\forall\, i\}$ *where* $F_i(\cdot)$*s are given continuous distributions. Then we have*:

$$w^s(\boldsymbol{\pi})=\max_{\mathbf{x}}\left\{\boldsymbol{\pi}^T\mathbf{x}+\sum_{i=1}^n\int_{1-x_i}^1 F_i^{-1}(t)\,dt\,\middle|\,\mathbf{x}\in\Delta_{n-1}\right\}. \tag{7}$$

Furthermore, the choice probabilities $\mathbf{q}^s(\boldsymbol{\pi})$ *can be obtained as the optimal solution* $\mathbf{x}^*$ *in* (7).

Another semiparametric model is the marginal moment model (MMM), in which only the first and second moments of the marginal distributions are known and Θ comprises all distributions that are consistent with the marginal moments. Natarajan et al. (2009) show that the MMM can also be represented as a RAM (without loss of generality, we assume that the marginal mean of ϵ_i is 0 for all i):

Proposition 4. *Suppose the marginal standard deviation of* ϵ_i *is* σ_i *for all* i. *Then we have*

$$w^s(\boldsymbol{\pi})=\max_{\mathbf{x}}\left\{\boldsymbol{\pi}^T\mathbf{x}+\sum_{i=1}^n\sigma_i\sqrt{x_i(1-x_i)}\,\middle|\,\mathbf{x}\in\Delta_{n-1}\right\}. \tag{8}$$

Furthermore, the choice probabilities $\mathbf{q}^s(\boldsymbol{\pi})$ *can be obtained as the optimal solution* $\mathbf{x}^*$ *in* (8).

To incorporate covariance information, Mishra et al. (2012) further propose a complete moment model (CMM), in which Θ is the set of distributions with known first and second moments $\boldsymbol{\Sigma}$ (covariance matrix). Ahipasaoglu et al. (2016) show that the CMM model can also be written as a RAM (again without loss of generality, we assume the first moments are 0):

Proposition 5. *Assume* $\boldsymbol{\Sigma}\succ 0$. *Then we have*:

$$w^s(\boldsymbol{\pi})=\max_{\mathbf{x}}\left\{\boldsymbol{\pi}^T\mathbf{x}+\text{trace}(\boldsymbol{\Sigma}^{1/2}S(\mathbf{x})\boldsymbol{\Sigma}^{1/2})^{1/2}\,|\,\mathbf{x}\in\Delta_{n-1}\right\}, \tag{9}$$

where $S(\mathbf{x}) = \text{Diag}(\mathbf{x}) - \mathbf{x}\mathbf{x}^T$ *and* $\text{trace}(X)$ *is the trace of the matrix* X. *Furthermore, the choice probabilities* $\mathbf{q}^s(\boldsymbol{\pi})$ *can be obtained as the optimal solution* $\mathbf{x}^*$ *in* (9).

Thus, all semiparametric models studied so far can be represented as RAMs. In the next section, we show that this is generally the case. Moreover, we show that, in fact, the reverse is also true and thus the set of RAMs is equivalent to that of semiparametric models.

Before we end this section, note that there are other types of choice models in the literature in addition to those mentioned above, such as the Markov chain-based choice model (see Blanchet et al. 2016), the two-stage choice model (see Jagabathula and Rusmevichientong 2013), the generalized attraction model (see Gallego et al. 2014) and the non-parametric model (see Farias et al. 2013). Some of those models are also more general than the RUM model. However, they are based on different ideas. In particular, they do not take the form of mapping a utility vector to a choice probability vector. Thus we choose not to include a detailed review of those models in this paper.

3. Relations Between Choice Models

In this section, we study the relations between the various choice models reviewed in Section 2. We first notice that although the choice models reviewed in Section 2 are based on different ideas, they are all essentially functions from a vector of utilities $\boldsymbol{\pi}$ to a vector of choice probabilities $\mathbf{q}(\boldsymbol{\pi})$. Moreover, each of these models allows a welfare function $w(\boldsymbol{\pi})$ that captures the expected utility an individual can get from the choice model, and the choice probability vector can be viewed as the gradient of $w(\boldsymbol{\pi})$ with respect to $\boldsymbol{\pi}$. Our proposed approach is based on these observations. We begin with the following definition:

Definition 1 (Choice Welfare Function). Let $w(\boldsymbol{\pi})$ be a mapping from $\mathcal{R}^n$ to $\bar{\mathcal{R}}$. We call $w(\boldsymbol{\pi})$ a choice welfare function if $w(\boldsymbol{\pi})$ satisfies the following properties:

1. (Monotonicity): For any $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2 \in \mathcal{R}^n$ and $\boldsymbol{\pi}_1 \geqslant \boldsymbol{\pi}_2$, $w(\boldsymbol{\pi}_1) \geqslant w(\boldsymbol{\pi}_2)$;
2. (Translation Invariance): For any $\boldsymbol{\pi} \in \mathcal{R}^n$, $t \in \mathcal{R}$, $w(\boldsymbol{\pi} + t\mathbf{e}) = w(\boldsymbol{\pi}) + t$;
3. (Convexity): For any $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2 \in \mathcal{R}^n$ and $0 \leqslant \lambda \leqslant 1$, $\lambda w(\boldsymbol{\pi}_1) + (1-\lambda) w(\boldsymbol{\pi}_2) \geqslant w(\lambda\boldsymbol{\pi}_1 + (1-\lambda)\boldsymbol{\pi}_2)$.

In addition to the three properties, if $w(\boldsymbol{\pi})$ is also differentiable, then we call $w(\boldsymbol{\pi})$ a differentiable choice welfare function.

Here we make a few comments on the three conditions in Definition 1. The monotonicity condition is straightforward. It requires that the welfare is higher if all alternatives have higher deterministic utilities. The translation invariance property requires that if the deterministic utilities of all alternatives increase by a certain amount t, then the choice welfare function will increase by the same amount. This is reasonable given that choice is about relative preferences: Increasing the utilities of all alternatives by the same amount will not change the relative preferences but will only increase the welfare by the amount of the increment. Later, we will see that this condition is necessary to guarantee well-defined choice probabilities. The last condition of convexity basically states that the average welfare at two utility vectors is greater than the welfare at the average utility vector. If we view the welfare as the maximal utility one can get among the alternatives, then this property is equivalent to saying that the weighted optimal value of two maximization problems (of the utilities of the alternatives) is larger than the optimal value of the weighted one, which is true since the maximal operator is convex .

Next we show that a choice welfare function has two equivalent representations, i.e., a convex optimization representation and a semiparametric representation. This result will be instrumental for us to derive the relationships between choice models.

Theorem 1. *The following statements are equivalent*:

1. $w(\boldsymbol{\pi})$ *is a choice welfare function*;
2. *There exists a convex function* $V(\mathbf{x})$: $\Delta_{n-1} \mapsto \bar{\mathcal{R}}$ *such that*

$$w(\boldsymbol{\pi}) = \max\{\boldsymbol{\pi}^T\mathbf{x} - V(\mathbf{x}) \mid \mathbf{x} \in \Delta_{n-1}\}; \tag{10}$$

3. *There exists a distribution set* Θ *such that*

$$w(\boldsymbol{\pi}) = \sup_{\theta\in\Theta} \mathbb{E}_{\boldsymbol{\epsilon}\sim\theta}\left[\max_{i\in\mathcal{N}} \pi_i + \epsilon_i\right]. \tag{11}$$

The proof of Theorem 1 uses several results in convex analysis and optimization. In the following, we establish its implication to discrete choice models. In this paper, we refer to discrete choice models as the entire set of functions $\mathbf{q}(\boldsymbol{\pi})$: $\mathcal{R}^n \mapsto \Delta_{n-1}$, mapping a utility vector to a choice probability vector. We first propose the following choice model based on the choice welfare function:

Definition 2 (Welfare-Based Choice Model). Suppose $w(\boldsymbol{\pi})$ is a differentiable choice welfare function. Then the welfare-based choice model derived from $w(\boldsymbol{\pi})$ is defined by

$$\mathbf{q}(\boldsymbol{\pi}) = \nabla w(\boldsymbol{\pi}). \tag{12}$$

Note that when $w(\cdot)$ is differentiable, we have $\nabla w(\boldsymbol{\pi}) \in \Delta_{n-1}$ by the monotonicity and the translation invariance property of $w(\boldsymbol{\pi})$. Therefore $\mathbf{q}(\boldsymbol{\pi})$ defined by (12) is indeed a valid choice model. Next we show the equivalence of various choice models. We first introduce the following definitions (see Rockafellar 1974):

Definition 3 (Proper Function). A function f: $X \mapsto \bar{\mathcal{R}}$ is proper if $f(\mathbf{x}) < \infty$ for at least one $\mathbf{x} \in X$ and $f(\mathbf{x}) > -\infty$ for all $\mathbf{x} \in X$.

Definition 4 (Essentially Strictly Convex Function). A proper convex function f on $\mathcal{R}^n$ is essentially strictly convex if f is strictly convex on every convex subset of

$$\mathrm{dom}(\partial f) = \{\mathbf{x} \mid \partial f(\mathbf{x}) \neq \varnothing\},$$

where $\partial f(\mathbf{x})$ is the set of subgradients of f at $\mathbf{x}$, and $\varnothing$ is the empty set.

Note that any strictly convex function is essentially strictly convex. Next, we have the following theorem:

Theorem 2. *For a choice model* $\mathbf{q}$: $\mathcal{R}^n \mapsto \Delta_{n-1}$, *the following statements are equivalent*:

1. *There exists a differentiable choice welfare function* $w(\boldsymbol{\pi})$ *such that* $\mathbf{q}(\boldsymbol{\pi}) = \nabla w(\boldsymbol{\pi})$;
2. *There exists an essentially strictly convex function* $V(\mathbf{x})$ *such that*

$$\mathbf{q}(\boldsymbol{\pi}) = \arg\max\{\boldsymbol{\pi}^T\mathbf{x} - V(\mathbf{x}) \mid \mathbf{x} \in \Delta_{n-1}\};$$

3. *There exists a distribution set* Θ *such that*

$$\mathbf{q}(\boldsymbol{\pi}) = \nabla_{\pi}\left\{\sup_{\theta\in\Theta} \mathbb{E}_{\theta}\left[\max_{i\in\mathcal{N}} \pi_i + \epsilon_i\right]\right\}.$$

In Theorems 1 and 2, with the help of the welfare-based choice model, we establish the connection between two existing choice models, the RAM and the semiparametric model. In particular, we show that those two classes of choice models are equivalent. This result explains the prior results that for most known semiparametric models, there exists an equivalent RAM. In addition, it asserts that the reverse is also true, which is somewhat surprising. Therefore, in terms of scope, those three classes of choice models (the welfare-based choice model, the RAM and the semiparametric model) are the same. We believe this result is useful for the theoretical study of discrete choice models.

In light of the equivalence of the three classes of choice models, we could have more versatile ways to construct a choice model. In particular, we can pick any of the three representations to begin. For the welfare-based choice model, one needs to choose a choice welfare function $w(\boldsymbol{\pi})$ which satisfies the three conditions. For the RAM, one needs to choose a (strictly) convex regularization function. For the semiparametric model, one needs to choose a set of distributions. In different situations, it might be easier to use one representation than the others to capture certain properties of the choice model.

The next theorem studies one desirable property of choice models and investigates how it can be reflected to the construction of the three choice models. We start with the following definition:

Definition 5 (Superlinear Choice Welfare Function). A differentiable choice welfare function $w(\boldsymbol{\pi})$ is called *superlinear* if there exist $b_i, i = 1, \ldots, n$, such that for any $\boldsymbol{\pi} \in \mathcal{R}^n$:

$$w(\boldsymbol{\pi}) \geqslant \pi_i + b_i, \quad \forall\, i = 1, \ldots, n.$$

This property is desirable in most applications. It requires that the utility one can get from a set of alternatives is not much less than the utility of each alternative. After all, for each alternative i, one can always choose it and obtain the corresponding utility. We have the following theorem:

Theorem 3. *For a choice model* $\mathbf{q}$: $\mathcal{R}^n \mapsto \Delta_{n-1}$, *the following statements are equivalent*:

1. *There exists a superlinear differentiable choice welfare function* $w(\boldsymbol{\pi})$ *such that* $\mathbf{q}(\boldsymbol{\pi}) = \nabla w(\boldsymbol{\pi})$;
2. *There exists an essentially strictly convex function* $V(\mathbf{x})$ *that is upper bounded on* Δ_{n-1} *such that*

$$\mathbf{q}(\boldsymbol{\pi}) = \arg\max\{\boldsymbol{\pi}^T\mathbf{x} - V(\mathbf{x}) \mid \mathbf{x} \in \Delta_{n-1}\};$$

3. *There exists a distribution set* Θ *containing only distributions with finite expectation* (*i.e.*, $\mathbb{E}_{\theta}|\epsilon_i| < \infty$ *for all* i *and* $\theta \in \Theta$) *such that*

$$\mathbf{q}(\boldsymbol{\pi}) = \nabla_{\pi}\left\{\sup_{\theta\in\Theta} \mathbb{E}_{\theta}\left[\max_{i\in\mathcal{N}} \pi_i + \epsilon_i\right]\right\}.$$

Moreover, if either of the above cases holds, then $\mathbf{q}(\boldsymbol{\pi})$ *can span the whole simplex, i.e., for all* $\mathbf{x}$ *in the interior of* Δ_{n-1}, *there exists* $\boldsymbol{\pi}$ *such that* $\mathbf{q}(\boldsymbol{\pi}) = \mathbf{x}$.

Theorem 3 further develops the equivalence of choice models obtained in Theorem 2 by narrowing the discussion to welfare-based choice models with the superlinear property. In particular, we find that a superlinear differentiable choice welfare function has a semiparametric representation, of which the distribution set only contains distributions with finite expectation; in practice, this is a desirable property. The last statement that $\mathbf{q}(\boldsymbol{\pi})$ spans the whole simplex is related to the results in Hofbauer and Sandholm (2002), Norets and Takahashi (2013) and Mishra et al. (2014). These papers provide conditions under which $\mathbf{q}(\boldsymbol{\pi})$ defined from the RUM or the MDM can span the whole simplex. Theorem 3 extends these results to more general conditions.

Next, we study the relationship between the welfare-based choice model (thus also the RAM and the SCM) and the RUM. In particular, we study under what conditions a welfare-based choice model can be equivalently represented by a RUM. This study will help us understand clearly the relations between various choice models and the RUM, and thus design new choice models that do not necessarily have a random utility representation.

First, we show that when there are only two alternatives, the class of RUMs is equivalent to the class of welfare-based choice models (thus also equivalent to the RAM and the SCM).

Theorem 4. *For any differentiable choice welfare function* $w(\pi_1, \pi_2)$, *there exists a distribution* θ *of* $\{\epsilon_1, \epsilon_2\}$ *such that*:

$$w(\pi_1, \pi_2) = \mathbb{E}_{\theta}[\max\{\pi_1 + \epsilon_1, \pi_2 + \epsilon_2\}]. \tag{13}$$

In addition, if $w(\pi_1, \pi_2)$ is superlinear, then there exists a distribution θ with finite expectation (i.e., $\mathbb{E}_\theta|\epsilon_1| < \infty$ and $\mathbb{E}_\theta|\epsilon_2| < \infty$) that satisfies (13).

Note that the first part of Theorem 4 can be partly derived from McFadden (1980). However, in McFadden (1980), the author requires w to be second-order differentiable while we only require w to be differentiable. In addition, we also derive the relation between w being superlinear and θ having finite expectation. In the appendix, we give a direct and complete proof for this theorem.

By Proposition 2 proved in Hofbauer and Sandholm (2002), when $n \geqslant 4$, the RAM strictly subsumes the RUM (thus by Theorem 1 also the SCM and the welfare-based choice model). We next show that this is also true when $n = 3$. We start with the following "substitutable" property for the RUM.

Proposition 6. *For any RUM $\mathbf{q}(\boldsymbol{\pi})$, we must have $q_j(\boldsymbol{\pi}) \leqslant q_j(\boldsymbol{\pi} + h\mathbf{e}_i)$ for all $\boldsymbol{\pi} \in \mathcal{R}^n$, $h \geqslant 0$ and $i \neq j$.*

Proposition 6 says that in a RUM, if we increase the deterministic utility of one alternative while holding all other utilities unchanged, then the choice probabilities for all other alternatives must not increase. In the following two examples, we provide two choice models with $n = 3$ that violate the "substitutable" property in Proposition 6, derived from the welfare-based choice model or the RAM. These examples show that the welfare-based choice model (thus also the RAM and the SCM) strictly subsumes the RUM for $n = 3$.

Example 1. Consider $\mathbf{q}(\boldsymbol{\pi}) = \arg\max\{\boldsymbol{\pi}^T\mathbf{x} - V(\mathbf{x}) \,|\, \mathbf{x} \in \Delta_{n-1}\}$, where $V(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$ with

$$\mathbf{A} = \begin{bmatrix} 3 & 2 & 0 \\ 2 & 3 & 2 \\ 0 & 2 & 3 \end{bmatrix} \succ \mathbf{0}.$$

When we fix $\pi_2 = \pi_3 = 0$ and plot the choice probabilities against π_1 in the range of values $[-2, 2]$ as shown in Figure 1, it is observed that q_3 increases in π_1 in the range of $[-1.5, -1]$, i.e., it does not satisfy the property stated in Proposition 6. Thus, there is no RUM that is equivalent to this choice model. □

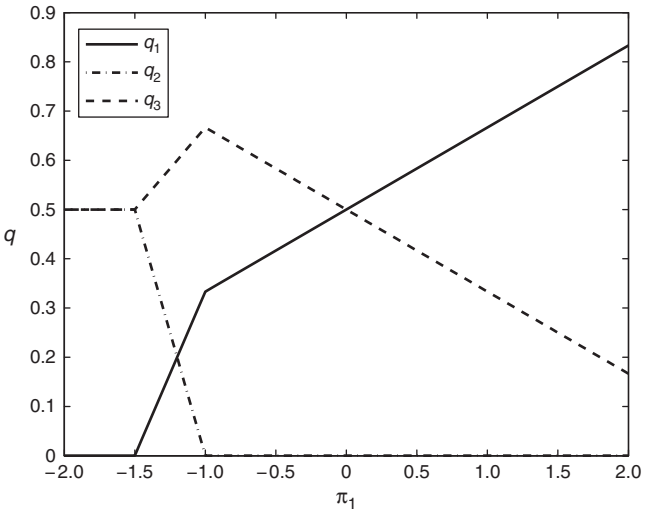
Example 2. Consider a function of three variables:

$$w(\boldsymbol{\pi}) = \log(e^{\pi_1} + e^{\pi_2} + e^{\pi_3} + e^{0.5(\pi_1+\pi_2)}).$$

Clearly, $w(\boldsymbol{\pi})$ is monotone, translation invariant and convex, therefore it is a choice welfare function. Also, it is differentiable with the corresponding choice probability:

$$\mathbf{q}(\boldsymbol{\pi}) = \frac{1}{e^{\pi_1} + e^{\pi_2} + e^{\pi_3} + e^{0.5(\pi_1+\pi_2)}} \cdot \left(e^{\pi_1} + \frac{1}{2}e^{0.5(\pi_1+\pi_2)}, e^{\pi_2} + \frac{1}{2}e^{0.5(\pi_1+\pi_2)}, e^{\pi_3}\right).$$

Figure 1. Choice Probabilities in Example 1 with $\pi_2 = \pi_3 = 0$



Furthermore, the second-order derivative of $w(\boldsymbol{\pi})$ with respect to π_1 and π_2 is

$$\frac{\partial^2 w(\boldsymbol{\pi})}{\partial\pi_1\partial\pi_2} = \frac{\partial q_1(\boldsymbol{\pi})}{\partial\pi_2} = \frac{\partial q_2(\boldsymbol{\pi})}{\partial\pi_1} = \frac{e^{0.5(\pi_1+\pi_2)}(-e^{\pi_1} - e^{\pi_2} + e^{\pi_3} - 4e^{0.5(\pi_1+\pi_2)})}{4(e^{\pi_1} + e^{\pi_2} + e^{\pi_3} + e^{0.5(\pi_1+\pi_2)})^2}.$$

It is non-positive if and only if $e^{\pi_3} \leqslant 4e^{0.5\pi_1+0.5\pi_2} + e^{\pi_1} + e^{\pi_2}$. Therefore, when π_3 is large while both π_1 and π_2 are small, this choice model will violate the property stated in Proposition 6. Thus, there is no RUM that is equivalent to this choice model. □

4. Conclusion

In this paper, we proposed a welfare-based approach for studying discrete choice models. We showed that the welfare-based choice model is equivalent to the RAM and the semiparametric model, thus establishing the equivalence between the latter two. We also showed the relationship between these choice models and the RUM. In particular, we showed that the welfare-based choice model (thus also the RAM and the SCM) strictly subsumes the RUM when there are three or more alternatives, while they are equivalent when there are only two alternatives.

Acknowledgments

The authors thank the area editor, the associate editor and the referees for their helpful comments and suggestions.

Appendix

Proof of Theorem 1. First we show that the $w(\boldsymbol{\pi})$ defined in (10) and (11) are choice welfare functions. To see this, note that the monotonicity and translation invariance properties are immediate from (10) and (11). For the convexity, note that $w(\boldsymbol{\pi})$ defined in (10) is the supremum of linear functions of $\boldsymbol{\pi}$

thus is convex in $\boldsymbol{\pi}$. In (11), for each $\boldsymbol{\epsilon}$, $\max_{i\in\mathcal{N}}\{\pi_i+\epsilon_i\}$ is a convex function in $\boldsymbol{\pi}$, and so is the expectation. Therefore, if $w(\boldsymbol{\pi})$ is defined by (10) or (11), then it must be a choice welfare function.

Next we show the other direction. That is, if $w(\boldsymbol{\pi})$ is a choice welfare function, then it can be represented in the form of (10) and (11). First, if a choice welfare function $w(\boldsymbol{\pi})=+\infty$ for some $\boldsymbol{\pi}$, then for any $\boldsymbol{\pi}'$, we have $w(\boldsymbol{\pi}')\geqslant w(\boldsymbol{\pi}+\min_i(\pi_i'-\pi_i)\mathbf{e})=w(\boldsymbol{\pi})+\min_i(\pi_i'-\pi_i)=+\infty$, where the first inequality uses the monotonicity property and the first equality uses the translation invariance property. Thus $w(\boldsymbol{\pi})=+\infty$ for all $\boldsymbol{\pi}$. In that case, we can choose $V(\mathbf{x})=-\infty$ and $\Theta=\{\theta_\infty\}$ where θ_∞ is a singleton distribution taking value on $(\infty,\ldots,\infty)$. Therefore, $w(\boldsymbol{\pi})$ can be represented by (10) and (11) in that case. Similarly, if $w(\boldsymbol{\pi})=-\infty$ for some $\boldsymbol{\pi}$, then it must be that $w(\boldsymbol{\pi})=-\infty$ for all $\boldsymbol{\pi}$, and we can take $V(\mathbf{x})=\infty$ and $\Theta=\{\theta_{-\infty}\}$, where $\theta_{-\infty}$ is a singleton distribution on $(-\infty,\ldots,-\infty)$. Therefore, $w(\boldsymbol{\pi})$ can be represented in (10) and (11) in this case too.

In the remainder of the proof, we focus on the case where $w(\boldsymbol{\pi})$ is finite for all $\boldsymbol{\pi}$. In this case, by Bertsekas (2003, Proposition 1.4.6), $w(\boldsymbol{\pi})$ must be continuous. The remaining proof is divided into two parts:

1. We show that any choice welfare function $w(\boldsymbol{\pi})$ can be represented by (10). Since $w(\boldsymbol{\pi})$ is monotone and translation invariant, the following holds:

$$\begin{aligned}w(\boldsymbol{\pi})&=\min_{\mathbf{y}}\left\{w(\mathbf{y})+\max_i\{\pi_i-y_i\}\right\}\\&=\min_{\mathbf{y}}\left\{w(\mathbf{y})+\max_{\mathbf{x}\in\Delta_{n-1}}(\boldsymbol{\pi}-\mathbf{y})^T\mathbf{x}\right\}.\end{aligned}$$

Here the first equality holds since for any $\mathbf{y}$, $w(\boldsymbol{\pi})=w(\boldsymbol{\pi}-\max_i\{\pi_i-y_i\}\mathbf{e})+\max_i\{\pi_i-y_i\}$ by the translation invariance property. Furthermore, by the monotonicity property, $w(\boldsymbol{\pi}-\max_i\{\pi_i-y_i\}\mathbf{e})\leqslant w(\mathbf{y})$ and the equality holds when $\mathbf{y}=\boldsymbol{\pi}$.

Next we define $L(\mathbf{x},\mathbf{y})=w(\mathbf{y})+(\boldsymbol{\pi}-\mathbf{y})^T\mathbf{x}$. We have for fixed $\mathbf{x}$, $L(\mathbf{x},\cdot)$ is convex in $\mathbf{y}$ (by the convexity of $w(\cdot)$); and for fixed $\mathbf{y}$, $L(\cdot,\mathbf{y})$ is convex and closed in $\mathbf{x}$. Furthermore, $\inf_{\mathbf{y}}\max_{\mathbf{x}\in\Delta_{n-1}}L(\mathbf{x},\mathbf{y})=w(\boldsymbol{\pi})<\infty$ and the function $p(\mathbf{u})=\inf_{\mathbf{y}}\max_{\mathbf{x}\in\Delta_{n-1}}\{L(\mathbf{x},\mathbf{y})-\mathbf{u}^T\mathbf{x}\}=w(\boldsymbol{\pi}-\mathbf{u})$ is continuous. Therefore, by Bertsekas (2003, Proposition 2.6.2), the minimax equality holds, i.e.,

$$\inf_{\mathbf{y}}\max_{\mathbf{x}\in\Delta_{n-1}}L(\mathbf{x},\mathbf{y})=\max_{\mathbf{x}\in\Delta_{n-1}}\inf_{\mathbf{y}}L(\mathbf{x},\mathbf{y}).$$

Therefore, we have:

$$w(\boldsymbol{\pi})=\max_{\mathbf{x}\in\Delta_{n-1}}\left\{\boldsymbol{\pi}^T\mathbf{x}+\inf_{\mathbf{y}}\{w(\mathbf{y})-\mathbf{y}^T\mathbf{x}\}\right\}=\max_{\mathbf{x}\in\Delta_{n-1}}\{\boldsymbol{\pi}^T\mathbf{x}-V(\mathbf{x})\}$$

where $V(\mathbf{x})=\sup_{\mathbf{y}}\{\mathbf{y}^T\mathbf{x}-w(\mathbf{y})\}$ is a convex function.

2. Next we show that any choice welfare function can be represented by (11). Since $w(\boldsymbol{\pi})$ is convex, there exists a subgradient for any $\boldsymbol{\pi}$. We denote the subgradient vector by $\mathbf{d}(\boldsymbol{\pi})=(d_1(\boldsymbol{\pi}),\ldots,d_n(\boldsymbol{\pi}))^T$. Here it is possible that the choice of $\mathbf{d}(\boldsymbol{\pi})$ is not unique, in which case, the choice can be arbitrary. Furthermore, by taking the derivative with respect to t in the translation invariance equation, and by applying the chain rule (see Bertsekas 2003, Proposition 4.2.5), we have for any subgradient $\mathbf{d}(\boldsymbol{\pi})$, it must hold that $\mathbf{e}^T\mathbf{d}(\boldsymbol{\pi})=1$. Similarly, by the monotonicity property of $w(\boldsymbol{\pi})$, we must have $\mathbf{d}(\boldsymbol{\pi})\geqslant\mathbf{0}$. By the definition of subgradient and the convexity of $w(\boldsymbol{\pi})$, we must have:

$$w(\boldsymbol{\pi})\geqslant(\boldsymbol{\pi}-\mathbf{z})^T\mathbf{d}(\mathbf{z})+w(\mathbf{z}),\quad\forall\,\mathbf{z}\in\mathcal{R}^n,$$

where the equality holds when $\mathbf{z}=\boldsymbol{\pi}$. Define $l(\mathbf{z})=w(\mathbf{z})-\mathbf{z}^T\mathbf{d}(\mathbf{z})$. By reorganizing terms, we have

$$w(\boldsymbol{\pi})=\sup_{\mathbf{z}}\{\boldsymbol{\pi}^T\mathbf{d}(\mathbf{z})+l(\mathbf{z})\}.\tag{A.1}$$

Now we define the distribution set as follows: Let $\Theta=\{\theta_{\mathbf{z}}\,|\,\mathbf{z}\in\mathcal{R}^n\}$, where $\theta_{\mathbf{z}}$ is an n-point distribution with

$$\mathbb{P}_{\theta_{\mathbf{z}}}(\boldsymbol{\epsilon}=\boldsymbol{\epsilon}_{\mathbf{z}}^i)=d_i(\mathbf{z}),\quad\text{for } i=1,\ldots,n$$

where

$$\epsilon_{\mathbf{z}}^i(j)=\begin{cases}l(\mathbf{z}) & \text{if } j=i,\\ -\infty & \text{if } j\neq i.\end{cases}$$

That is, $\epsilon_{\mathbf{z}}^i$ is a vector of all $-\infty$'s except $l(\mathbf{z})$ at the ith entry. Therefore, for any $\mathbf{z}$, we have

$$\mathbb{E}_{\theta_{\mathbf{z}}}\left[\max_i(\pi_i+\epsilon_i)\right]=\sum_{i=1}^n d_i(\mathbf{z})(\pi_i+l(\mathbf{z}))=\boldsymbol{\pi}^T\mathbf{d}(\mathbf{z})+l(\mathbf{z}).$$

Then by (A.1), we have

$$\begin{aligned}w(\boldsymbol{\pi})&=\sup_{\mathbf{z}}\{\boldsymbol{\pi}^T\mathbf{d}(\mathbf{z})+l(\mathbf{z})\}=\sup_{\mathbf{z}}\mathbb{E}_{\theta_{\mathbf{z}}}\left[\max_i(\pi_i+\epsilon_i)\right]\\&=\sup_{\theta\in\Theta}\mathbb{E}_{\theta}\left[\max_i(\pi_i+\epsilon_i)\right].\end{aligned}$$

Therefore, the theorem is proved. □

Proof of Theorem 2. The equivalence between 1 and 3 directly follows from Theorem 1. Next we show that 1 ⇒ 2. If $w(\boldsymbol{\pi})$ is a differentiable choice welfare function, by Theorem 1, we know that

$$w(\boldsymbol{\pi})=\max\{\boldsymbol{\pi}^T\mathbf{x}-V(\mathbf{x})\,|\,\mathbf{x}\in\Delta_{n-1}\},$$

where $V(\mathbf{x})=\sup_{\mathbf{y}}\{\mathbf{y}^T\mathbf{x}-w(\mathbf{y})\}$. Therefore, $V(\mathbf{x})$ is the convex conjugate of $w(\boldsymbol{\pi})$. By Rockafellar (1974, Theorem 6.3), we know that $w(\boldsymbol{\pi})$ is essentially differentiable if and only if $V(\mathbf{x})$ is essentially strictly convex. Also, from the envelope theorem (see Mas-Colell et al. 1995),

$$\nabla w(\boldsymbol{\pi})=\nabla_{\boldsymbol{\pi}}(\boldsymbol{\pi}^T\mathbf{x}-V(\mathbf{x}))|_{\mathbf{x}=\mathbf{x}^*}=\mathbf{x}^*,$$

where $\mathbf{x}^*=\arg\max\{\boldsymbol{\pi}^T\mathbf{x}-V(\mathbf{x})\,|\,\mathbf{x}\in\Delta_{n-1}\}$. Therefore,

$$\mathbf{q}(\boldsymbol{\pi})=\nabla w(\boldsymbol{\pi})=\arg\max\{\boldsymbol{\pi}^T\mathbf{x}-V(\mathbf{x})\,|\,\mathbf{x}\in\Delta_{n-1}\}.$$

Last, we show that 2 ⇒ 1. Given an essentially strictly convex $V(\mathbf{x})$, by Theorem 1, we know that

$$w(\boldsymbol{\pi})=\max\{\boldsymbol{\pi}^T\mathbf{x}-V(\mathbf{x})\,|\,\mathbf{x}\in\Delta_{n-1}\}$$

is a choice welfare function. Again, by Rockafellar (1974, Theorem 6.3), we know that $w(\boldsymbol{\pi})$ is essentially differentiable. Moreover, in our case, $w(\boldsymbol{\pi})$ is a convex and finite-valued function in $\mathcal{R}^n$, thus essentially differentiability is equivalent to differentiability. Again, by applying the envelope theorem, $\mathbf{q}(\boldsymbol{\pi})=\nabla w(\boldsymbol{\pi})$. Therefore the theorem is proved. □

Proof of Theorem 3. First, we show the equivalence between 1 and 2. Based on Theorem 2, it suffices to prove that $w(\boldsymbol{\pi})$ is superlinear if and only if $V(\mathbf{x})$ defined by $\max_{\mathbf{y}}\{\mathbf{y}^T\mathbf{x} - w(\mathbf{y})\}$ is upper bounded. If $w(\boldsymbol{\pi})$ is superlinear, we have, for any $\mathbf{x} \in \Delta_{n-1}$,

$$w(\boldsymbol{\pi}) \geqslant \sum_{i\in\mathcal{N}} x_i(\pi_i + b_i) = \mathbf{x}^T\boldsymbol{\pi} + \mathbf{x}^T\mathbf{b} \geqslant \mathbf{x}^T\boldsymbol{\pi} + \min_i b_i.$$

By reorganizing terms, we have

$$\mathbf{x}^T\boldsymbol{\pi} - w(\boldsymbol{\pi}) \leqslant -\min_i\{b_i\} = \max_i\{-b_i\}.$$

Therefore, $V(\mathbf{x}) = \max_{\mathbf{y}}\{\mathbf{y}^T\mathbf{x} - w(\mathbf{y})\} \leqslant \max_i\{-b_i\}$, i.e., $V(\mathbf{x})$ is upper bounded.

To show the other direction, if $V(\mathbf{x})$ is upper bounded by a constant u, then we have

$$w(\boldsymbol{\pi}) \geqslant \max\{\boldsymbol{\pi}^T\mathbf{x} - u \mid \mathbf{x} \in \Delta_{n-1}\} \geqslant \pi_i - u, \quad \forall\, i,$$

i.e., $w(\boldsymbol{\pi})$ is superlinear. Therefore, the equivalence between 1 and 2 is proved.

Next we show the equivalence between 1 and 3. We first show that for any superlinear differentiable choice welfare function $w(\boldsymbol{\pi})$, we can find a distribution set Θ consisting of only distributions with finite expectation such that $w(\boldsymbol{\pi})$ can be represented as $w(\boldsymbol{\pi}) = \sup_{\theta\in\Theta}\mathbb{E}_\theta[\max_{i\in\mathcal{N}}\pi_i + \epsilon_i]$.

First, since $w(\boldsymbol{\pi})$ is convex with $\mathbf{q}(\boldsymbol{\pi}) = \nabla w(\boldsymbol{\pi})$, we have

$$w(\boldsymbol{\pi}) = \sup_{\mathbf{z}}\{\boldsymbol{\pi}^T\mathbf{q}(\mathbf{z}) + l(\mathbf{z})\}, \tag{A.2}$$

where $l(\mathbf{z}) = w(\mathbf{z}) - \mathbf{z}^T\mathbf{q}(\mathbf{z})$. Now we define a distribution set Θ that is slightly different from that of Theorem 1. Specifically, let $\Theta = \{\theta_{\mathbf{z}} \mid \mathbf{z} \in \mathcal{R}^n\}$, where $\theta_{\mathbf{z}}$ is an n-point distribution with $\mathbb{P}_{\theta_{\mathbf{z}}}(\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^i_{\mathbf{z}}) = q_i(\mathbf{z})$, $\forall\, i \in \mathcal{N}$ (Note that by the monotonicity and the translation invariance properties, $\mathbf{q}(\mathbf{z}) = \nabla w(\mathbf{z})$ must satisfy $\mathbf{q}(\mathbf{z}) \geqslant \mathbf{0}$ and $\mathbf{e}^T\mathbf{q}(\mathbf{z}) = 1$). Here,

$$\epsilon^i_{\mathbf{z}}(j) = \begin{cases} l(\mathbf{z}) & \text{if } j = i, \\ l(\mathbf{z}) - M(\mathbf{z}) & \text{if } j \neq i. \end{cases}$$

where

$$M(\mathbf{z}) = \max\left\{1 + \max_{i,j}\{z_i - z_j\}, \frac{l(\mathbf{z}) - \min_i\{b_i\}}{t^*(\mathbf{z})}\right\}, \tag{A.3}$$

with

$$t^*(\mathbf{z}) = \min\{q_i(\mathbf{z}) \mid q_i(\mathbf{z}) > 0\}. \tag{A.4}$$

Since $M(\mathbf{z}) > z_i - z_j$, for all i, j, we have $i = \arg\max_j(z_j + \epsilon^i_{\mathbf{z}}(j))$. Therefore,

$$\mathbb{E}_{\theta_{\mathbf{z}}}\left[\max_j\ z_j + \epsilon_j\right] = \sum_{i=1}^n q_i(\mathbf{z})(z_i + l(\mathbf{z})) = \mathbf{z}^T\mathbf{q}(\mathbf{z}) + l(\mathbf{z}) = w(\mathbf{z}).$$

Next we show that:

$$\mathbb{E}_{\theta_{\mathbf{z}}}\left[\max_i\ \pi_i + \epsilon_i\right] \leqslant w(\boldsymbol{\pi}), \quad \forall\, \boldsymbol{\pi}.$$

For any given $\boldsymbol{\pi}$, define $k(i) \triangleq \arg\max_j\,(\pi_j + \epsilon^i_{\mathbf{z}}(j))$ (we break ties arbitrarily). There are two cases:

1. For all i such that $q_i(\mathbf{z}) > 0$, $k(i) = i$. In this case, we have

$$\mathbb{E}_{\theta_{\mathbf{z}}}\left[\max_j \pi_j + \epsilon_j\right] = \sum_{i\in\mathcal{N}} q_i(\mathbf{z})(\pi_i + l(\mathbf{z})) = \boldsymbol{\pi}^T\mathbf{q}(\mathbf{z}) + l(\mathbf{z}) \leqslant w(\boldsymbol{\pi}),$$

in which the last inequality is because of the convexity of $w(\cdot)$.

2. There exists some i such that $q_i(\mathbf{z}) > 0$, but $k(i) \neq i$. In this case, from the construction of $\theta_{\mathbf{z}}$, we have

$$\begin{aligned}\mathbb{E}_{\theta_{\mathbf{z}}}\left[\max_j \pi_j + \epsilon_j\right] &= \sum_{i\in\mathcal{N},\, q_i(\mathbf{z})>0} q_i(\mathbf{z})(\pi_{k(i)} + l(\mathbf{z}) - M(\mathbf{z})\mathbb{I}_{\{k(i)\neq i\}}) \\ &\leqslant \max_i\{\pi_i\} + l(\mathbf{z}) - t^*(\mathbf{z})M(\mathbf{z}) \\ &\leqslant \max_i\{\pi_i\} + \min_j\{b_j\} \\ &\leqslant \max_i\ \{\pi_i + b_i\} \\ &\leqslant w(\boldsymbol{\pi}),\end{aligned}$$

where the first inequality follows from the fact that $M(\mathbf{z}) > 0$ and $\sum_{i\in\mathcal{N}} q_i(\mathbf{z})\mathbb{I}_{\{q_i(\mathbf{z})>0, k(i)\neq i\}} \geqslant t^*(\mathbf{z})$, the second inequality is because of the definition of $M(\mathbf{z})$ and the last inequality follows from the definition of superlinear function.

Based on the analysis of these two cases, we have

$$\mathbb{E}_{\boldsymbol{\epsilon}\sim\theta_{\mathbf{z}}}\left[\max_i \pi_i + \epsilon_i\right] \leqslant w(\boldsymbol{\pi}), \quad \forall\, \boldsymbol{\pi}.$$

Then by Equation (A.2) we have

$$\begin{aligned}w(\boldsymbol{\pi}) &= \sup_{\mathbf{z}}\{\boldsymbol{\pi}^T\mathbf{q}(\mathbf{z}) + l(\mathbf{z})\} = \sup_{\mathbf{z}}\mathbb{E}_{\theta_{\mathbf{z}}}\left[\max_i \pi_i + \epsilon_i\right] \\ &= \sup_{\theta\in\Theta}\mathbb{E}_\theta\left[\max_i \pi_i + \epsilon_i\right].\end{aligned}$$

Therefore, we have proved that statement 1 implies statement 3.

Finally, we prove that statement 3 implies statement 1. Suppose there exists a distribution $\hat{\theta} \in \Theta$ such that $\mathbb{E}_{\hat{\theta}}|\epsilon_i| < +\infty$ for $\forall\, i \in \mathcal{N}$, then for $\boldsymbol{\pi} \in \mathcal{R}^n$ we have

$$\begin{aligned}\sup_{\theta\in\Theta}\mathbb{E}_\theta\left[\max_{i\in\mathcal{N}} \pi_i + \epsilon_i\right] &\geqslant \mathbb{E}_{\hat{\theta}}\left[\max_{i\in\mathcal{N}} \pi_i + \epsilon_i\right] \\ &\geqslant \mathbb{E}_{\hat{\theta}}[\pi_j + \epsilon_j] = \pi_j + \mathbb{E}_{\hat{\theta}}[\epsilon_j], \quad \forall\, j.\end{aligned}$$

Therefore we can conclude that

$$w(\boldsymbol{\pi}) = \sup_{\theta\in\Theta}\mathbb{E}_\theta\left[\max_{i\in\mathcal{N}}(\pi_i + \epsilon_i)\right]$$

is superlinear.

It remains to prove the last statement. We show that for any

$$\mathbf{x} \in \Delta^\circ_{n-1} \triangleq \{\mathbf{x} \mid \mathbf{e}^T\mathbf{x} = 1, x_i > 0, \forall\, i \in \mathcal{N}\},$$

there exists $\boldsymbol{\pi}_{\mathbf{x}}$ such that $\mathbf{q}(\boldsymbol{\pi}_{\mathbf{x}}) = \nabla w(\boldsymbol{\pi}_{\mathbf{x}}) = \mathbf{x}$. Fix $\mathbf{x} \in \Delta^\circ_{n-1}$, we consider

$$V(\mathbf{x}) = \max_{\boldsymbol{\pi}}\{\boldsymbol{\pi}^T\mathbf{x} - w(\boldsymbol{\pi})\}. \tag{A.5}$$

Clearly, $V(\mathbf{x}) \geqslant -w(\mathbf{0})$, since $\boldsymbol{\pi} = \mathbf{0}$ is a feasible solution. Moreover, since $w(\boldsymbol{\pi})$ is translation invariant, we can restrict the feasible region of (A.5) to $\mathcal{L} \triangleq \{\boldsymbol{\pi} | \mathbf{e}^T\boldsymbol{\pi} = 0\}$. For all $\boldsymbol{\pi} \in \mathcal{L}$, we have $\pi_j \leqslant 0$ for some $j \in \mathcal{N}$. Thus

$$\boldsymbol{\pi}^T\mathbf{x} \leqslant \sum_{i\neq j} \pi_i x_i \leqslant \sum_{i\neq j} x_i \max_k\{\pi_k\} \leqslant \left(1 - \min_i\{x_i\}\right)\max_k\{\pi_k\}.$$

However, by superlinearity of $w(\boldsymbol{\pi})$, we have:

$$w(\boldsymbol{\pi}) \geqslant \max_k\{\pi_k + b_k\} \geqslant \max_k\{\pi_k\} + \min_k\{b_k\}.$$

Thus, for all $\boldsymbol{\pi} \in \mathcal{L}$, we have:

$$\boldsymbol{\pi}^T\mathbf{x} - w(\boldsymbol{\pi}) \leqslant -\min_i\{x_i\}\max_k\{\pi_k\} - \min_k\{b_k\}.$$

Let $K = (w(\mathbf{0}) - \min_k\{b_k\})/\min_i\{x_i\}$. For $\boldsymbol{\pi}$ to be optimal to (A.5), by the above arguments, we would have $\pi_i \leqslant K$ for all i. Thus we can further restrict the feasible set of (A.5) to $\{\boldsymbol{\pi} \mid \mathbf{e}^T\boldsymbol{\pi} = 0, \pi_i \leqslant K\ \forall\, i \in \mathcal{N}\}$, which is a compact set. Since $w(\boldsymbol{\pi})$ is continuous, there exists $\boldsymbol{\pi}_{\mathbf{x}} \in \{\boldsymbol{\pi} \mid \mathbf{e}^T\boldsymbol{\pi} = 0, \pi_i \leqslant K\ \forall\, i \in \mathcal{N}\}$ that attains maximum in problem (A.5). By the first-order necessary condition, $\nabla w(\boldsymbol{\pi}_{\mathbf{x}}) = \mathbf{x}$. This concludes the proof. □

Proof of Theorem 4. Define $v(x) \triangleq w(x, 0)$. Since $w(\cdot)$ is differentiable, by the chain rule, we have

$$v'(x) = \frac{\partial w}{\partial \pi_1}(x, 0).$$

Since $w(\pi_1, \pi_2)$ is convex and satisfies the translation invariance property, we have $v'(x) \in [0, 1]$ and is increasing. We define a distribution θ of $\{\epsilon_1, \epsilon_2\}$ as follows:

$$\{\epsilon_1, \epsilon_2\} = \{v_0 - \max\{\xi, 0\}, v_0 - \max\{-\xi, 0\}\},$$

where $v_0 = v(0) = w(0, 0)$ and ξ is a random variable with c.d.f. $F_\xi(x) = \mathbb{P}(\xi \leqslant x) = v'(x)$. Note $F(\cdot)$ is a well-defined c.d.f. since $w(\cdot)$ is convex and differentiable, thus $v'(x)$ must be continuous and increasing (Rockafellar 1974).

Now we compute $\mathbb{E}_\theta[\max\{\pi_1 + \epsilon_1, \pi_2 + \epsilon_2\}]$. We have

$$\begin{aligned}
&\mathbb{E}_\theta[\max\{\pi_1 + \epsilon_1, \pi_2 + \epsilon_2\}] \\
&\quad = \pi_1 + v_0 + \mathbb{E}_\theta[\max\{-\max\{\xi, 0\}, \pi_2 - \pi_1 - \max\{-\xi, 0\}\}] \\
&\quad = \pi_1 + v_0 + \mathbb{E}_\theta[\max\{0, \pi_2 - \pi_1 + \xi\} - \max\{\xi, 0\}],
\end{aligned}$$

where the last step can be verified by considering $\xi \geqslant 0$ and $\xi \leqslant 0$, respectively.

Now we compute the last term. For $x \geqslant 0$, we have (let $\mathbb{I}(\cdot)$ be the indicator function):

$$\begin{aligned}
&\mathbb{E}_\theta[\max\{0, x + \xi\} - \max\{0, \xi\}] \\
&\quad = x\mathbb{P}(\xi > 0) + \mathbb{E}_\theta[(x + \xi) \cdot \mathbb{I}(-x < \xi \leqslant 0)] \\
&\quad = x\mathbb{P}(\xi > 0) + \int_{-x}^{0} (x + \xi)\, dv'(\xi) \\
&\quad = x(1 - v'(0)) + (x + \xi)v'(\xi)\big|_{-x}^{0} - \int_{-x}^{0} v'(\xi)\, d\xi \\
&\quad = x - v_0 + v(-x).
\end{aligned}$$

Similarly, for $x \leqslant 0$, we have

$$\begin{aligned}
&\mathbb{E}_\theta[\max\{0, x + \xi\} - \max\{0, \xi\}] \\
&\quad = x\mathbb{P}(\xi > -x) + \mathbb{E}_\theta[-\xi \cdot \mathbb{I}(0 < \xi \leqslant -x)] \\
&\quad = x\mathbb{P}(\xi > -x) - \int_{0}^{-x} \xi\, dv'(\xi) \\
&\quad = x(1 - v'(-x)) - \xi v'(\xi)\big|_{0}^{-x} + \int_{0}^{-x} v'(\xi)\, d\xi \\
&\quad = x - v_0 + v(-x).
\end{aligned}$$

Therefore, $\mathbb{E}_\theta[\max\{\pi_1 + \epsilon_1, \pi_2 + \epsilon_2\}] = \pi_1 + v_0 + (\pi_2 - \pi_1) - v_0 + v(\pi_1 - \pi_2) = w(\pi_1, \pi_2)$.

To prove the last statement, it suffices to show that both $\mathbb{E}_\theta[\max\{0, \xi\}]$ and $\mathbb{E}_\theta[\max\{0, -\xi\}]$ are finite if $w(\boldsymbol{\pi})$ is superlinear. If $w(\cdot)$ is superlinear, then we have $v(t) - t = w(0, -t)$ is decreasing in t and lower bounded, thus $L_1 = \lim_{t\to+\infty}(v(t) - t)$ exists and is finite. Similarly, $v(t) = w(t, 0)$ is increasing in t and lower bounded, thus $L_2 = \lim_{t\to-\infty} v(t)$ exists and is finite. Therefore, we have:

$$\begin{aligned}
\mathbb{E}_\theta[\max\{0, \xi\}] &= \int_0^{+\infty} \mathbb{P}_\theta(\xi \geqslant t)\, dt = \int_0^{+\infty} (1 - v'(t))\, dt \\
&= (t - v(t))\big|_0^{+\infty} = v(0) - L_1,
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_\theta[\max\{0, -\xi\}] &= \int_0^{+\infty} \mathbb{P}_\theta(-\xi \geqslant t)\, dt = \int_0^{+\infty} v'(-t)\, dt \\
&= \int_{-\infty}^{0} v'(t)\, dt = v(0) - L_2.
\end{aligned}$$

Thus, the theorem is proved. □

Proof of Proposition 6. In a RUM, the probability of choosing alternative j is $q_j(\boldsymbol{\mu}) = \mathbb{P}_\theta(j = \arg\max_{k\in\mathcal{N}}(\mu_k + \epsilon_k))$. Since the choice probability is based on the comparison, it is clear that $q_j(\boldsymbol{\pi}) \leqslant q_j(\boldsymbol{\pi} + h\mathbf{e}_i)$ for all $\boldsymbol{\pi} \in \mathcal{R}^n$, $h \geqslant 0$ and $i \neq j$. □

Endnote

[1] If $V(\mathbf{x})$ is not convex or lower semicontinuous, then we can replace $V(\mathbf{x})$ by a convex and lower semicontinuous function $V^{**}(\mathbf{x}) = \sup_{\mathbf{y}}\{\mathbf{y}^T\mathbf{x} - w^r(\mathbf{y})\}$ and the Equation (3) still holds (Borwein and Lewis 2010). Therefore, it is without loss of generality to assume $V(\mathbf{x})$ is convex and lower semicontinuous.

References

Ahipasaoglu SD, Li X, Natarajan K (2016) A convex optimization approach for computing correlated choice probabilities with many alternatives. Working paper, Singapore University of Technology and Design.

Alptekinoğlu A, Semple J (2016) The exponomial choice model: A new alternative for assortment and price optimization. *Oper. Res.* 64(1):79–93.

Anderson SP, De Palma A, Thisse JF (1988) A representative customer theory of the logit model. *Internat. Econom. Rev.* 29(3):461–466.

Anderson SP, De Palma A, Thisse JF (1992) *Discrete Choice Theory of Product Differentiation* (The MIT Press, Cambridge, MA).

Ben-Akiva M, Lerman SR (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand* (The MIT Press, Cambridge, MA).

Bertsekas D (2003) *Convex Analysis and Optimization* (Athena Scientific, Belmont, MA).

Blanchet J, Gallego G, Goyal V (2016) A markov chain approximation to choice modeling. *Oper. Res.* 64(4):886–905.

Borwein JM, Lewis AS (2010) *Convex Analysis and Nonlinear Optimization: Theory and Examples* (Springer, New York).

Daganzo C (1980) *Multinomial Probit: The Theory and Its Application to Demand Forecasting* (Academic Press, New York).

Farias VF, Jagabathula S, Shah D (2013) A nonparametric approach to modeling choice with limited data. *Management Sci.* 59(2): 305–322.

Gallego G, Ratliff R, Shebalov S (2014) A general attraction model and sales-based linear program for network revenue management under customer choice. *Oper. Res.* 63(1):212–232.

Hofbauer J, Sandholm WH (2002) On the global convergence of stochastic fictitious play. *Econometrica* 70(6):2265–2294.

Jagabathula S, Rusmevichientong P (2013) A two-stage model of consideration set and choice: Learning, revenue prediction, and applications. Working paper, New York University.

Mas-Colell A, Whinston MD, Green JR (1995) *Microeconomic Theory* (Oxford University Press, New York).

McFadden D (1974) Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics* (Academic Press, New York), 105–142.

McFadden D (1980) Econometric models for probabilistic choice among products. *J. Bus.* 53(3):13–29.

Mishra VK, Natarajan K, Tao H, Teo C-P (2012) Choice prediction with semidefinite optimization when utilities are correlated. *IEEE Trans. Automatic Control* 57(10):2450–2463.

Mishra VK, Natarajan K, Padmanabhan D, Teo C-P, Li X (2014) On theoretical and empirical aspects of marginal distribution choice models. *Management Sci.* 60(6):1511–1531.

Natarajan K, Song M, Teo C-P (2009) Persistency model and its applications in choice modeling. *Management Sci.* 55(3):453–469.

Norets A, Takahashi S (2013) On the surjectivity of the mapping between utilities and choice probabilities. *Quant. Econom.* 4(1):149–155.

Rockafellar T (1974) *Conjugate Duality and Optimization* (SIAM, Philadelphia).

Thurstone L (1927) A law of comparative judgment. *Psych. Rev.* 34(4):273–286.

Train KE (2009) *Discrete Choice Methods with Simulation* (Cambridge University Press, New York).

Guiyun Feng is a Ph.D. student in the Department of Industrial and Systems Engineering at the University of Minnesota. Her research interest includes stochastic simulation and operations management.

Xiaobo Li is a Ph.D. student in the Department of Industrial and Systems Engineering at the University of Minnesota. His research interest includes robust optimization, optimization algorithms, and data-driven decision making.

Zizhuo Wang is an assistant professor in the Department of Industrial and Systems Engineering at the University of Minnesota. His research interest includes pricing and revenue management, operations management, and data-driven decision making problems.