Analysis of Discrete Choice Models: A Welfare-Based Framework

Guiyun Feng  
Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN 55455, fengx421@umn.edu

Xiaobo Li  
Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN 55455, lixx3195@umn.edu

Zizhuo Wang  
Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN 55455, zwang@umn.edu

Based on the observation that many existing discrete choice models admit a welfare function of utilities whose gradient gives the choice probability vector, we propose a new representation of discrete choice model which we call the welfare-based choice model. The welfare-based choice model is meaningful on its own by providing a new way of constructing choice models. More importantly, it provides great analysis convenience for establishing connections among existing choice models. We prove by using convex analysis theory, that the welfare-based choice model is equivalent to the representative agent choice model and the semi-parametric choice model, establishing the equivalence of the latter two. We show that these three models are all strictly more general than the random utility model, while when there are only two alternatives, those four models are equivalent. In particular, we show that the distinction between the welfare-based choice model and the random utility model lies in the requirement of the higher-order derivatives of the welfare function. We then define a new concept in choice models: substitutability/complementarity between alternatives. We show that the random utility model only allows substitutability between different alternatives; while the welfare-based choice model allows flexible substitutability/complementarity patterns. We argue that such flexibility could be desirable in capturing certain practical choice patterns and expanding the scope of discrete choice models. Examples are given of new choice models proposed under our framework.

Key words: welfare-based choice model, random utility model, representative agent model, semi-parametric choice model, substitutability, complementarity, convex optimization

History: This version: March 9, 2015

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. For example, they are used to examine which product to purchase for a consumer, which mode of transportation to take for a passenger, among many other choice scenarios people face everyday. 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 (see Anderson et al. 1992 and Ben-Akiva and Lerman 1985). Among them, the most popular one 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 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 semi-parametric models has been proposed (see 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 they choose one extreme distribution in that set to determine the choice probabilities. There are other choice models based on the dynamics of choice decisions or other non-parametric ideas. We will provide a more detailed review of these models in Section 2.

Although these models have all provided excellent explanations, both theoretically and empirically, for how people make choices in practice, some gaps in the literature still exist regarding the relations between those popular choice models. In particular, the following questions are not answered in the prior literature:

1. What is the relation between the representative agent model and the semi-parametric model?
   It has been shown that for many special cases, the semi-parametric model can be represented as a representative agent model. However, it is unknown whether this is generally true.

2. It is known that both the representative agent model and the semi-parametric model are more general than the random utility model. What exactly is the distinction between these models?

3. What choice pattern is restricted in the random utility model? Can we easily construct choice models that relax those restrictions?

In this paper, we present precise answers to the above questions. To answer those questions, we propose a new class of choice models, which we call the welfare-based choice model. The welfare-based choice model 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. Therefore, by directly proposing desirable conditions on the welfare functions, we define the class of welfare-based choice models. We show that the welfare-based choice model is not only meaningful on its own, but also provides great analysis convenience for establishing connections between existing choice models.

First, we show that by using the welfare-based choice model as an intermediate model, the classes of choice models defined by: 1) the welfare-based choice model, 2) the representative agent model and 3) the semi-parametric model, are the same. More precisely, under mild regularity assumptions, given any of the following three: a choice welfare function (which defines a welfare-based choice model), a regularization function (which defines a representative agent model) or a distribution set (which defines a semi-parametric model), one can construct the other two to define exactly the same choice model. This means that the class of representative agent models and the class of semi-parametric models are equivalent to each other, which 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 methods to convert one model to another in an explicit way, potentially alleviating the need to construct correspondences in a case by case manner as is done in the current research.

Second, we study the relation 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 more than two alternatives, in which case the above three models strictly subsume the random utility model. In particular, we point out the exact distinction between these three models and the random utility model, which lies in the higher-order derivatives of the choice function. Our result gives precise relations among those models.

Finally, by examining the difference between the welfare-based choice model and the random utility model, we identify an important property that is restricted in the random utility model but is flexible in the other three models. We call the property substitutability and complementarity of alternatives. Specifically, this property examines whether the choice probability of another alternative will increase or decrease when the utility of one alternative increases. We show that random utility models only allow substitutability between alternatives. Although this is natural in many practical situations, we argue that in certain applications, it might be appealing to allow some alternatives to exhibit complementarity in certain range, especially when the utility is based on scores and certain alternatives share the same feature (e.g., brand or certain component) on
which the score is based. We derive conditions under which different choice models exhibit substitutable/complementary properties. As far as we know, we are the first to study such properties in choice models, and we believe that this study will open new possibilities in the design of choice models by enlarging its horizon and capturing more practical choice patterns. In fact, we show a few examples of new choice models that allow complementarity among choices (in a certain range) and explain the practicality of those models.

It is worth mentioning that the analysis technique used in this paper is novel and interesting. In particular, we adopt several convex analysis tools that are not commonly used in the study of choice models. Such tools enable us to uncover deep connections between seemingly unrelated models and are key to our findings. We believe that such analysis methods may be of independent interest in the future study of choice models.

The remainder of this paper is organized as follows: In Section 2, we review discrete choice models that are relevant to our study. In Section 3, we propose the welfare-based choice model and study its relation with other choice models. In Section 4, we study the relation between the welfare-based choice model and the random utility model. In Section 5, we propose the concept of substitutability and complementarity between choice alternatives and derive conditions under which each model exhibits such properties. We discuss the issue of constructing new choice models from existing ones in Section 6. We conclude the paper in Section 7.

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 $i$th 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 $\mathbf{x} \geq \mathbf{y}$ to denote a componentwise relationship and $\Delta_{n-1}$ to denote the $n-1$-dimensional simplex, i.e., $\Delta_{n-1} = \{\mathbf{x} | \mathbf{e}^T \mathbf{x} = 1, \mathbf{x} \geq \mathbf{0}\}$. In our discussions, ordinary lowercase letters $x, y, \ldots$ denote scalars, 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), proposed first 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 could be due to the lack of information of the alternatives for a particular individual or to the idiosyncrasies of preferences among a population. As the output, the random utility model predicts a vector of choice probabilities among the alternatives, rather than a single deterministic choice. Mathematically, suppose there are \( n \) alternatives denoted by \( \mathcal{N} = \{1, 2, ..., n\} \), then the random utility model assumes that the utility of alternative \( i \) takes the following form:

\[
    u_i = \mu_i + \epsilon_i, \quad \forall i \in \mathcal{N},
\]

(1)

where \( \mu = (\mu_1, ..., \mu_n) \) is the deterministic part of the utility and \( \epsilon = (\epsilon_1, ..., \epsilon_n) \) is the random part. In the random utility model, it is assumed that the joint distribution \( \theta \) of \( \epsilon = (\epsilon_1, ..., \epsilon_n) \) is known. Then the probability that alternative \( i \) will be chosen is (to ensure the following equation is well-defined, we assume \( \theta \) is absolutely continuous, an assumption we make for all the random utility models we discuss later):

\[
    q_i(\mu) = P_{\epsilon \sim \theta} \left( i = \text{argmax}_{k \in \mathcal{N}} (\mu_k + \epsilon_k) \right).
\]

(2)

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

\[
    q_i^{\text{mnl}}(\mu) = \frac{\exp(\mu_i/\eta)}{\sum_{k \in \mathcal{N}} \exp(\mu_k/\eta)}.
\]

It can also be computed that the expected utility an individual can get under the MNL model is:

\[
    w^{\text{mnl}}(\mu) = E_{\epsilon \sim \theta} \left[ \max_{i \in \mathcal{N}} \mu_i + \epsilon_i \right] = \eta \log \left( \sum_{i \in \mathcal{N}} \exp(\mu_i/\eta) \right).
\]

The existence of closed-form formulae for the MNL model makes it a very popular choice model. We refer the readers to 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 model, there are other choices of the random part in (1) that lead to alternative choice models. Some popular ones among them are the probit model (in which \( \epsilon \) is chosen to be a joint normal distribution, see, e.g., Daganzo 1980), the nested logit model (in which \( \epsilon \) is chosen to be correlated general extreme value distributions, see, e.g., McFadden 1980), the mixed logit model (where \( \epsilon \) is chosen to be Gumbel distributions with a correlated term, see, e.g., McFadden and Train 2000, and Train 2009) and the exponomial choice model (in which \( \epsilon \) is chosen to be negative exponential distributions, see Alptekinoglu and Temple 2013).
2.2. Representative Agent Model

Another popular way to model choice 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 choice is a vector $x = (x_1, ..., x_n)$ on $\Delta_{n-1}$. To make his 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:

$$\max_{x \in \Delta_{n-1}} \mu^T x - V(x).$$

Here $\mu = (\mu_1, ..., \mu_n)$ is the deterministic utility of each alternative, which is similar to that in the random utility model. $V(x) : \mathbb{R}^n \mapsto \mathbb{R}$ is a regularization term that rewards diversification. Later, we denote the optimal value of (3) by $w^r(\mu)$, which is the utility a representative agent can obtain if the deterministic utility vector is $\mu$. Moreover, if for any $\mu$, there is a unique solution to (3), then we define

$$q^r(\mu) = \arg \max \left\{ \mu^T x - V(x) \mid x \in \Delta_{n-1} \right\}$$

to be the choice probability vector given by the representative agent model. (To ensure the maximum is attainable, $V(x)$ is required to be lower semi-continuous. We make this assumption for all the representative agent models in our discussions.)

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

$$q^\text{mnl}(\mu) = \arg \max \left\{ \mu^T x - \eta \sum_{i=1}^{n} x_i \log x_i \mid x \in \Delta_{n-1} \right\}.$$
They also show that the reverse statement of Proposition 1 is not true:

**Proposition 2** (Proposition 2.2 in Hofbauer and Sandholm 2002). When \( n \geq 4 \), there does not exist a random utility model that is equivalent to the representative agent model with \( V(x) = -\sum_{i=1}^{n} \log x_i \).

Based on the above two propositions, we know that the representative agent model strictly subsumes the random utility model as a special case.

### 2.3. Semi-Parametric Choice Model

Recently, a new class of semi-parametric choice model (SCM) was proposed by Natarajan et al. (2009). Unlike the random utility model where a certain distribution of the random utility \( \epsilon \) is specified, in the semi-parametric choice model, one considers a set of distributions \( \Theta \) for \( \epsilon \). Given the deterministic utility vector \( \mu \), one defines the maximum expected utility function \( w^*(\mu) \) as follows:

\[
w^*(\mu) = \sup_{\theta \in \Theta} \mathbb{E}_{\epsilon \sim \theta} \left[ \max_{i \in N} \mu_i + \epsilon_i \right].
\]

(5)

Note that in the random utility model, the maximum expected utility function can be defined in a similar way, but only with a single distribution \( \theta \). Thus the semi-parametric choice model can be viewed as an extension of the random utility model. Let \( \theta^*(\mu) \) denote the extreme distribution (or a limit of a sequence of distributions) that attains the optimal solution in (5). The choice probability for alternative \( i \) under this model is given by (provided it is well-defined):

\[
q^*_i(\mu) = \mathbb{P}_{\theta^*(\mu)} \left( i = \arg\max_{k \in N} (\mu_k + \epsilon_k) \right).
\]

(6)

Several special cases of semi-parametric choice models 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 \( \Theta \) contains all the distributions that have certain marginal distributions. The following proposition proved in Natarajan et al. (2009) shows that the marginal distribution model can be equivalently represented by a representative agent model:

**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:

\[
w^*(\mu) = \max_{x} \left\{ \mu^T x + \sum_{i=1}^{n} \int_{1-x_i}^{1} F_i^{-1}(t) dt \bigg| x \in \Delta_{n-1} \right\}.
\]

(7)

Furthermore, the choice probabilities \( q^*(\mu) \) can be obtained as the optimal solution \( x^* \) in (7).
Another semi-parametric model is the marginal moment model (MMM), in which only the first and second moments of the marginal distributions are known and $\Theta$ comprises all distributions that are consistent with the marginal moments. Natarajan et al. (2009) show that the MMM can also be represented as a representative agent model (without loss of generality, we assume that the marginal mean of $\epsilon_i$ is 0 for all $i$):

**Proposition 4.** Suppose the marginal variance of $\epsilon_i$ is $\sigma_i$ for all $i$. Then we have

$$w^*(\mu) = \max_{x} \left\{ \mu^T x + \sum_{i=1}^{n} \sigma_i \sqrt{x_i(1-x_i)} \mid x \in \Delta_{n-1} \right\}. \tag{8}$$

Furthermore, the choice probabilities $q^*(\mu)$ can be obtained as the optimal solution $x^*$ in (8).

In order to incorporate covariance information, Mishra et al. (2012) further propose a complete moment model (CMM), in which $\Theta$ is the set of distributions with known first and second moments $\Sigma$ (covariance matrix). It is shown in Ahipasaoglu et al. (2013) that the CMM model can also be written as a representative agent model (again without loss of generality, we assume the first moments are 0):

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

$$w^*(\mu) = \max_{x} \left\{ \mu^T x + \text{trace} \left( \Sigma^{1/2} S(x) \Sigma^{1/2} \right)^{1/2} \mid x \in \Delta_{n-1} \right\}, \tag{9}$$

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

Thus, all semi-parametric models studied so far can be represented as representative agent models. In the next section, we will show that this is generally the case. Moreover, we show that in fact, the set of representative agent models is equivalent to that of semi-parametric models.

Before we end this section, we comment 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. 2013), 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). However, they are based on different ideas and are less related to our study. Therefore, we choose not to include a detailed review of those models in this paper.
3. Welfare-Based Choice Model

In this section, we propose a new framework for discrete choice models and show that it provides a way to unify the various choice models reviewed in Section 2. To introduce our new model, we first notice that although various choice models reviewed in Section 2 are based on different ideas, they are all essentially functions from a vector of utilities \( \mu \) to a vector of choice probabilities \( q(\mu) \). Moreover, each of these models allows a welfare function \( w(\mu) \) that captures the expected utility that an individual can get from the choice model, and the choice probability vector can be viewed as the gradient of \( w(\mu) \) with respect to \( \mu \). Our proposed model is based on these observations. In particular, we first construct a class of welfare functions by defining what properties such functions should satisfy. Then we discuss the relation between our model and the previous ones. We start by making the following definition:

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

1. (Monotonicity): For any \( \mu_1, \mu_2 \in \mathbb{R}^n \) and \( \mu_1 \geq \mu_2 \), \( w(\mu_1) \geq w(\mu_2) \);
2. (Translation Invariance): For any \( \mu \in \mathbb{R}^n \), \( t \in \mathbb{R} \), \( w(\mu + te) = w(\mu) + t \);
3. (Convexity): For any \( \mu_1, \mu_2 \in \mathbb{R}^n \) and \( 0 \leq \lambda \leq 1 \), \( \lambda w(\mu_1) + (1 - \lambda)w(\mu_2) \geq w(\lambda \mu_1 + (1 - \lambda)\mu_2) \).

In addition to the three properties, if \( w(\mu) \) is also differentiable, then we call \( w(\mu) \) 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, therefore, 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 welfare is higher when there is an alternative with high utility rather than several mediocre alternatives. This is again plausible in reality and as we will see later, all previously reviewed choice models satisfy this condition.

In the following, we show that a choice welfare function has two equivalent representations: a convex optimization representation and a semi-parametric representation. This result will be instrumental for us to derive the relations among choice models.

**Theorem 1.** The following statements are equivalent:
1  \( w(\mu) \) is a choice welfare function;

2 There exists a convex function \( V(x) : \Delta_{n-1} \mapsto \bar{\mathbb{R}} \) such that

\[
w(\mu) = \max \left\{ \mu^T x - V(x) \middle| x \in \Delta_{n-1} \right\} ; \tag{10}\]

3 There exists a distribution set \( \Theta \) such that

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

**Proof.** First we show that the \( w(\mu) \) defined in (10) and (11) are choice welfare functions. To see this, we note that the monotonicity and translation invariance properties are immediate from (10) and (11). For the convexity, we note that \( w(\mu) \) defined in (10) is the supremum of linear functions of \( \mu \) thus is convex in \( \mu \). In (11), for each \( \epsilon \), \( \max_{i \in N} \{\mu_i + \epsilon_i\} \) is a convex function in \( \mu \), and so is the expectation. Therefore, if \( w(\mu) \) is defined by (10) or (11), then it must be a choice welfare function.

Next we show the other direction. That is, if \( w(\mu) \) is a choice welfare function, then it can be represented in the form of (10) and (11). First we note that if a choice welfare function \( w(\mu) = \infty \) for some \( \mu \), then by the translation invariance property and the monotonicity property, it must be that \( w(\mu) = \infty \) for any \( \mu \). In that case, we can choose \( V(x) = -\infty \) and \( \Theta = \{\theta_\infty\} \) where \( \theta_\infty \) is a singleton distribution taking value on \((\infty, ..., \infty)\). Therefore, \( w(\mu) \) can be represented by (10) and (11) in that case. Similarly, if \( w(\mu) = -\infty \) for some \( \mu \), then it must be that \( w(\mu) = -\infty \) for all \( \mu \), and we can take \( V(x) = \infty \) and \( \Theta = \{\theta_\infty\} \), where \( \theta_\infty \) is a singleton distribution on \((-\infty, ..., -\infty)\). Therefore, \( w(\mu) \) can be represented in (10) and (11) in this case too.

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

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

\[
w(\mu) = \min_y \left\{ w(y) + \max_i \{\mu_i - y_i\} \right\} = \min_y \left\{ w(y) + \max_{x \in \Delta_{n-1}} (\mu - y)^T x \right\} .
\]

Here the first equality holds since for any \( y \), \( w(\mu) = w(\mu - \max_i \{\mu_i - y_i\} e) + \max_i \{\mu_i - y_i\} \) by the translation invariance property. Furthermore, by the monotonicity property, \( w(\mu - \max_i \{\mu_i - y_i\} e) \leq w(y) \) and the equality holds when \( y = \mu \).

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

\[ \inf y \max_{x \in \Delta_{n-1}} L(x, y) = \max_{x \in \Delta_{n-1}} \inf y L(x, y). \]

Therefore, we have:

\[ w(\mu) = \max_{x \in \Delta_{n-1}} \left\{ \mu^T x + \inf_{y} \{ w(y) - y^T x \} \right\} = \max_{x \in \Delta_{n-1}} \{ \mu^T x - V(x) \} \]

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

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

\[ w(\mu) \geq (\mu - z)^T d(z) + w(z), \quad \forall z \in \mathbb{R}^n, \]

where the equality holds when \( z = \mu \). Define \( l(z) = w(z) - z^T d(z) \). By reorganizing terms, we have

\[ w(\mu) = \sup_z \{ \mu^T d(z) + l(z) \}. \] (12)

Now we define the distribution set as follows: Let \( \Theta = \{ \theta z \mid z \in \mathbb{R}^n \} \), where \( \theta z \) is an \( n \)-point distribution with

\[ \mathbb{P}_{\theta z} (\epsilon = \epsilon_z^i) = d_i(z), \quad \text{for} \ i = 1, \ldots, n \]

where

\[ \epsilon_z^i(j) = \begin{cases} l(z) & \text{if } j = i \\ -\infty & \text{if } j \neq i. \end{cases} \]

That is, \( \epsilon_z \) is a vector of all \( -\infty \)'s except \( l(z) \) at the \( i \)th entry. Therefore, for any \( z \), we have

\[ \mathbb{E}_{\theta z} [\max_i \{ \mu_i + \epsilon_i \}] = \sum_{i=1}^n d_i(z)(\mu_i + l(z)) = \mu^T d(z) + l(z). \]
Then by (12), we have
\[
    w(\mu) = \sup_z \{\mu^T d(z) + l(z)\} = \sup_z \mathbb{E}_q[z, \max_i \{\mu_i + \epsilon_i\}] = \sup_{\theta \in \Theta} \mathbb{E}_\theta[\max_i \{\mu_i + \epsilon_i\}].
\]
Therefore, the theorem is proved.

Note that the above discussion focuses on the equivalent representations of the choice welfare function. 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 \(q(\mu) : \mathbb{R}^n \mapsto \Delta_{n-1}\), 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 \(w(\mu)\) is a differentiable choice welfare function. Then the welfare-based choice model derived from \(w(\mu)\) is defined by
\[
    q(\mu) = \nabla w(\mu).
\]
(13)

Note that when \(w(\cdot)\) is differentiable, we have \(\nabla w(\mu) \in \Delta_{n-1}\) by the translation invariance property of \(w(\mu)\). Therefore \(q(\mu)\) defined by (13) 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 \(f : X \mapsto \bar{\mathbb{R}}\) is proper if \(f(x) < \infty\) for at least one \(x \in X\) and \(f(x) > -\infty\) for all \(x \in X\).

**Definition 4 (Essentially Strictly Convex Function).** A proper convex function \(f\) on \(\mathbb{R}^n\) is essentially strictly convex if \(f\) is strictly convex on every convex subset of
\[
    \text{dom}(\partial f) = \{x | \partial f(x) \neq \emptyset\}.
\]
where \(\partial f(x)\) is the set of subgradients of \(f\) at \(x\), and \(\emptyset\) is the empty set.

Note that any strictly convex function is essentially strictly convex. Next we have the following theorem, whose proof is relegated to the Appendix:

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

1. There exists a differentiable choice welfare function \(w(\mu)\) such that \(q(\mu) = \nabla w(\mu)\);
2. There exists an essentially strictly convex function \(V(x)\) such that \(q(\mu) = \arg \max_\mu \left\{ \mu^T x - V(x) \middle| x \in \Delta_{n-1} \right\}\);
3. There exists a distribution set \(\Theta\) such that
\[
    q(\mu) = \nabla \mu \left\{ \sup_{\theta \in \Theta} \mathbb{E}_\theta[\max_{i \in N} \mu_i + \epsilon_i] \right\}.
\]
The next corollary follows immediately from Theorem 2 and Propositions 1 and 2.

**Corollary 1.** Let $q(\mu)$ be a random utility model with absolutely continuous distribution $\theta$ and $w(\mu)$ be the corresponding expected utility an individual can get under this model. Then $w(\mu)$ is a differentiable choice welfare function, and $q(\mu) = \nabla w(\mu)$. Moreover, the reverse statement is not true, i.e., there exists a differentiable choice welfare function $w(\mu)$ such that there is no random utility model that gives the choice probability $q(\mu) = \nabla w(\mu)$.

The significance of Theorems 1 and 2 is mainly twofold. First, we propose a new framework for discrete choice model, the welfare-based choice model, which is based on the desired functional properties of the expected utility function. With the help of the new framework, we establish the connection between two existing choice models, the representative agent model and the semi-parametric model. In particular, we show that those two classes of choice models are equivalent. This result explains the prior results that for every known semi-parametric model, there is a corresponding representative agent model. In addition, it asserts that the reverse is also true, which is quite surprising in some sense. Therefore, in terms of the scope of choice models that can be captured, those three models (the welfare-based choice model, the representative agent model and the semi-parametric model) are the same. We believe this result is useful for the theoretical study of discrete choice models.

Second, by establishing the equivalence of the three classes of choice models, we can allow more versatile ways to construct a choice model. In particular, we can pick any of the three representations to start with. For the welfare-based choice model, one needs to choose a choice welfare function $w(\mu)$ which satisfies the three conditions. For the representative agent model, one needs to choose a (strictly) convex regularization function. And for the semi-parametric model, one needs to choose a set of distributions. In different situations, it might be easier to use one representation than the other in order to capture certain properties of the choice model. In addition, by Corollary 1, the welfare-based choice model strictly subsumes the random utility model, thus it is possible to construct new choice models that have certain interesting properties that a random utility model could not accommodate. We will further study this issue in Sections 4 and 5.

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 $w(\mu)$ is called superlinear if there exist $b_i$, $i = 1, ..., n$, such that for any $\mu \in \mathbb{R}^n$:

$$w(\mu) \geq \mu_i + b_i, \ \forall \ i = 1, ..., n.$$
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 \( q : \mathbb{R}^n \mapsto \Delta_{n-1} \), the following statements are equivalent:

1. There exists a superlinear differentiable choice welfare function \( w(\mu) \) such that \( q(\mu) = \nabla w(\mu) \);
2. There exists an essentially strictly convex function \( V(x) \) that is upper bounded on \( \Delta_{n-1} \) such that
   
   \[
   q(\mu) = \arg \max \left\{ \mu^T x - V(x) \mid x \in \Delta_{n-1} \right\};
   \]
3. There exists a distribution set \( \Theta \) containing only distributions with finite expectation (i.e., \( E_{\theta} |\epsilon_i| < \infty \) for all \( i \) and \( \theta \in \Theta \)) such that
   
   \[
   q(\mu) = \nabla \mu \left\{ \sup_{\theta \in \Theta} E_{\theta} \left[ \max_{i \in N} \mu_i + \epsilon_i \right] \right\},
   \]

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

We present the proof of Theorem 3 in the Appendix. We can see that Theorem 3 further develops the equivalence of choice models obtained in Theorem 2 by narrowing down the discussion to welfare-based choice models with the desirable superlinear property. In particular, we find that a superlinear differentiable choice welfare function has a semi-parametric representation, of which the distribution set contains at least one bounded distribution. The distribution set containing bounded distribution is also desirable due to its potential practical application. The last statement that \( q(\mu) \) 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 \( q(\mu) \) defined from the RUM or the MDM can span the whole simplex. Theorem 3 extends these results to more general conditions.

**4. Relation to the Random Utility Model**

In the last section, we proposed a new framework for choice models: the welfare-based choice model. In particular, by Corollary 1, the class of welfare-based choice models strictly subsumes the random utility model. In this section, we investigate further the relation between the welfare-based choice model and the random utility model. In particular, we study under what conditions a welfare-based choice model can be equivalently represented by a random utility model. This study will help us
understand clearly the relations between various choice models and the random utility model and 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 random utility models is equivalent to the class of welfare-based choice models.

**Theorem 4.** For any differentiable choice welfare function $w(\mu_1, \mu_2)$, there exists a distribution $\theta$ of $\{\epsilon_1, \epsilon_2\}$ such that:

$$w(\mu_1, \mu_2) = \mathbb{E}_\theta[\max\{\mu_1 + \epsilon_1, \mu_2 + \epsilon_2\}].$$

(14)

In addition, if $w(\mu_1, \mu_2)$ is superlinear, then there exists a distribution $\theta$ with finite expectation (i.e., $\mathbb{E}_\theta|\epsilon_1| < \infty$ and $\mathbb{E}_\theta|\epsilon_2| < \infty$) that satisfies (14).

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

$$v'(x) = \frac{\partial w}{\partial \mu_1}(x, 0).$$

Since $w(\mu_1, \mu_2)$ is convex and satisfies the translation invariance property, we have $v'(x) \in [0, 1]$ and is increasing. We define a distribution $\theta$ 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 $\xi$ is a random variable with c.d.f. $F_\xi(x) = \mathbb{P}(\xi \leq x) = v'(x)$. 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 $\mathbb{E}_\theta[\max\{\mu_1 + \epsilon_1, \mu_2 + \epsilon_2\}]$. We have

$$\mathbb{E}_\theta[\max\{\mu_1 + \epsilon_1, \mu_2 + \epsilon_2\}] = \mu_1 + v_0 + \mathbb{E}_\theta[\max\{-\max\{\xi, 0\}, \mu_2 - \mu_1 - \max\{-\xi, 0\}\}]$$

$$= \mu_1 + v_0 + \mathbb{E}_\theta[\max\{0, \mu_2 - \mu_1 + \xi\} - \max\{\xi, 0\}],$$

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

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

$$\mathbb{E}_\theta[\max\{0, x + \xi\} - \max\{0, \xi\}] = x \mathbb{P}(\xi > 0) + \mathbb{E}_\theta[(x + \xi) \cdot \mathbb{I}(-x < \xi \leq 0)]$$

$$= x \mathbb{P}(\xi > 0) + \int_{-x}^{0} (x + \xi) dv'(\xi)$$

$$= x(1 - v'(0)) + (x + \xi)v'(\xi) \big|_{-x}^{0} - \int_{-x}^{0} v'(\xi) d\xi$$

$$= x - v_0 + v(-x).$$
Similarly, for $x \leq 0$, we have
\[
E_\theta[\max\{0, x + \xi\} - \max\{0, \xi\}] = x P(\xi > -x) + E_\theta[-\xi \cdot 1(0 < \xi \leq -x)] \\
= x P(\xi > -x) - \int_0^{-x} \xi dv'(\xi) \\
= x(1 - v'(-x)) - \xi v'(-x) \mid_0^{-x} + \int_0^{-x} v'(\xi)d\xi \\
= x - v_0 + v(-x).
\]
Therefore, $E_\theta[\max\{\mu_1 + \epsilon_1, \mu_2 + \epsilon_2\}] = \mu_1 + v_0 + (\mu_2 - \mu_1) - v_0 + v(\mu_1 - \mu_2) = w(\mu_1, \mu_2).

To prove the last statement, it suffices to show that both $E_\theta[\max\{0, \xi\}]$ and $E_\theta[\max\{0, -\xi\}]$ are finite if $w(\mu)$ 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:
\[
E_\theta[\max\{0, \xi\}] = \int_0^{+\infty} P_\theta(\xi \geq t) dt = \int_0^{+\infty} (1 - v'(t))dt = (t - v(t)) \mid_0^{+\infty} = v(0) - L_1,
\]
and
\[
E_\theta[\max\{0, -\xi\}] = \int_0^{+\infty} P_\theta(-\xi \geq t) dt = \int_0^{+\infty} v'(-t) dt = \int_{-\infty}^{0} v'(t) dt = v(0) - L_2.
\]
Thus, the theorem is proved. \hfill \Box

By Proposition 2 when $n \geq 4$, the welfare-based choice model strictly subsumes the random utility model. In fact, as we will see in some examples later (Examples 2 and 3 in Section 5), this is also true for $n = 3$. In light of this relation between these two classes of choice models, it would be interesting to know the exact difference between them. In other words, it would be interesting to know what property is restricted in the random utility model but not in the welfare-based choice model. In the following, we pinpoint this difference. The following result is a direct consequence of the result in [McFadden 1980]:

**Proposition 6.** Let $w(\mu) : \mathcal{R}^n \mapsto \mathcal{R}$ be a differentiable function. Then $\nabla w(\mu)$ is consistent with a random utility model if and only if $w(\cdot)$ satisfies the monotonicity, translation invariance, convexity properties, and for any $k \geq 1$ and $i_1, ..., i_k$ all distinct,
\[
(-1)^k \frac{\partial^k w(\mu)}{\partial \mu_{i_1} ... \partial \mu_{i_k}} \leq 0.
\]

By Proposition 6 and the above discussions, we point out that the difference between a random utility model and a welfare-based choice model (thus also the representative agent model and the semi-parametric model by Theorem 2) lies in the requirement on the higher-order derivatives of
In particular, a random utility model requires that the higher-order cross-partial derivatives of \( w(\cdot) \) have alternating signs, while in the welfare-based choice model, it only requires that the Hessian matrix of \( w(\cdot) \) be positive semidefinite, and there is no requirement on other higher-order derivatives. This difference will enable us to better understand the difference between those models and later construct choice models with new properties.

We next consider an important subclass of the random utility model: the generalized extreme value (GEV) model. The GEV model was first proposed by McFadden et al. [1978]. It is a special case of the random utility model in which the random part of the utility \( \epsilon_i \)s take a joint generalized extreme value distribution. The GEV model covers various popular models, including the MNL model, the nested logit model, etc. An equivalent definition of the GEV model is given as follows (McFadden [1980]):

**Definition 6 (GEV model).** A choice model \( q(\mu) \) is a GEV model if and only if there exists a function \( H(y) : \mathbb{R}^n_+ \rightarrow \mathbb{R} \) such that

\[
q(\mu) = \eta \nabla_{\mu} \log H(e^{\mu_1}, \ldots, e^{\mu_n}),
\]

where \( H(y) \) satisfies the following properties:

1. \( H(y) \geq 0 \) for all \( y \in \mathbb{R}^n_+ \).
2. \( H(y) \) is homogeneous of degree \( 1/\eta \), i.e., \( H(\alpha y) = \alpha^{1/\eta} H(y) \).
3. \( H(y) \rightarrow \infty \) as \( y_j \rightarrow \infty \) for any \( j \).
4. The \( k \)th-order cross-partial derivatives of \( H(y) \) exist for all \( 1 \leq k \leq n \), and for all distinct \( i_1, \ldots, i_k \),

\[
(-1)^k \frac{\partial^k H(y)}{\partial y_{i_1} \cdots \partial y_{i_k}} \leq 0.
\]

Under appropriate specifications of \( H(\cdot) \), various known choice models can be obtained from the GEV model. We list the MNL model and the nested logit model as examples (Train [2009]).

- **MNL model.** If one chooses \( H(y) = \sum_{i \in N} y_i^{1/\eta} \), then the corresponding choice model is the MNL model with choice probabilities:

\[
q_i(\mu) = \frac{\exp(\mu_i/\eta)}{\sum_{k \in N} \exp(\mu_k/\eta)}.
\]

- **Nested Logit model.** Suppose the \( n \) alternatives are partitioned into \( K \) nests labeled \( B_1, \ldots, B_K \). If one chooses \( H(y) = \sum_{l=1}^{K} \left( \sum_{i \in B_l} y_i^{1/\lambda_l} \right)^{\lambda_l} \), then the corresponding choice model is the nested logit model with choice probabilities:

\[
q_i(\mu) = \frac{\exp(\mu_i/\lambda_k) \left( \sum_{j \in B_k} \exp(\mu_j/\lambda_k) \right)^{\lambda_k-1}}{\sum_{l=1}^{K} \left( \sum_{j \in B_l} \exp(\mu_j/\lambda_l) \right)^{\lambda_l}}.
\]
Since the welfare-based choice model strictly subsumes the random utility model, we know that the GEV model can be equivalently represented by welfare-based choice models. By \( q(\mu) = \nabla_{\mu} w(\mu) = \eta \nabla_{\mu} \log H(e^{\mu_1}, \ldots, e^{\mu_n}) \), it is implied that a GEV model derived from a specific \( H(\cdot) \) is equivalent to a welfare-based choice model with welfare function \( w(\mu) = \eta \log H(e^{\mu_1}, \ldots, e^{\mu_n}) \) and such \( w(\cdot) \) must satisfy the properties in Proposition 6.

Conversely, for any \( w(\mu) \) being a differentiable choice welfare function, we can define a function \( H(y) = \exp(w(\log y_1, \ldots, \log y_n)/\eta) \). The following discussions point out what properties such an \( H(\cdot) \) would satisfy:

1. By definition, \( H(z) \geq 0 \) for all \( z \).
2. Since \( w(\cdot) \) is translation invariant, we have that \( H(\alpha z) = \alpha^{1/n} H(z) \), i.e., \( H(z) \) is homogeneous of degree \( 1/\eta \).
3. Since \( w(\cdot) \) is monotone, we have
   \[
   \frac{\partial w(\mu)}{\partial \mu_i} = \frac{\eta \exp(\mu_i) H^{(1)}_i(z)}{H(z)} \geq 0, \quad \forall i \in \mathcal{N}.
   \]
   where \( z = (e^{\mu_1}, \ldots, e^{\mu_n}) \) and \( H^{(1)}_i(\cdot) \) is the partial derivative of \( H \) with respect to \( i \). Therefore, all first-order partial derivatives of \( H(z) \) are non-negative.
4. Last, in order for \( w(\cdot) \) to be convex, we need the Hessian matrix of \( w(\cdot) \) defined as follows to be positive semidefinite:
   \[
   \frac{\partial^2 w(\mu)}{\partial \mu_i \partial \mu_j} = \frac{\eta \exp(\mu_i + \mu_j) \left( H(z) \cdot H^{(2)}_{ij}(z) - H^{(1)}_i(z) \cdot H^{(1)}_j(z) \right)}{H^2(z)} \quad \text{and} \quad \frac{\partial^2 w(\mu)}{\partial \mu_i^2} = -\sum_{j \neq i} \frac{\partial^2 w(\mu)}{\partial \mu_i \partial \mu_j}.
   \]
   where \( z = (e^{\mu_1}, \ldots, e^{\mu_n}) \), \( H^{(1)}_i(\cdot) \) is the partial derivative of \( H(\cdot) \) with respect to \( i \), and \( H^{(2)}_{ij}(\cdot) \) is the second-order partial derivative of \( H(\cdot) \) with respect to \( i \) and \( j \).

It is worth pointing out that the last condition holds if all second-order cross-partial derivatives of \( H \) are negative, but the reverse is not necessarily true (the equivalent condition involves all the zero-, first- and second-order derivatives of \( H \)). Therefore, the GEV model requires an even stronger condition that the higher-order derivatives of \( \exp(w(\log y_1, \ldots, \log y_n)/\eta) \) have alternating signs, while in the welfare-based choice model, we only need the first-order derivative to be positive and some condition that is weaker than requiring all the cross second-order derivatives be negative.

5. **Substitutability and Complementarity of Choices**

In the previous section, we have seen that the distinction between the welfare-based choice model and the random utility model lies in the property of the higher-order derivatives of the choice
welfare function. In particular, the random utility model has stronger requirements on the higher-order derivatives. In this section, we will discuss more in depth about the practical meaning of such properties. We introduce two concepts, which we call the _substitutability_ and _complementarity_ of choices and then discuss the practical relevance of these two concepts. We show that if a choice model is derived from a random utility model, then the alternatives can only exhibit substitutability. However, using our welfare-based framework, we can design choice models that have more flexible substitutability or complementarity patterns. We also show how this property can be reflected through the regularization function in a representative agent model. Before we formally define these two concepts, we first introduce the definition of local monotonicity:

**Definition 7 (Local Monotonicity).** A function \( f(x) : \mathbb{R} \rightarrow \mathbb{R} \) is locally increasing at \( x \) if there exists \( \delta > 0 \) such that

\[
f(x - h) \leq f(x) \leq f(x + h), \quad \forall \, 0 < h < \delta.
\]

Similarly, \( f(x) \) is locally decreasing at \( x \) if there exists \( \delta > 0 \) such that

\[
f(x - h) \geq f(x) \geq f(x + h), \quad \forall \, 0 < h < \delta.
\]

Now we introduce the definition of substitutability and complementarity in choice models:

**Definition 8.** Consider a choice model \( q(\mu) : \mathbb{R}^n \rightarrow \Delta_{n-1} \). For any fixed \( \mu \) and \( i, j \in \mathbb{N} \):

1. **(Substitutability)** If \( q_j(\mu) \) is locally decreasing in \( \mu_i \) at \( \mu \), then we say alternative \( i \) is substitutable to alternative \( j \) at \( \mu \). Furthermore, if \( q_j(\mu) \) is locally decreasing in \( \mu_i \) for all \( \mu \), then we say alternative \( i \) is substitutable to alternative \( j \);

2. **(Complementarity)** If \( q_j(\mu) \) is locally increasing in \( \mu_i \) at \( \mu \), then we say alternative \( i \) is complementary to alternative \( j \) at \( \mu \). Furthermore, if \( q_j(\mu) \) is locally increasing in \( \mu_i \) for all \( \mu \), then we say alternative \( i \) is complementary to alternative \( j \);

3. **(Substitutable Choice Model)** For all \( i \neq j \), if alternative \( i \) is substitutable to alternative \( j \), then we say \( q(\mu) \) is a substitutable choice model.

The definition of substitutability and complementarity of two alternatives is similar to that of two consumer goods (see Mankiw 1997). However, in Definition 8 the independent variable is not the price, and the dependent variable is choice probability rather than demand. We first investigate some basic properties of substitutability and complementarity.

**Proposition 7.** Consider a choice model \( q(\mu) : \mathbb{R}^n \rightarrow \Delta_{n-1} \) that is derived from a differentiable choice welfare function \( w(\mu) \). For any \( i \), alternative \( i \) must be complementary to itself. Furthermore,
if \( w(\mu) \) is second-order continuously differentiable and alternative \( i \) is substitutable (complementary, resp.) to alternative \( j \) at \( \mu \), then alternative \( j \) must be substitutable (complementary, resp.) to alternative \( i \) at \( \mu \).

The proof of Proposition \( \text{[6]} \) uses some basic properties of continuous and convex function and is delegated to the Appendix. The proposition shows that when \( w(\mu) \) is second-order continuously differentiable, the substitutability (complementarity, resp.) property is a reciprocal property. In these cases, we shall say \( i \) and \( j \) are substitutable (complementary, resp.) in the following discussions.

In the following, we investigate the substitutability and complementarity of choice models. First we show that random utility models are all substitutable:

**Theorem 5.** Any random utility model \( q(\mu) \) is a substitutable choice model.

Theorem \( \text{[5]} \) directly follows from Proposition \( \text{[6]} \). It states that in a random utility model, if the utility of one alternative increases while the utilities of all other alternatives stay the same, then it must be that the choice probabilities of all other alternatives decrease. This is certainly plausible in practice, especially if \( \mu \) is interpreted as how much a consumer values each product. However, as we show in the following example, sometimes it might be desirable to allow different alternatives to exhibit certain degrees of complementarity. This is especially true if we allow more versatile meanings of the utility \( \mu \).

**Example 1.** Suppose a customer is considering to buy a camera from the following three alternatives: a Canon-A model, a Canon-B model and a Sony-C model. On a certain website, there are customer reviews that rate each model, which we denote by \( v_1 \), \( v_2 \) and \( v_3 \), respectively. We assume that the customer’s choice is solely based on those review scores (suppose other factors are fixed). That is, the choice probability \( q \) is a function of \( v = (v_1, v_2, v_3) \). Suppose at a certain time, a new review for the Canon-A model comes in, rating it favorably. How would it change the purchase probability of the Canon-B model?

The answer to the above questions may depend. There might be two forces. On one hand, due to a new favorable rating given to the Canon brand, the probability of choosing the Canon-B model might increase. On the other hand, the favorable rating for the Canon-A model might switch some customers from the Canon-B model to the Canon-A model. Either force might be dominant in practice. If the former force is stronger, then it is plausible that one additional favorable rating for the Canon-A model might increase the choice probability of the Canon-B model.

The above example illustrates that sometimes it might be desirable to have a choice model in which a certain pair of alternatives exhibit complementarity. One may notice that the above
example may be reminiscent of the nested logit model, in which the customers first choose a nest (in this case, the brand), and then choose a particular product. When increasing the utility of another product in the same nest, the tradeoff is between the probability of choosing the nest (which will be higher) and the individual product (which will be lower). However, we note that the nested logit model is essentially a random utility model (with the randomness $\epsilon$ chosen to be an extreme value distribution). Therefore, it is impossible to capture complementarity between alternatives through a nested logit model. Next, we show that we can capture the substitutability/complementarity of alternatives through our welfare-based framework.

In the following discussion, we only consider choice models $q(\mu)$ that are derived from differentiable choice welfare functions $w(\mu)$. We study necessary and sufficient conditions on the model parameters for a choice model to be substitutable. We first review the concepts of supermodularity and submodularity:

**Definition 9 (Supermodularity and Submodularity).** A function $f: \mathbb{R}^n \mapsto \mathbb{R} \cup \{\infty\}$ is called supermodular if for any $x, y \in \mathbb{R}^n$, $f(x \vee y) + f(x \wedge y) \geq f(x) + f(y)$, where $x \vee y$ and $x \wedge y$ denote the componentwise maximum and minimum of $x$ and $y$, respectively. A function $f: \mathbb{R}^n \mapsto \mathbb{R} \cup \{-\infty\}$ is called submodular if $-f$ is supermodular.

We have the following theorem:

**Theorem 6.** Consider a choice model $q(\mu): \mathbb{R}^n \mapsto \Delta_{n-1}$ that is derived from a differentiable choice welfare function $w(\mu)$. Then

1. $q(\mu)$ is a substitutable choice model if and only if $w(\mu)$ is submodular.
2. If $q(\mu)$ is a substitutable choice model, then there exists an essentially strictly convex $V(\cdot)$ with $\bar{V}_i(\cdot)$ supermodular on $\mathbb{R}^{n-1}$ for all $i$, such that

$$q(\mu) = \arg \max \left\{ \mu^T x - V(x) \middle| x \in \Delta_{n-1} \right\},$$

where

$$\bar{V}_i(z) = \begin{cases} V \left( z_1, z_2, \ldots, z_{i-1}, 1 - \sum_{j=1}^{n-1} z_j, z_i, \ldots, z_{n-1} \right), & \text{if } e^T z \leq 1 \text{ and } z \geq 0, \\ +\infty, & \text{otherwise.} \end{cases} \tag{16}$$

Furthermore, the reverse is true if $n = 3$.

We present the proof of Theorem 6 in the Appendix. Theorem 6 provides some sufficient and necessary conditions for $q(\mu)$ to be substitutable. We note that the supermodularity of $\bar{V}_i$ has nothing to do with the supermodularity of $V$. In fact, since $V(x)$ is only meaningful on $\Delta_{n-1}$, it can always be modified to be supermodular by defining $V(x) = +\infty$ for all $x \notin \Delta_{n-1}$. The definition of $\bar{V}_i(\cdot)$ reduces a redundant variable in $V$, making the operations “$x \vee y$” and “$x \wedge y$” meaningful.
Next we provide an easy-to-check sufficient condition for a substitutable choice model. We note that in the MDM and the MMM introduced in Propositions 3 and 4, the corresponding $V(\cdot)$s are separable. The following theorem shows that the choice models derived from such $V(\cdot)$s are always substitutable:

**Theorem 7.** If $V(x) = \sum_{i \in N} V_i(x_i)$ on $\Delta_{n-1}$ where $V_i(x_i) : [0,1] \to \mathbb{R}^n$ is a strictly convex function for all $i \in N$. Then $q(\mu)$ defined by

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

is a substitutable choice model.

Another possible choice of $V(\cdot)$ is a quadratic function. In that case, we have the following results:

**Theorem 8.** Consider $q(\mu) = \arg\max \{ \mu^T x - V(x) | x \in \Delta_{n-1} \}$, where $V(x) = x^T Ax$ is strictly convex with $A \succ 0$. Then $\bar{V}_i(z)$ for all $i \in N$ are supermodular if and only if $A_{jk} - A_{ik} - A_{ij} + A_{ii} \geq 0$ for all distinct $i, j, k \in N$, where $A_{ij}$ is the $(i,j)$-th entry of $A$.

Combining Theorems 6 and 8, we know that when $n = 3$ and $V(x) = x^T Ax$ with $A \succ 0$, the choice model defined by $q(\mu) = \arg\max \{ \mu^T x - V(x) | x \in \Delta_{n-1} \}$ is substitutable if and only if

$$A_{12} + A_{33} \geq A_{13} + A_{23}, \quad A_{13} + A_{22} \geq A_{12} + A_{23} \quad \text{and} \quad A_{23} + A_{11} \geq A_{12} + A_{13}.$$ 

Note that the above condition is different from $A$ being positive semidefinite. Indeed, the following example shows a case where the choice model is not substitutable even if $V(x)$ is strictly convex and supermodular:

**Example 2.** Consider $q(\mu) = \arg\max \{ \mu^T x - V(x) | x \in \Delta_{n-1} \}$, where $V(x) = x^T Ax$ with

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

It is easy to see that $V(x)$ is strictly convex and supermodular. However, it doesn’t satisfy that $A_{13} + A_{22} \geq A_{12} + A_{23}$. By some further calculations, we obtain that

$$\bar{V}_2(z) = z^T \begin{pmatrix} 2 \\ -1 \\ -2 \end{pmatrix} z - [-2; -2]^T z + 3,$$

which is not supermodular.

Therefore $q(\mu)$ is not a substitutable choice model by Theorem 6. In fact, when we fix $\mu_2 = \mu_3 = 0$ and plot the choice probabilities against $\mu_1$ in the range of values $[-2, 2]$ as shown in Figure 1 it is observed that $q_3$ increases in $\mu_1$ in the range of $[-1.5, -1]$, i.e., alternative 3 is complementary to alternative 1 in that interval. □
In addition to the quadratic example above, we can also easily generate a non-substitutable choice model through a proper choice welfare function $w(\cdot)$:

**Example 3.** Consider the following function:

$$w(\mu) = \log \left( e^{\mu_1} + e^{\mu_2} + e^{\mu_3} + e^{0.5(\mu_1+\mu_2)} \right).$$

It is easy to see that $w(\mu)$ is monotone, translation invariant and convex, therefore it is a choice welfare function. Also, it is differentiable. The corresponding choice probability is:

$$q(\mu) = \frac{1}{e^{\mu_1} + e^{\mu_2} + e^{\mu_3} + e^{0.5(\mu_1+\mu_2)}} \left( e^{\mu_1} + \frac{1}{2} e^{0.5(\mu_1+\mu_2)} \right).$$

Furthermore, the second-order derivative of $w(\mu)$ with respect to $\mu_1$ and $\mu_2$ is

$$\frac{\partial^2 w(\mu)}{\partial \mu_1 \partial \mu_2} = \frac{\partial q_1(\mu)}{\partial \mu_2} = \frac{\partial q_2(\mu)}{\partial \mu_1} = \frac{e^{0.5(\mu_1+\mu_2)}(-e^{\mu_1} - e^{\mu_2} + e^{\mu_3} - 4e^{0.5(\mu_1+\mu_2)})}{4(e^{\mu_1} + e^{\mu_2} + e^{\mu_3} + e^{0.5(\mu_1+\mu_2)})^2}.$$

It is positive if and only if $e^{\mu_3} \geq 4e^{0.5\mu_1+0.5\mu_2} + e^{\mu_1} + e^{\mu_2}$. Therefore, under this choice model, when $\mu_3$ is large enough (compared to $\mu_1$ and $\mu_2$), then alternatives 1 and 2 will exhibit complementarity. On the other hand, if $\mu_1$ or $\mu_2$ (or both) are comparable to $\mu_3$, then they will exhibit substitutability.

Now we give a plausible explanation for this model. Suppose as in Example 1, we explain $\mu$ as the number of positive reviews for each product. Then the above substitutability pattern could be reasonable if alternatives 1 and 2 are both from some relatively unknown brand (which has very few positive reviews in the history), while alternative 3 is from a well-known brand (which in contrast, has a lot of positive reviews). Then a few more positive reviews on either alternative 1 or 2 will be likely to positively impact the purchase probability of the other one, since it increases the overall attractiveness of this brand. On the other hand, if alternatives 1 and 2 have already gained
enough positive reviews in the past, then further increasing the number of positive reviews of one of them will be more likely to attract the demand from the other one, rather than from alternative 3.

To numerically illustrate the above model, we fix $\mu_2 = 0$, $\mu_3 = 3$ and plot the choice probability of alternative 2 as a function of $\mu_1$ in the range of $[-10,5]$ in Figure 2. From Figure 2 we can see that alternative 2 is complementary to alternative 1 when $\mu_1 \in [-10,2]$.

![Figure 2](image)

**Figure 2**  $q_2$ as a Function of $\mu_1$ in Example 3

The above two examples show that by using the welfare-based framework, it is possible to construct choice models with more versatile substitution patterns. In addition, the above two examples further verify that even when $n = 3$, we can construct choice models that do not have a random utility representation (remember all random utility models are substitutable choice models). Therefore, the welfare-based choice model (thus also the representative agent model and the semi-parametric models) strictly subsumes the random utility model, even for $n = 3$. This result is an extension of the result obtained by Hofbauer and Sandholm (2002), which only showed the result for $n \geq 4$.

### 6. Constructing New Choice Models from Existing Ones

In this section, we show that by using the welfare-based choice model framework, one can easily construct new choice models from existing ones. In particular, we provide three transformations below by which new choice models can be derived from existing ones. In the following discussions, we use $\tilde{q}(\cdot)$ to denote existing welfare-based choice models with choice welfare function $\tilde{w}(\cdot)$, and use $q(\cdot)$ and $w(\cdot)$ to denote the choice probability and the choice welfare function of the new model, respectively.
1. **Scaling.** Given any existing choice welfare function $\bar{w}(\cdot)$ and any $\eta > 0$, one can easily verify that $w(\mu) = \eta \bar{w}(\mu/\eta)$ is still a choice welfare function. The corresponding choice model is $q(\mu) = \bar{q}(\mu/\eta)$. We note that if $\bar{q}(\cdot)$ has an RUM representation, then $q(\cdot)$ also has an RUM representation with $\epsilon = \eta \bar{\epsilon}$. As $\eta$ becomes larger, the difference among $\mu/\eta$ is smaller and the choice will be more evenly distributed. As pointed out in Gigerenzer and Selten (2002), such $\eta$ can be used to model the level of rationality of the individual.

2. **Mixing.** Let $B_k, k = 1, \ldots, m$ be a cover of $N$, i.e., $\cup_k B_k = N$. Let $\bar{w}_k(\mu_k)$ be the choice welfare function on alternatives in $B_k$, with choice probabilities $\bar{q}_k(\mu_k)$. Define

$$w(\mu) = \sum_{k=1}^{m} \lambda_k \bar{w}_k(\mu_k),$$

where $\lambda_k \geq 0$, $\sum_{k=1}^{m} \lambda_k = 1$. We can verify that $w(\mu)$ is a choice welfare function and its corresponding choice model is

$$q_i(\mu) = \sum_{k: \mu_i \in B_k} \lambda_k \bar{q}_k(\mu_k), \quad \forall i \in N.$$

If $\bar{q}_k(\cdot)$ has an RUM representation for all $k$, then $q(\cdot)$ also has an RUM representation by assuming $\epsilon$ has a mixed distribution of $\epsilon_k$, each with probability $\lambda_k$. This model can be used to model choice scenarios where there are different segments of customers. Customers of different segments may only care about a subset of the products and choose according to a certain choice model. Then the mixed model is the choice model for the entire population.

3. **Crossing.** Let $A$ be an $m \times n$ matrix with $A_{ij} \geq 0$ and $Ae^n = e^m$, where $e^\ell$ refers to an $\ell$-dimensional column vector of ones. Given an existing choice welfare function $\bar{w}(\cdot)$ and its choice probabilities $\bar{q}(\cdot)$, we can easily verify that

$$w(\mu) = \bar{w}(A\mu)$$

is still a choice welfare function and the corresponding welfare-based choice model is

$$q(\mu) = \nabla_\mu w(\mu) = A^T \nabla \bar{w}(A\mu) = A^T \bar{q}(A\mu).$$

An example of such a transformation was in fact shown in Example 3, where $\bar{w}(\mu)$ is an MNL model for 4 alternatives with $\eta = 1$ and

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0.5 & 0.5 & 0 \end{bmatrix}.$$ 

In light of this example, we note that RUM is not closed under cross-transformation, i.e., even if $\bar{q}(\cdot)$ has an RUM representation, $q(\mu)$ may not. Thus, the cross-transformation provides us a way of generating new choice models.
7. Conclusion
In this paper, we proposed a new framework for discrete choice models: the welfare-based choice model, which is based on the idea of considering the expected utility an individual can get when facing a set of alternatives. We showed that the welfare-based choice model is equivalent to the representative agent model and the semi-parametric model, thus establishing the equivalence between the latter two. We also showed that the welfare-based choice model subsumes the random utility model by relaxing its requirement on properties of higher-order cross-partial derivatives of the choice welfare function. In particular, we showed that when there are only two alternatives, the welfare-based choice model is equivalent to the random utility model. We defined a new concept for choice models – substitutability and complementarity – and showed that under the new framework, we can construct choice models with complementary alternatives, thus enabling us to capture new choice patterns. We believe that this framework is useful for future studies of choice models.

References
Ahipasaoglu, S. D., X. Li, K. Natarajan. 2013. A convex optimization approach for computing correlated choice probabilities with many alternatives. Working paper.

Alptekinoglu, A., J. Temple. 2013. The exponential choice model: A new alternative for assortment and price optimization. Working paper.

Anderson, S. P., A. De Palma, J. F. Thisse. 1988. A representative customer theory of the logit model. International Economic Review 29(3) 461–466.

Anderson, S. P., A. De Palma, J. F. Thisse. 1992. Discrete Choice Theory of Product Differentiation. The MIT Press.

Ben-Akiva, M., S. R. Lerman. 1985. Discrete Choice Analysis: Theory and Application to Travel Demand. The MIT Press.

Bertsekas, D. 2003. Convex Analysis and Optimization. Athena Scientific.

Blanchet, J., G. Gallego, V. Goyal. 2013. A Markov chain approximation to choice modeling. Working paper.

Daganzo, C. 1980. Multinomial Probit: The Theory and Its Application to Demand Forecasting. Academic Press.

Farias, V. F., S. Jagabathula, D. Shah. 2013. A nonparametric approach to modeling choice with limited data. Management Science 59(2) 305–322.

Gallego, G., R. Ratliff, S. Shebalov. 2014. A general attraction model and sales-based linear program for network revenue management under customer choice. Operations Research.

Gigerenzer, G., R. Selten. 2002. Bounded Rationality: The Adaptive Toolbox. The MIT Press.
Hofbauer, J., W. H. Sandholm. 2002. On the global convergence of stochastic fictitious play. *Econometrica* 70(6) 2265–2294.

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

Mankiw, G. 1997. *Principles of Microeconomics*. Cengage Learning.

Mas-Colell, A., M. D. Whinston, J. R. Green. 1995. *Microeconomic Theory*. Oxford University Press.

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

McFadden, D. 1980. Econometric models for probabilistic choice among products. *The Journal of Business* 53(3) 13–29.

McFadden, D., K. Train. 2000. Mixed MNL models for discrete responses. *Journal of Applied Econometrics* 15 447–470.

McFadden, Daniel, et al. 1978. *Modelling the choice of residential location*. Institute of Transportation Studies, University of California.

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

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

Murota, K. 2003. *Discrete Convex Analysis*. Society for Industrial and Applied Mathematics.

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

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

Rockafellar, T. 1974. *Conjugate Duality and Optimization*. Society for Industrial and Applied Mathematics.

Simchi-Levi, D., X. Chen, J. Bramel. 2014. Convexity and supermodularity. *The Logic of Logistics*. Springer, 15–44.

Thurstone, L. 1927. A law of comparative judgment. *Psychological Review* 34(4) 273–286.

Train, K. E. 2009. *Discrete Choice Methods with Simulation*. Cambridge University Press.
Appendix

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

\[
w(\mu) = \max \left\{ \mu^T x - V(x) \mid x \in \Delta_{n-1} \right\},
\]

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

\[
\nabla w(\mu) = \nabla \mu \left( \mu^T x - V(x) \right) \big|_{x=x^*} = x^*,
\]

where \( x^* = \arg \max \left\{ \mu^T x - V(x) \mid x \in \Delta_{n-1} \right\} \). Therefore,

\[
q(\mu) = \nabla w(\mu) = \arg \max \left\{ \mu^T x - V(x) \mid x \in \Delta_{n-1} \right\}.
\]

Last, we show that 2 ⇒ 1. Given an essentially strictly convex \( V(x) \), by Theorem 1, we know that \( w(\mu) \) is a choice welfare function. Again, by Theorem 6.3 in Rockafellar (1974), we know that \( w(\mu) \) is essentially differentiable. Moreover, in our case, \( w(\mu) \) is a convex and finitely valued function in \( \mathbb{R}^n \), thus essentially differentiability is equivalent to differentiability. Again, by applying the envelope theorem, \( q(\mu) = \nabla w(\mu) \). 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 \( w(\mu) \) is superlinear if and only if \( V(x) \) defined by \( \max_y \{ y^T x - w(y) \} \) is upper bounded. If \( w(\mu) \) is superlinear, we have, for any \( x \in \Delta_{n-1}, \)

\[
w(\mu) \geq \sum_{i \in \mathcal{N}} x_i (\mu_i + b_i) = x^T \mu + x^T b \geq x^T \mu + \min_i \{ b_i \}.
\]

By reorganizing terms, we have

\[
x^T \mu - w(\mu) \leq - \min_i \{ b_i \} = \max_i \{ -b_i \}.
\]

Therefore, \( V(x) = \max_y \{ y^T x - w(y) \} \leq \max_i \{ -b_i \} \), i.e., \( V(x) \) is upper bounded.

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

\[
w(\mu) \geq \max \{ \mu^T x - u \mid x \in \Delta_{n-1} \} \geq \mu_i - u, \quad \forall i,
\]
i.e., \( w(\mu) \) 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 \( w(\mu) \), we can find a distribution set \( \Theta \) consisting of only distributions with finite expectation such that \( w(\mu) \) can be represented as \( w(\mu) = \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[ \max_{i \in N} \mu_i + \epsilon_i \right] \).

First, since \( w(\mu) \) is convex with \( q(\mu) = \nabla w(\mu) \), we have

\[
w(\mu) = \sup_z \left\{ \mu^T q(z) + l(z) \right\},
\]

where \( l(z) = w(z) - z^T q(z) \). Now we define a distribution set \( \Theta \) that is slightly different from that of Theorem\( \ref{thm:2} \). Specifically, let \( \Theta = \{ \theta_{z} \mid z \in \mathbb{R}^n \} \), where \( \theta_{z} \) is an \( n \)-point distribution with \( \mathbb{P}_{e} (\epsilon = e^i) = q_i(z), \forall i \in N \). (Note that by the monotonicity and the translation invariance properties, \( q(z) = \nabla w(z) \) must satisfy \( q(z) \geq 0 \) and \( q^T q(z) = 1 \).

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

where

\[
M(z) = \max \left\{ 1 + \max_{i,j} \{ z_i - z_j \}, \frac{l(z) - \min_i \{ b_i \}}{t^*(z)} \right\},
\]

with

\[
t^*(z) = \min \{ q_i(z) \mid q_i(z) > 0 \}.
\]

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

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

Next we show that:

\[
\mathbb{E}_{\theta_{z}} \left[ \max_{i} \mu_i + \epsilon_i \right] \leq w(\mu), \quad \forall \mu.
\]

For any given \( \mu \), define \( k(i) = \arg \max_j \{ \mu_j + \epsilon^i_{z}(j) \} \) (we break ties arbitrarily). There are two cases:

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

\[
\mathbb{E}_{\theta_{z}} \left[ \max_{j} \{ \mu_j + \epsilon_j \} \right] = \sum_{i \in N} q_i(z)(\mu_i + l(z)) = \mu^T q(z) + l(z) \leq w(\mu),
\]

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

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

\[
\mathbb{E}_{\theta_{z}} \left[ \max_{j} \mu_j + \epsilon_j \right] = \sum_{i \in N, q_i(z) > 0} q_i(z)(\mu_{k(i)} + l(z) - M(z) \mathbb{I}_{k(i) \neq i})
\]


\[ \leq \max_i \{\mu_i\} + l(z) - t^*(z)M(z) \]
\[ \leq \max_i \{\mu_i\} + \min_j \{b_j\} \]
\[ \leq \max_i \{\mu_i + b_i\} \]
\[ \leq w(\mu), \]

where the first inequality follows from the fact that \( M(z) > 0 \) and \( \sum_{i \in N} q_i(z) 1_{q_i(z) > 0, k(i) \neq i} \geq t^*(z) \), the second inequality is because of the definition of \( M(z) \) and the last inequality follows from the definition of superlinear function.

Based on the analysis of these two cases, we have

\[ \mathbb{E}_{\epsilon \sim \theta} \left[ \max_i \mu_i + \epsilon_i \right] \leq w(\mu), \quad \forall \mu. \]

Then by equation (18) we have

\[ w(\mu) = \sup_z \{\mu^T q(z) + l(z)\} = \sup_z \mathbb{E}_{\theta \in \Theta} \left[ \max_i \mu_i + \epsilon_i \right] = \sup_{\theta \in \Theta} \mathbb{E}_{\theta} \left[ \max_i \mu_i + \epsilon_i \right]. \]

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 \( \forall i \in N \), then for \( \mu \in \mathbb{R}^n \) we have

\[ \sup_{\theta \in \Theta} \mathbb{E}_{\theta} \left[ \max_{i \in N} \mu_i + \epsilon_i \right] \geq \mathbb{E}_{\hat{\theta}} \left[ \max_{i \in N} \mu_i + \epsilon_i \right] = \mathbb{E}_{\hat{\theta}} [\mu_j + \epsilon_j] = \mu_j + \mathbb{E}_{\hat{\theta}} [\epsilon_j], \quad \forall j. \]

Therefore we can conclude that \( w(\mu) = \sup_{\theta \in \Theta} \mathbb{E}_{\theta} \left[ \max_{i \in N} \mu_i + \epsilon_i \right] \) is superlinear.

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

\[ x \in \Delta_{n-1}^0 \triangleq \{x \mid e^T x = 1, x_i > 0, \forall i \in N\}, \]

there exists \( \mu_x \) such that \( q(\mu_x) = \nabla w(\mu_x) = x \). Fix \( x \in \Delta_{n-1}^0 \), we consider

\[ V(x) = \max_{\mu} \{\mu^T x - w(\mu)\}. \quad (21) \]

Clearly, \( V(x) \geq -w(0) \), since \( \mu = 0 \) is a feasible solution. Moreover, since \( w(\mu) \) is translation invariant, we can restrict the feasible region of (21) to \( L \triangleq \{\mu \mid e^T \mu = 0\} \). For all \( \mu \in L \), we have \( \mu_j \leq 0 \) for some \( j \in N \). Thus

\[ \mu^T x \leq \sum_{i \neq j} \mu_i x_i \leq \sum_{i \neq j} x_i \max_k \{\mu_k\} \leq (1 - \min \{x_i\}) \max_k \{\mu_k\}. \]

However, by superlinearity of \( w(\mu) \), we have:

\[ w(\mu) \geq \max_k \{\mu_k + b_k\} \geq \max_k \{\mu_k\} + \min \{b_k\}. \]
Thus, for all \( \mu \in \mathcal{L} \), we have:
\[
\mu^T x - w(\mu) \leq -\min_i \{x_i\} \max_k \{\mu_k\} - \min_k \{b_k\}.
\]

Let \( K = \frac{w(0) - \min_{x \in \mathcal{L}} \{b_k\}}{\min_{x \in \mathcal{L}} \{x_i\}} \). In order for \( \mu \) to be optimal to (21), by the above arguments, we would have \( \mu_i \leq K \) for all \( i \). Thus we can further restrict the feasible set of (21) to \( \{\mu \in \mathcal{L} \mid e^T \mu = 0, \mu_i \leq K \forall i \in N\} \), which is a compact set. Since \( w(\mu) \) is continuous, there exists \( \mu_x \in \{\mu \in \mathcal{L} \mid e^T \mu = 0, \mu_i \leq K \forall i \in N\} \) that attains maximum in problem (21). By the first-order necessary condition, \( \nabla w(\mu_x) = x \). This concludes the proof.

Proof of Proposition 7 Since \( w(\mu) \) is convex and differentiable, for any \( \mu \in \mathcal{L} \) and any \( t > 0 \), we have
\[
w(\mu + te_i) - w(\mu) \geq te_i^T \nabla w(\mu) = t q_i(\mu),
w(\mu) - w(\mu + te_i) \geq -te_i^T \nabla w(\mu + te_i) = -t q_i(\mu + te_i).
\]

From these two inequalities, we have \( q_i(\mu + te_i) - q_i(\mu) \geq 0 \), for all \( t > 0 \) and \( \mu \). Thus, alternative \( i \) is complementary to itself.

Furthermore, if \( w(\mu) \) is second-order continuously differentiable, then we have
\[
\frac{\partial w}{\partial \mu_i} = \frac{\partial^2 w}{\partial \mu_i \partial \mu_j},
\]
\[
\frac{\partial^2 w}{\partial \mu_i \partial \mu_j} = \frac{\partial q_j}{\partial \mu_i}.
\]

Thus, if alternative \( i \) is substitutable (complementary, resp.) to alternative \( j \) at \( \mu \), then alternative \( j \) is substitutable (complementary, resp.) to alternative \( i \) at \( \mu \).

Proof of Theorem 6 In this proof, we use the following lemma from [Murota (2003)].

**Lemma 1.** Let \( f : \mathcal{R}^n \rightarrow \mathcal{R} \cup \{\infty\} \) be a function such that there exists at least one \( \mu \) such that \( f(\mu) < \infty \). Let \( g(\mu) = \max_{\mu} \{\mu^T x - f(\mu)\} \) be the convex conjugate of \( f \). We have
1. If \( f \) is submodular, then \( g \) is supermodular.
2. If \( n = 2 \) and \( f \) is supermodular, then \( g \) is submodular.

Now we use this lemma to prove the theorem. To prove the first part, by [Simchi-Levi et al. (2014)]), a differentiable function \( w(\mu) \) is submodular in \( \mu \) if and only if \( \frac{\partial w(\mu)}{\partial \mu_i} \) is decreasing in \( \mu_j \) for all \( i \neq j \). By the definition of \( q(\mu) = \nabla w(\mu) \), the result holds.

For the second part, let \( V(x) = \max_{\mu} \{\mu^T x - w(\mu)\} \) be the convex conjugate of \( w(\mu) \). From Theorem 2 \( V(x) \) is essentially strictly convex and
\[
q(\mu) = \arg \max_{x \in \Delta_n} \{\mu^T x - V(x)\}.
\]

For any \( y \in \mathcal{R}^{n-1} \) and \( i \in N \), define \( f_i(y) = w(y_1, y_2, ..., y_{i-1}, 0, y_i, ..., y_{n-1}) \). Also define \( \mu_{-i} = (\mu_1, ..., \mu_{i-1}, \mu_{i+1}, ..., \mu_n) \), then we have
\[
\hat{V}_i(z) = \max_{\mu} \{\mu^T_i z + \mu_i (1 - e^T z - w(\mu))\}.
\]
where the second equality is due to the translation invariance property of $w(\mu)$. The submodularity of $w(\mu)$ implies the submodularity of $f_i(y)$ for all $i \in \mathcal{N}$. Thus $\bar{V}_i(z)$, as the convex conjugate of $f_i(y)$, is supermodular by Lemma 11.

For the last statement, since $V(\cdot)$ is an essentially strictly convex function, $q(\mu) = \arg \max \left\{ \mu^T x - V(x) \bigg| x \in \Delta_{n-1} \right\}$ is well-defined. By Theorem 2, $q(\mu) = \nabla w(\mu)$ where $w(\mu) = \sup \left\{ \mu^T x - V(x) \bigg| x \in \Delta_{n-1} \right\}$. For any $y \in \mathbb{R}^{n-1}$ and $i \in \mathcal{N}$, define $f_i(y) = w(y_1, y_2, ..., y_{i-1}, 0, y_i, ..., y_{n-1})$. Also define $x_{\sim i} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)$, then we have

$$f_i(y) = \max_{x \in \Delta_{n-1}} \left\{ x_i y + 0(1 - e_i^T x_{\sim i}) - V(x) \right\}$$

$$= \max_{x \in \Delta_{n-1}} \left\{ x_i y - \bar{V}_i(x_{\sim i}) \right\}$$

$$= \max_x \left\{ y^T x_{\sim i} - \bar{V}_i(x_{\sim i}) \right\}$$

$$= \max_z \left\{ y^T z - \bar{V}_i(z) \right\},$$

where the third equality holds since $\bar{V}_i(x_{\sim i}) = +\infty$ for all $x \not\in \Delta_{n-1}$. From Lemma 11 given that $n = 3$ and thus $y \in \mathbb{R}^2$, $f_i(y)$ is submodular. It remains to show that $w(\mu)$ is also submodular. According to Theorem 6, it suffices to show that $q_i(\mu)$ is locally decreasing with $\mu_j$ for all $j \neq i$ for all $\mu$. Fix $i, j$ and let $k \neq i, j$. We assume $i > j$ without loss of generality. We have $q_i(\mu - \mu_k e) = q_i(\mu)$ from translation invariance property. But $q_i(\mu - \mu_k e) = \frac{\partial f_k(\mu - \mu_k e, \mu_j - \mu_k)}{\partial \mu_i}$ is non-decreasing with $\mu_j$ due to the submodularity of $f_k$. Thus $w(\mu)$ is submodular and $q(\mu) = \nabla w(\mu)$ is a substitutable choice model.

**Proof of Theorem 7** We first consider the case where $V_i(x_i)$ is differentiable for all $i \in \mathcal{N}$. Let $\lambda(\mu)$ be the Lagrangian multiplier of the constraint $\sum_i x_i = 1$. The KKT conditions (see Bertsekas 2003) for problem 17 can be written as:

$$\mu_i - V_i'(q_i(\mu)) - \lambda(\mu) \leq 0, \ \forall i \in \mathcal{N};$$

$$\mu_i - V_i'(q_i(\mu)) - \lambda(\mu) = 0, \ \forall i \text{ s.t. } q_i(\mu) \neq 0;$$

$$q_i(\mu) \geq 0, \ \forall i \in \mathcal{N};$$

$$\sum_{i \in \mathcal{N}} q_i(\mu) = 1.$$  

Now we consider any two points $\mu_0$ and $\mu_0 + te_i$ where $e_i$ is a unit vector along the $i$-th coordinate axis and $t > 0$. Suppose that there exists a $j \neq i$ such that $q_j(\mu_0 + te_i) > q_j(\mu_0)$. Since $V_j$ is strictly convex, $V_j'(q_j(\mu_0 + te_i)) > V_j'(q_j(\mu_0))$. There are two possible cases for $q_j(\mu_0)$:
• $q_j(\mu_0) > 0$: In this case, we have $\mu_j - V_j'(q_j(\mu_0 + te_i)) - \lambda(\mu_0 + te_i) = 0$ and $\mu_j - V_j'(q_j(\mu_0)) - \lambda(\mu_0) = 0$, therefore, we have $\lambda(\mu_0 + te_i) < \lambda(\mu_0)$.

• $q_j(\mu_0) = 0$: In this case, $\mu_j - V_j'(q_j(\mu_0)) - \lambda(\mu_0) \leq 0$, which implies that $\mu_j - V_j'(q_j(\mu_0 + te_i)) - \lambda(\mu_0 + te_i) = 0$, we have $\lambda(\mu_0 + te_i) < \lambda(\mu_0)$.

In both cases, $\lambda(\mu_0 + te_i) < \lambda(\mu_0)$. This implies that $q_j(\mu_0 + te_i) \geq q_j(\mu_0)$ for all $j \neq i$. Note that we also have $q_i(\mu_0 + te_i) > q_i(\mu_0)$ by Proposition 7. Therefore, we have $\sum_{j \in \mathcal{N}} q_j(\mu_0 + te_i) > \sum_{j \in \mathcal{N}} q_j(\mu_0) = 1$, which contradicts with that $q(\mu_0 + te_i) \in \Delta_{n-1}$. Thus we have $q_j(\mu_0 + te_i) \leq q_j(\mu_0)$ for all $j \neq i$. Since this is true for all $\mu_0$ and $t > 0$, $q$ is substitutable.

If $V_i(x_i)$ is not differentiable, we need to replace the derivative with the subgradient in the above argument. Since $V_i$ is strictly convex, $g_1 > g_2$ for all $g_1 \in \partial V_i(x_1)$ and $g_2 \in \partial V_i(x_2)$ if $x_1 > x_2$, the above argument is still valid.

**Proof of Theorem 8**: For $i \in \mathcal{N}$, $\tilde{V}_i$ is an $n - 1$ variate quadratic function. Let $H^i$ denote the Hessian matrix of $\tilde{V}_i(z)$. For $j, k \in \{1, 2, ..., n - 1\}$ and $j \neq k$, the off-diagonal element $H^i_{j,k} = A^i_{\tilde{j},\tilde{k}} - A^i_{i,\tilde{k}} - A^i_{\tilde{j},i} + A^i_{i,i}$, where

$$\tilde{j} = \begin{cases} j, & \text{if } j < i, \\ j + 1, & \text{if } j \geq i; \end{cases} \quad \text{and} \quad \tilde{k} = \begin{cases} k, & \text{if } k < i, \\ k + 1, & \text{if } k \geq i. \end{cases}$$

Thus, $\tilde{V}_i(z)$ is supermodular if and only if $H^i_{j,k} \geq 0$ for all $j, k \in \{1, 2, ..., n - 1\}$ and $j \neq k$, which is equivalent to $A^i_{j,k} - A^i_{i,k} - A^i_{i,j} + A^i_{i,i} \geq 0$ for all distinct $i, j, k \in \mathcal{N}$. 

$\square$