Convex Optimization and Online Learning: Their Applications in Discrete Choice Modeling and Pricing

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Xiaobo Li

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advisors: Prof. Shuzhong Zhang and Prof. Zizhuo Wang

May, 2018

Acknowledgements

First of all, I would like to express my deep gratitude to my advisor Professor Shuzhong Zhang for his mentoring, support and guidance throughout my study at the University of Minnesota as a Ph.D. student. I have been constantly inspired by his deep understanding of optimization and related fields. I have learned a lot from Professor Zhang that would not have been learned if I were not a student of him. Besides Professor Zhang's technical guidance, I am very much motivated by his passion for research, academic integrity and endless innovative ideas. Thanks to Professor Zhang's patience and student-centric mentoring, I gradually developed my own research styles and built my own research path. I am also very grateful for Professor Zhang's invaluable suggestions and tremendous help when I face difficulties during my Ph.D. study. In addition, Professor Zhang has influenced my personable growth by sharing with me his philosophy and wisdom of life and knowledge of the world.

My great gratitude also goes to my co-advisor, Professor Zizhuo Wang, who provides me great academic support at the University of Minnesota. Without Professor Wang, this dissertation would not have been possible. I learned a lot from him on how to develop a topic, how to write an excellent paper, and how to better apply theories to real-world practice. For his efficient mode of working, broad research interest, great technical ability, and wide range of knowledge, Professor Wang have been an excellent academic role model for me to follow. I also benefit a lot from him through his great effort on paving the way for my academic journey, which includes, but is not limited to, paying for my conference presentations, referring me to world-class researchers, sharing with me updates on research trends and improving my writing skills.

I am very grateful to Professor Saif Benjaafar for his excellent guidance and support. Professor Benjaafar has greatly broadened my research vision on operations management

through his great course and our personal discussions. He never stops inspiring me with his great vision on many operations management areas. Moreover, Professor Benjaafar has shared a lot of resources and information with me for my academic life. Thanks to his persistent help and stimulating suggestions, I have been able to finish a paper with him. But my harvest of being closely connected with Professor Benjaafar is far more than that paper.

I would like to thank other faculty members in the Department of Industrial and Systems Engineering for creating a nice academic environment for my Ph.D. study. In particular, my special gratitude goes to Professor William Cooper, who provides me a lot of help as a Director of Graduate Studies. His help includes, but is not limited to, coordinating my courses, applying for scholarships/fellowships, polishing my documents for job searching, and serving as my thesis committee member. I am also grateful to Professor Tony Cui at Carlson School of Management, for devoting the time to be my thesis committee member, and for his brilliant comments and suggestions during my preliminary and final thesis defense.

My Ph.D. study would not have started without Professor Karthik Natarajan at Singapore University of Technology and Design (SUTD), who brought me into the operations research and robust optimization area. As my Master's advisor, Professor Natarajan has spent countless hours to help me with every single component of being a researcher, such as proving theorems, developing topics, finding research questions and writing papers. After my graduation as a Master student, Professor Natarajan kindly offered me a one-year research assistant position at SUTD to help me smoothly transition into the Ph.D. study. I feel very lucky to have finished several research projects with Professor Natarajan. During my Ph.D. study, I still receive a lot of help and feedback from him. I always find myself improved after every single discussion with him.

I would like to thank all my coauthors for their wonderful ideas, great effort, and excellent teamwork. Chapter 3 is based on a joint work with Professor Karthik Natarajan and Professor Selin Damla Ahipasaoglu. Chapter 4 is based on a joint work with Professor Zizhuo Wang and Guiyun Feng. Chapter 5 is based on a joint work with Professor Shuzhong Zhang and Xiang Gao. Other coauthors of mine include Professor Chung-Piaw Teo, Professor Xuan Vinh Doan, Professor Zhichao Zheng and Dr. Xiang

Li. Here, I would like to thank Professor Ahipasaoglu additionally for her support and help during my stay in Singapore.

My special gratitude also goes to Professor Simai He, Professor Bo Jiang, Professor Andrew Lim, Professor Guangwu Liu, Professor Rowan Wang and Professor Yimin Yu, for their help, suggestions, and encouragements for my graduate studies in Hong Kong, Singapore and Minneapolis.

Last but not least, I am very grateful to my parents, Jialin Li and Qiuling Liao, and grandparents, Yuzhen Li and Xiafang Guo, for their unconditional love, support and trust. Finally, I owe my sincere appreciation to my wife, Guiyun Feng, for her constant love and company. Without her, my life would not have been such colorful and enjoyable.

Dedication

To my parents, my grandparents and my wife.

Abstract

The discrete choice model has been an important tool to model customers' demand when facing a set of substitutable choices. The random utility model, which is the most commonly used discrete choice framework, assumes that the utility of each alternative is random and follows a prescribed distribution. Due to the popularity of the random utility model, the probabilistic approach has been the major method to construct and analyze choice models. In recent years, several choice frameworks that are based on convex optimization are studied. Among them, the most widely used frameworks are the representative agent model and the semi-parametric choice model. In this dissertation, we first study a special class of the semi-parametric choice model – the cross moment model (CMM) – and reformulate it as a representative agent model. We also propose an efficient algorithm to calculate the choice probabilities in the CMM model. Then, motivated by the reformulation of the CMM model, we propose a new choice framework – the welfare-based choice model – and establish the equivalence between this framework and the other two choice frameworks: the representative agent model and the semi-parametric choice model. Lastly, motivated by the multi-product pricing problem, which is an important application of discrete choice models, we develop an online learning framework where the learning problem shares some similarities with the multi-product pricing problem. We propose efficient online learning algorithms and establish convergence rate results for these algorithms. The main techniques underlying our studies are continuous optimization and convex analysis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Discrete choice models are widely used to describe people's behaviors when they have to make choices 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 much research attention in 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 one is the random utility model, in which a utility value is assigned to each alternative. In the random utility model, the utility value is composed of a deterministic part and a random part. Each individual then chooses the alternative with the highest utility value, 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 value of the choice (which is a vector of proportions for each alternative) plus a regularization term, which typically encourages diversification of the choice (see e.g. Anderson et al. 1988). More recently,

a class of semi-parametric choice models has been proposed (see Natarajan et al. 2009). This model is similar to the random utility model in that it also assumes that the utility value is composed of a deterministic part and a random part. 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 Chapter 2.

In the classical discrete choice modeling literature, convex optimization is mainly applied to the estimation of choice models. For example, McFadden et al. (1973) showed that the maximum likelihood estimation (MLE) under the multinomial logit choice model is a convex program. Daganzo and Kusnic (1993) showed that the MLE under the nested logit model is also a convex program if the scale parameters are fixed. In recent years, there have been numerous research papers which study the applications of convex optimization in discrete choice modeling. For instance, Evgeniou et al. (2007) proposed a semidefinite program based convex optimization formulation to address customer heterogeneity in choice estimation; Natarajan et al. (2009), Mishra et al. (2012) proposed several classes of semi-parametric models and reformulated the models into convex programs; Mishra et al. (2014) studied the marginal distribution model (MDM), a special class of the semi-parametric choice model, and proved the convexity of the MLE under various MDM models with some mild conditions.

In this dissertation, we further apply convex optimization to several different aspects of discrete choice modeling, the details of which are listed in Section 1.1. These approaches of applying convex optimization provide some new insights on how the development of convex optimization theories can help the development of discrete choice modeling.

1.1 Organization

In Chapter 2, we review three important classes of discrete choice models: the random utility model, the representative agent model and the semi-parametric choice model. Though proposed many years ago, the representative agent models and the

semi-parametric models have not been studied or applied satisfactorily in the academic literature, as compared to the random utility models. A thorough understanding of these models is an underlying motivation of this dissertation.

In Chapter 3, we propose a representative agent representation for the CMM model, which provides a different angle for the CMM model, rather than just an approximation of the multinomial probit model. More importantly, we find some desirable properties of the objective function in the representative agent form. With these properties, we prove that in terms of computing the choice probabilities, the gradient ascent algorithm with inexact line search enjoys local linear convergence. Numerical results show that the new algorithm can compute the choice probabilities for a choice set of a size up to several thousands, while the existing approach could not deal with the choice set as large as 200. More justifications on the CMM model when the number of alternatives is large are provided.

In Chapter 4, we propose a welfare-based framework to analyze the discrete choice model. The welfare-based framework could be viewed as a new choice framework that is derived from a welfare (potential) function. We show that it is equivalent to the representative agent model and the semi-parametric choice model, thus establishing the equivalence between the latter two existing choice frameworks. The equivalence can explain the known fact that most existing semi-parametric models can be reformulated as representative agent models. We further study the relations between these choice frameworks and the random utility model. We find that these frameworks are indeed more flexible in modeling the complementarity behavior. Moreover, it provides a theoretical justification for the general nested logit model where some scale parameters may exceed one, which is known to be inconsistent with the random utility model.

In Chapter 5, we consider the online learning problem where the loss functions could be non-convex. We apply the non-stationary regret as the performance metric. We propose different algorithms under different assumptions on the information available regarding the loss functions. A sublinear regret bound for online learning with non-convex loss functions and non-stationary regret measure is established.

1.2 Notation

Throughout the dissertation, the following notations will be used. We use $\mathbb{R}$ to denote the set of real numbers, and $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ 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 $\boldsymbol{x} \geq \boldsymbol{y}$ to denote a componentwise relationship and Δ_{n-1} to denote the $(n-1)$-dimensional simplex, i.e., $\Delta_{n-1} = \{\boldsymbol{x} \mid \mathbf{e}^T\boldsymbol{x} = 1, \boldsymbol{x} \geq \mathbf{0}\}$. In our discussions, ordinary lowercase letters $x, y, \ldots$ denote scalars, boldfaced lowercase letters $\boldsymbol{x}, \boldsymbol{y}, \ldots$ denote vectors.

Chapter 2

Discrete Choice Models

In this chapter, we review several prevailing classes of discrete choice models that are related to the study presented in this proposal.

2.1 The Random Utility Model

Perhaps the most popular class of discrete choice models is the random utility model (RUM), proposed first in 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 in a choice set denoted by $\mathcal{N} = \{1, 2, ..., n\}$, then the random utility model assumes that the utility of alternative i takes the following form:

$$u_i = \mu_i + \epsilon_i, \quad \forall i \in \mathcal{N}, \tag{2.1}$$

where $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)$ is the deterministic part of the utility and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)$ is the random part. In the random utility model, it is assumed that the joint distribution θ

of $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)$ is known. Then the probability that alternative i will be chosen is:

$$q_i(\boldsymbol{\mu}) = \mathbb{P}_{\boldsymbol{\epsilon}\sim\theta}\left(i = \underset{k\in\mathcal{N}}{\operatorname{argmax}}\ (\mu_k + \epsilon_k)\right), \quad \forall i \in \mathcal{N}. \tag{2.2}$$

To ensure the above equation is well-defined, we assume θ is absolutely continuous, an assumption we make for all the random utility models we discuss later.

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 in McFadden (1974). The MNL model is derived by assuming that $(\epsilon_1, ..., \epsilon_n)$ follow independent and identically distributed Gumbel distributions with scale parameter η. Given that assumption, the choice probability in (2.2) can be further written as follows:

$$q_i^{\text{mnl}}(\boldsymbol{\mu}) = \frac{\exp(\mu_i/\eta)}{\sum\limits_{k\in\mathcal{N}} \exp(\mu_k/\eta)}, \quad \forall i \in \mathcal{N}.$$

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

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

The existence of closed-form formula 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 properties of the MNL model.

While the MNL choice probability is known in closed form and possesses desirable properties such as concavity of the log-likelihood function, it also suffers from drawbacks. One of the well-known properties of MNL is the Independence of Irrelevant Alternatives (IIA) property which implies that the ratio of the choice probabilities for any two alternatives is independent of the utilities of the other alternatives:

$$\frac{p_i^{\text{mnl}}}{p_j^{\text{mnl}}} = e^{\mu_i - \mu_j}, \quad \forall i \neq j.$$

When the alternatives have correlated utilities, the IIA property of MNL gives rise to

misleading choice predictions.

A popular choice model that can incorporate correlation information is the Generalized Extreme Value (GEV) model, which is a generalization of the MNL model. The GEV model was first proposed in McFadden (1978). The random part of the utility vector $\boldsymbol{\epsilon}$ is assumed to follow the generalized extreme value distribution. The GEV model includes various popular models, including the MNL model, the nested logit model, etc. An equivalent definition of the GEV model is given as follows (see McFadden 1980):

Definition 1 (GEV model) *A choice model $\boldsymbol{q}(\boldsymbol{\mu})$ is a GEV model if and only if there exists a function $H(\boldsymbol{y}) : \mathbb{R}_+^n \mapsto \mathbb{R}$ such that*

$$\boldsymbol{q}(\boldsymbol{\mu}) = \eta \nabla_{\boldsymbol{\mu}} \log H(e^{\mu_1}, \ldots, e^{\mu_n}), \tag{2.3}$$

where $H(\boldsymbol{y})$ satisfies the following properties:

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

$$(-1)^k \frac{\partial^k H(\boldsymbol{y})}{\partial y_{i_1} ... \partial y_{i_k}} \leq 0.$$

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 (see Train 2009).

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

$$q_i(\boldsymbol{\mu}) = \frac{\exp(\mu_i/\eta)}{\sum_{k \in \mathcal{N}} \exp(\mu_k/\eta)}, \quad \forall i \in \mathcal{N}.$$

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

$$q_i(\boldsymbol{\mu}) = \frac{\exp(\mu_i/\lambda_k)(\sum_{j \in B_k} \exp(\mu_j/\lambda_k))^{\lambda_k - 1}}{\sum_{l=1}^{K} \left(\sum_{j \in B_l} \exp(\mu_j/\lambda_l) \right)^{\lambda_l}}, \quad \forall i \in \mathcal{N}.$$

Although the correlation structure of the random term in the GEV model is flexible, one can not easily incorporate the correlation information in the GEV model. A model that accounts for any valid correlation matrix is the Multinomial Probit (MNP) model in which $\boldsymbol{\epsilon}$ is assumed to be normally distributed with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, namely $\boldsymbol{u} \sim \text{Normal}\,(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The MNP model is flexible in terms of modeling dependence and does not possess the IIA property. However the choice probabilities do not have a closed-form expression and Monte Carlo simulation is the most commonly used method to find the choice probabilities. The reader is referred to Hajivassiliou et al. (1996) for an in-depth discussion of simulation techniques used to approximate the choice probabilities in MNP models with the Geweke-Hajivassiliou-Keane (GHK) simulator being the most commonly used technique among them (see Geweke 1989, Hajivassiliou and McFadden 1998, Keane 1994) .

Besides these models, there are other choices of the random part in (2.1) that lead to alternative choice models. The examples includes the mixed logit model (where $\boldsymbol{\epsilon}$ 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 $\boldsymbol{\epsilon}$ is chosen to be negative exponential distributions, see Alptekinoğlu and Semple 2016).

2.2 The 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 $\boldsymbol{x} = (x_1, ..., x_n)$ on Δ_{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:

$$w^r(\boldsymbol{\mu}) = \text{maximize}_{\boldsymbol{x}\in\Delta_{n-1}} \quad \boldsymbol{\mu}^T\boldsymbol{x} - V(\boldsymbol{x}). \tag{2.4}$$

Here $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)$ is the deterministic utility of each alternative, which is similar to that in the random utility model and $V(\boldsymbol{x}) : \Delta_{n-1} \mapsto \mathbb{R}$ is a regularization term that rewards diversification. We denote the optimal value of (2.4) by $w^r(\boldsymbol{\mu})$, which is the utility a representative agent can obtain if the deterministic utility vector is $\boldsymbol{\mu}$. In this dissertation, without loss of generality, we assume $V(\boldsymbol{x})$ is convex and lower semi-continuous. [1] Moreover, if for any $\boldsymbol{\mu}$, there is a unique solution to (2.4), then we define

$$\boldsymbol{q}^r(\boldsymbol{\mu}) = \arg\max \left\{ \boldsymbol{\mu}^T\boldsymbol{x} - V(\boldsymbol{x}) \middle| \boldsymbol{x} \in \Delta_{n-1} \right\} \tag{2.5}$$

to be the choice probability vector given by the representative agent model.

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 η can be equally derived from a representative agent model with $V(\boldsymbol{x}) = \eta \sum_{i=1}^n x_i \log x_i$. Or equivalently, we can write

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

In Hofbauer and Sandholm (2002), the authors further extend the result to general random utility models. They show that for any random utility model with continuously distributed random utility, there exists a representative agent model that yields the same choice probability. The precise statement of their result is as follows:

Proposition 2.2.1 *Let $\boldsymbol{q}(\boldsymbol{\mu}) : \mathbb{R}^n \mapsto \Delta_{n-1}$ be the choice probability function defined in (2.2) where the random vector $\boldsymbol{\epsilon}$ admits a strictly positive density on $\mathbb{R}^n$ and the*

[1] If $V(\boldsymbol{x})$ is not convex or lower semi-continuous, then we can replace $V(\boldsymbol{x})$ by a convex and lower semi-continuous function $V^{**}(\boldsymbol{x}) = \sup_{\boldsymbol{y}}\{\boldsymbol{y}^T\boldsymbol{x} - w^r(\boldsymbol{y})\}$ and the equation (2.4) still holds (see Borwein and Lewis 2010). Therefore, it is without loss of generality to assume that $V(\boldsymbol{x})$ is convex and lower semi-continuous.

function $\boldsymbol{q}(\boldsymbol{\mu})$ *is continuously differentiable. Then for all* $\boldsymbol{\mu}$ *there exists* $V(\cdot)$ *such that:*

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

They also show that the reverse statement of Proposition 2.2.1 is not true:

Proposition 2.2.2 (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* $V(\boldsymbol{x}) = -\sum_{i=1}^{n} \log x_i$*.*

It follows from the above two propositions that the representative agent model strictly subsumes the random utility model as a special case.

Recently, inspired from the entropy formulation for the MNL model, Fudenberg et al. (2015) have proposed the use of a general additive perturbation function in the representative agent model as a simple and tractable approach to model choice under uncertainty. Under this approach, the choice probability vector is defined as the optimal solution to the following problem:

$$\boldsymbol{q} = \operatorname{argmax} \left\{ \boldsymbol{\mu}^T \boldsymbol{x} - \sum_{i \in \mathcal{N}} V_i(x_i) \;\middle|\; \boldsymbol{x} \in \boldsymbol{\Delta}_{n-1} \right\}, \tag{2.6}$$

where the functions $V_i(\cdot)$s are assumed to be strictly convex in the interval $[0, 1]$, continuously differentiable in $(0, 1)$ with $\lim_{x_i \to 0} V_i'(x_i) = -\infty$. The entropy maximization problem is a special case of this model. Under the assumptions on the functions $V_i(\cdot)$, the choice probability vector is unique and lies in the relative interior of the simplex (see Rockafellar 1970). Furthermore, Fudenberg et al. (2015) provide an axiomatic justification of formulation (2.6) by showing its equivalence to two conditions, one condition generalizing an acyclicity condition derived from the strong axiom of revealed preferences and the second condition generalizing Luce's IIA condition. A weaker form of the perturbation function by relaxing the condition that $\lim_{x_i \to 0} V_i'(x_i) = -\infty$ is also studied in Fudenberg et al. (2015), which allows for some choice probabilities to take a value of 0. Fudenberg et al. (2015) show that such framework rules out some random utility models even with independent and identically distributed random terms. Natarajan et al. (2009) provide an alternative justification for formulation (2.6) by relaxing the

standard assumption that the joint distribution of the random components in the utility model is known, as we shall discuss in the next section.

2.3 The Semi-Parametric Choice Model

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

$$w^s(\boldsymbol{\mu}) = \sup_{\theta \in \Theta} \ \mathbb{E}_{\boldsymbol{\epsilon} \sim \theta} \left[\max_{i \in \mathcal{N}} \ \mu_i + \epsilon_i \right]. \tag{2.7}$$

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 θ. Thus the semi-parametric choice model can be viewed as an extension of the random utility model. Let $\theta^*(\boldsymbol{\mu})$ denote the distribution (or a limit of a sequence of distributions) that attains the optimal value in (2.7). The choice probability for alternative i under this model is given by (provided it is well-defined):

$$q_i^s(\boldsymbol{\mu}) = \mathbb{P}_{\theta^*(\boldsymbol{\mu})} \left(i = \operatorname*{argmax}_{k \in \mathcal{N}} \ (\mu_k + \epsilon_k) \right). \tag{2.8}$$

Several special cases of semi-parametric choice models have been studied recently. One such model, called the marginal distribution model (MDM), is proposed in Natarajan et al. (2009). In the MDM, the distribution set Θ 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 2.3.1 *Suppose $\Theta = \{\theta \mid \epsilon_i \sim F_i(\cdot), \forall i\}$ where $F_i(\cdot)$s are given continuous distribution functions. Then we have:*

$$w^s(\boldsymbol{\mu}) = \max_{\boldsymbol{x}} \ \left\{ \boldsymbol{\mu}^T \boldsymbol{x} + \sum_{i=1}^{n} \int_{1-x_i}^{1} F_i^{-1}(t) dt \ \middle| \ \boldsymbol{x} \in \Delta_{n-1} \right\}. \tag{2.9}$$

Furthermore, the choice probabilities $\boldsymbol{q}^s(\boldsymbol{\mu})$ in (2.8) can be obtained as the optimal solution $\boldsymbol{x}^*$ in (2.9).

Another semi-parametric model proposed by Natarajan et al. (2009) 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. In Natarajan et al. (2009), the authors show that the MMM can also be represented as a representative agent model (without loss of generality, we assume that the marginal mean of ϵ_i is 0 for all i):

Proposition 2.3.2 *Suppose the marginal standard deviation of ϵ_i is σ_i for all i. Then we have*

$$w^s(\boldsymbol{\mu}) = \max_{\boldsymbol{x}} \left\{ \boldsymbol{\mu}^T \boldsymbol{x} + \sum_{i=1}^{n} \sigma_i \sqrt{x_i(1-x_i)} \;\middle|\; \boldsymbol{x} \in \Delta_{n-1} \right\}. \tag{2.10}$$

Furthermore, the choice probabilities $\boldsymbol{q}^s(\boldsymbol{\mu})$ can be obtained as the optimal solution $\boldsymbol{x}^*$ in (2.10).

In order to incorporate covariance information, Mishra et al. (2012) propose the cross moment model (CMM), in which Θ is the set of distributions with mean zero and known covariance matrix $\boldsymbol{\Sigma}$. In this choice model, the joint distribution of $\boldsymbol{\epsilon}$ is assumed to be only partially specified to the modeler. Specifically, the available information on the joint distribution is the first two moments of $\boldsymbol{\epsilon}$. Let $\boldsymbol{\epsilon} \sim_\theta (\mathbf{0}, \boldsymbol{\Sigma})$ denote the set of probability distributions for $\boldsymbol{\epsilon}$ that satisfies the following two conditions: $\mathbb{E}_\theta[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{Cov}_\theta[\boldsymbol{\epsilon}] = \boldsymbol{\Sigma}$. The modeler is then assumed to solve the optimization problem:

$$\text{(CMM)} \quad w^{cmm} = \max_{\boldsymbol{\epsilon} \sim_\theta (\mathbf{0}, \boldsymbol{\Sigma})} \mathbb{E}_\theta \left(\max_{i \in \mathcal{N}} (\mu_i + \epsilon_i) \right). \tag{2.11}$$

The outer optimization in (2.11) is over all joint distributions of the random components that are consistent with the two moment information. Hence, problem (2.11) is equivalent to finding a joint distribution for the random components that maximizes the expected agent utility[2].

[2]The problem of finding the joint distribution of the random components that minimizes the expected agent utility with the first two moment information reduces to Jensen's bound. This is uninteresting

Mishra et al. (2012) solve the moment problem (2.11) by reformulating it as the following semidefinite program:

$$\begin{aligned} \text{(CMM)} \quad w^{cmm} = \max \quad & \sum_{i\in\mathcal{N}} \boldsymbol{e}_i^T \boldsymbol{v}_i \\ \text{s.t.} \quad & \sum_{i\in\mathcal{N}} \begin{pmatrix} \boldsymbol{W}_i & \boldsymbol{v}_i \\ \boldsymbol{y}_i^T & x_i \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T & \boldsymbol{\mu} \\ \boldsymbol{\mu}^T & 1 \end{pmatrix} \\ & \begin{pmatrix} \boldsymbol{W}_i & \boldsymbol{v}_i \\ \boldsymbol{v}_i^T & x_i \end{pmatrix} \succeq 0 \qquad\qquad \forall i \in \mathcal{N}, \end{aligned} \tag{2.12}$$

where $\boldsymbol{e}_i$ is a vector with 1 in the ith position and 0 otherwise. Let $\{\boldsymbol{W}_i^*, \boldsymbol{v}_i^*, x_i^*\}$ for $i \in \mathcal{N}$ be an optimal solution to the above semidefinite program. The joint distribution of the random utilities $\boldsymbol{u}$ that maximizes the expected utility is a mixture of multivariate normal distributions given as:

$$\boldsymbol{u} = \left\{ \text{ Normal} \left(\frac{\boldsymbol{v}_i^*}{x_i^*}, \frac{\boldsymbol{W}_i^*}{x_i^*} - \frac{\boldsymbol{v}_i^*}{x_i^*}\frac{\boldsymbol{v}_i^{*T}}{x_i^*} \right), \quad \text{with probability } x_i^*, \quad \forall i \in \mathcal{N}. \right\} \tag{2.13}$$

More importantly, they show that the optimal decision vector $\boldsymbol{x}^*$ in the SDP formulation is the choice probability vector for the mixture of multivariate normal distributions in (2.13) which maximizes the expected agent utility. Mishra et al. (2012) provide applications of this formulation to problems in route choice, random walk theory and product line selection with the number of alternatives up to a hundred. Numerical experiments in Mishra et al. (2012) show that the CMM model captures correlation information in predicting choices and provides insights often qualitatively similar to MNP. Natarajan and Teo (2017) propose an equivalent SDP formulation with smaller size. The details would be reviewed in next chapter.

2.4 Other Choice Models

Before we end this chapter, 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. 2016), the two-stage choice model (see Jagabathula

from a discrete choice modelling perspective since all the agents then choose the alternative with the highest mean.

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 suit different purposes. In particular, they do not take the form of mapping a utility vector to a choice probability vector. Thus we choose not to discuss those models in this dissertation.

Chapter 3

The Reformulation of the CMM Choice Model

As shown in Chapter 2, a popular discrete choice model that incorporates correlation information is the Multinomial Probit (MNP) model where the random utilities of the alternatives are chosen from a multivariate normal distribution. Computing the choice probabilities is challenging in the MNP model when the number of alternatives is large and simulation is a popular technique used to approximate the choice probabilities. As described in Section 2.3, Mishra et al. (2012) have recently proposed a semidefinite optimization approach to compute choice probabilities for the distribution of the random utilities that maximizes expected agent utility given only the mean and covariance information. Their model is referred to as the Cross Moment (CMM) model. Computing the choice probabilities with many alternatives is challenging in the CMM model since one needs to solve large scale semidefinite programs. In this chapter, we develop a simpler formulation of CMM as a representative agent model, which maximizes over the choice probabilities in the unit simplex where the objective function is the sum of the expected utilities and a strongly concave perturbation function. By characterizing the perturbation function for the CMM model and its gradient, we develop a simple first-order gradient method with inexact line search to compute choice probabilities. We establish local linear convergence of this algorithm under mild assumptions on the

choice probabilities. An implication of our results is that inverting the choice probabilities to compute the mean utilities is straightforward given any positive definite covariance matrix. Numerical experiments show that this method can compute the choice probabilities for a large number of alternatives within a reasonable amount of time. Numerical results also show CMM explicitly captures the correlation information, regardless of the number of alternatives. Comparisons with simulation methods for MNP and semidefinite programming methods for CMM indicate the efficacy of the method.

3.1 A Representative Agent Formulation for the Cross Moment (CMM) Model

Recall that Mishra et al. (2012) solve the moment problem (2.11) by reformulating it as the following semidefinite program:

$$\begin{aligned}
\text{(CMM)} \quad w^{cmm} = \max \quad & \sum_{i\in\mathcal{N}} \boldsymbol{e}_i^T \boldsymbol{v}_i \\
\text{s.t.} \quad & \sum_{i\in\mathcal{N}} \begin{pmatrix} \boldsymbol{W}_i & \boldsymbol{v}_i \\ \boldsymbol{y}_i^T & x_i \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T & \boldsymbol{\mu} \\ \boldsymbol{\mu}^T & 1 \end{pmatrix} \\
& \begin{pmatrix} \boldsymbol{W}_i & \boldsymbol{v}_i \\ \boldsymbol{v}_i^T & x_i \end{pmatrix} \succeq 0 \qquad\qquad \forall i \in \mathcal{N},
\end{aligned} \tag{3.1}$$

where $\boldsymbol{e}_i$ is a vector with 1 in the ith position and 0 otherwise. Natarajan and Teo (2017) further reduce the size of formulation (3.1) and show that it can be equivalently computed as the following formulation:

$$\begin{aligned}
\text{(CMM)} \quad w^{cmm} = \max \quad & \text{trace}(\boldsymbol{Y}) \\
\text{s.t.} \quad & \boldsymbol{x} \in \Delta_{n-1} \\
& \begin{pmatrix} \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T & \boldsymbol{Y}^T & \boldsymbol{\mu} \\ \boldsymbol{Y} & \text{Diag}(\boldsymbol{x}) & \boldsymbol{x} \\ \boldsymbol{\mu}^T & \boldsymbol{x}^T & 1 \end{pmatrix} \succeq 0,
\end{aligned} \tag{3.2}$$

where trace($\boldsymbol{Y}$) is the trace of the matrix $\boldsymbol{Y}$ and Diag($\boldsymbol{x}$) is a diagonal matrix with the entries of $\boldsymbol{x}$ along the diagonal. For completeness, we provide the proof of the equivalence of the semidefinite programs (3.1) and (3.2) in the Section 3.5. Since the multivariate normal distribution is a feasible distribution in the CMM formulation, w^{cmm} is an upper bound on the expected consumer utility in MNP. Computationally these models differ in the way the choice probabilities are computed. In the MNP model, simulation techniques are usually used to compute the choice probabilities. On the other hand, the CMM model uses convex optimization techniques to solve the semidefinite program.

There has been an increasing interest in the literature on discrete choice models that deal with a large number of alternatives. Examples that have been studied includes the choice of lake recreation sites in the state of Wisconsin involving 589 alternatives (see Parsons and Kealy 1992), choice of car models involving 689 alternatives (see Brownstone et al. 2000) and choice of messenger bags involving 3584 alternatives (see Toubia et al. 2003). Models that treat products as bundles of characteristics with an additive error term that accounts for variation in the taste for the products in conjunction with variation in taste for the characteristics of the products results in choices where the number of products (alternatives) is exponential in the number of characteristics. In terms of CMM, although the size of the SDP is reduced, the computation of formulation (3.2) is still challenging when the size of alternatives is large. In a recent paper, Ahipaşaoğlu et al. (2015) used the CMM model as an alternative to MNP for computing choice probabilities in a traffic equilibrium problem and showed that it provides a practical alternative to MNP in estimating traffic flows. The correlation information in their model arises from origin-destination paths (alternatives) sharing common roads (characteristics). The number of paths in such networks might be exponential in the number of roads. In our computational experiments, we have found that solving the semidefinite program (3.2) using state of art interior point method based solvers such as SDPNAL+ version 0.3 (see Yang et al. 2015) in MATLAB R2014 on a laptop with an Intel(R) i7-5600U CPU processor (2.6 GHz) with 4GB RAM works well when the number of alternatives is up to two hundred roughly. Solving large semidefinite programs with matrix size up to a few thousands still remains a computational challenge and is a subject of intense research in the optimization community.

3.1.1 Optimization over the Unit Simplex

In this subsection, we develop a representative agent formulation for the CMM model that transforms the semidefinite program to a nonlinear maximization problem over the unit simplex.

Theorem 3.1.1 *Assume that* $\boldsymbol{\Sigma} \succ 0$. *Then the maximum expected consumer utility* w^{cmm} *in the CMM model is the optimal objective value to the following nonlinear optimization problem over the unit simplex:*

$$\textit{(CMM)} \quad w^{cmm} = \max \left\{ \boldsymbol{\mu}^T \boldsymbol{x} + \textit{trace} \left(\left(\boldsymbol{\Sigma}^{1/2} \ \boldsymbol{S}(\boldsymbol{x}) \ \boldsymbol{\Sigma}^{1/2} \right)^{1/2} \right) \ \Big| \ \boldsymbol{x} \in \Delta_{n-1} \right\}, \quad (3.3)$$

where $\boldsymbol{S}(\boldsymbol{x}) = \textit{Diag}(\boldsymbol{x}) - \boldsymbol{x}\boldsymbol{x}^T \succeq 0$ *and* $\boldsymbol{B} = \boldsymbol{A}^{1/2}$ *is the unique positive semidefinite square root of a matrix* $\boldsymbol{A} \succeq 0$ *such that* $\boldsymbol{A} = \boldsymbol{B}^2$. *Furthermore the optimal decision vector* $\boldsymbol{x}^*$ *is the choice probability vector for the distribution that maximizes the expected agent utility.*

Proof:
Applying Schur's lemma to the positive semidefinite matrix in formulation (3.2), we obtain the equivalent nonlinear semidefinite program:

$$\begin{aligned} w^{cmm} = \max \quad & \text{trace}(\boldsymbol{Y}) \\ \text{s.t.} \quad & \boldsymbol{x} \in \Delta_{n-1} \\ & \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{Y}^T - \boldsymbol{\mu}\boldsymbol{x}^T \\ \boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\mu}^T & \text{Diag}(\boldsymbol{x}) - \boldsymbol{x}\boldsymbol{x}^T \end{pmatrix} \succeq 0. \end{aligned} \quad (3.4)$$

Define a transformation of the variables by letting $\hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\mu}^T$. Then, $\text{trace}(\hat{\boldsymbol{Y}}) = \text{trace}(\boldsymbol{Y}) - \boldsymbol{\mu}^T\boldsymbol{x}$. This transforms the problem to the equivalent nonlinear semidefinite programming formulation:

$$\begin{aligned} w^{cmm} = \max \quad & \boldsymbol{\mu}^T\boldsymbol{x} + \text{trace}(\hat{\boldsymbol{Y}}) \\ \text{s.t.} \quad & \boldsymbol{x} \in \Delta_{n-1} \\ & \begin{pmatrix} \boldsymbol{\Sigma} & \hat{\boldsymbol{Y}}^T \\ \hat{\boldsymbol{Y}} & \boldsymbol{S}(\boldsymbol{x}) \end{pmatrix} \succeq 0, \end{aligned} \quad (3.5)$$

where $\boldsymbol{S}(\boldsymbol{x}) = \text{Diag}(\boldsymbol{x}) - \boldsymbol{x}\boldsymbol{x}^T$. The matrix $\boldsymbol{S}(\boldsymbol{x})$ is positive semidefinite for all $\boldsymbol{x} \in \Delta_{n-1}$ since:

$$\begin{aligned} \boldsymbol{v}^T \boldsymbol{S}(\boldsymbol{x}) \boldsymbol{v} &= \sum_{i \in \mathcal{N}} v_i^2 x_i - \left(\sum_{i \in \mathcal{N}} v_i x_i \right)^2 \\ &\geq 0, \qquad \forall \boldsymbol{v} \in \mathbb{R}^n, \end{aligned}$$

where the last inequality comes from $\mathbb{E}(v^2) \geq \mathbb{E}(v)^2$ where the random variable v is defined to take value v_i with probability x_i for $i \in \mathcal{N}$. The semidefinite program in (3.5) can be reformulated as a two-stage optimization problem of the form:

$$w^{cmm} = \max \left\{ \boldsymbol{\mu}^T \boldsymbol{x} - V(\boldsymbol{x}) \;\middle|\; \boldsymbol{x} \in \Delta_{n-1} \right\}, \tag{3.6}$$

where $\boldsymbol{x} \in \Delta_{n-1}$ is the first stage decision vector and $V(\boldsymbol{x})$ is the optimal value to the following second stage problem where $\hat{\boldsymbol{Y}}$ is the second stage matrix decision variable:

$$\begin{aligned} V(\boldsymbol{x}) = \min \quad & -\text{trace}(\hat{\boldsymbol{Y}}) \\ \text{s.t.} \quad & \begin{pmatrix} \boldsymbol{\Sigma} & \hat{\boldsymbol{Y}}^T \\ \hat{\boldsymbol{Y}} & \boldsymbol{S}(\boldsymbol{x}) \end{pmatrix} \succeq 0. \end{aligned} \tag{3.7}$$

The second stage semidefinite program in (3.7) for a given value of $\boldsymbol{x}$ has a closed-form solution (see Dowson and Landau 1982, Olkin and Pukelsheim 1982 and Shapiro 1985). Applying this result since $range(\boldsymbol{S}(\boldsymbol{x})) \subseteq range(\boldsymbol{\Sigma})$, the optimal second stage solution is given as:

$$\hat{\boldsymbol{Y}}^{*T} = \boldsymbol{\Sigma} \boldsymbol{S}(\boldsymbol{x})^{1/2} \left(\left(\boldsymbol{S}(\boldsymbol{x})^{1/2} \boldsymbol{\Sigma} \boldsymbol{S}(\boldsymbol{x})^{1/2} \right)^{1/2} \right)^{\dagger} \boldsymbol{S}(\boldsymbol{x})^{1/2}, \tag{3.8}$$

where $(\cdot)^\dagger$ denotes the Moore-Penrose pseudoinverse of a matrix. Hence, the optimal value of formulation (3.7) is:

$$\begin{aligned} V(\boldsymbol{x}) = & -\text{trace}\left(\boldsymbol{\Sigma}\boldsymbol{S}(\boldsymbol{x})^{1/2}\left(\left(\boldsymbol{S}(\boldsymbol{x})^{1/2}\boldsymbol{\Sigma}\boldsymbol{S}(\boldsymbol{x})^{1/2}\right)^{1/2}\right)^{\dagger}\boldsymbol{S}(\boldsymbol{x})^{1/2}\right) \\ = & -\text{trace}\left(\left(\boldsymbol{S}(\boldsymbol{x})^{1/2}\boldsymbol{\Sigma}\boldsymbol{S}(\boldsymbol{x})^{1/2}\right)\left(\left(\boldsymbol{S}(\boldsymbol{x})^{1/2}\boldsymbol{\Sigma}\boldsymbol{S}(\boldsymbol{x})^{1/2}\right)^{1/2}\right)^{\dagger}\right) \\ = & -\text{trace}\left(\left(\boldsymbol{S}(\boldsymbol{x})^{1/2}\boldsymbol{\Sigma}\boldsymbol{S}(\boldsymbol{x})^{1/2}\right)^{1/2}\right) \\ = & -\text{trace}\left(\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{S}(\boldsymbol{x})\boldsymbol{\Sigma}^{1/2}\right)^{1/2}\right), \end{aligned} \tag{3.9}$$

where the second equality comes from the invariance of the trace under cyclic permutations, the third equality comes from the property of the pseudo-inverse that $\boldsymbol{A}(\boldsymbol{A}^{1/2})^\dagger = \boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2}(\boldsymbol{A}^{1/2})^\dagger = \boldsymbol{A}^{1/2}$ and the last equality comes from the observation that for any $n\times n$ real square matrix $\boldsymbol{A}$, the matrices $\boldsymbol{A}\boldsymbol{A}^T$ and $\boldsymbol{A}^T\boldsymbol{A}$ have the same set of eigenvalues (see Horn and Johnson 1985). By substituting into (3.6), we obtain:

$$w^{cmm} = \max\left\{\boldsymbol{\mu}^T\boldsymbol{x} + \text{trace}\left(\left(\boldsymbol{\Sigma}^{1/2}\ \boldsymbol{S}(\boldsymbol{x})\ \boldsymbol{\Sigma}^{1/2}\right)^{1/2}\right)\ \Big|\ \boldsymbol{x}\in\Delta_{n-1}\right\}.$$

□

Remark 3.1.2 *The second stage problem in (3.7) has been studied in Dowson and Landau (1982); Olkin and Pukelsheim (1982); Shapiro (1985) in the following context: Given two n-dimensional random vectors with covariance matrices* $\boldsymbol{\Sigma}$ *and* $\boldsymbol{S}(\boldsymbol{x})$*, find the cross moment matrix* $\hat{\boldsymbol{Y}}^T$ *between the two random vectors that minimizes the expected* L_2 *distance between the vectors. In the proof of Theorem 3.1.1, the two vectors in the second stage correspond to the random component of the utility vector* $\boldsymbol{\epsilon}$ *and the random choice vector which chooses* $\boldsymbol{e}_i$ *(alternative i) with probability* x_i*. The first stage problem corresponds to finding the best probability vector* $\boldsymbol{x}$*. For completeness, we provide a proof for the optimality of the solution in (3.8) in Section 3.5.*

3.1.2 Strong Concavity of the Objective Function

In this section, we prove strong concavity of the objective function in the representative agent formulation for the CMM model. The result is based on the following definitions of functions of (positive semidefinite) matrices. Consider a symmetric positive semidefinite matrix $\boldsymbol{A}$ with an eigendecomposition $\boldsymbol{Q}\text{Diag}(\boldsymbol{\lambda})\boldsymbol{Q}^T$ where $\boldsymbol{Q}$ is an orthonormal matrix and $\boldsymbol{\lambda}$ is the vector of nonnegative eigenvalues. Given a function $h(\cdot) : [0, \infty) \to [0, \infty)$, the matrix function is defined as $h(\boldsymbol{A}) = \boldsymbol{Q}\text{Diag}(h(\boldsymbol{\lambda}))\boldsymbol{Q}^T$ where the function $h(\cdot)$ is applied to the eigenvalues in the diagonal matrix. As is the convention, we use $\boldsymbol{A} \succeq \boldsymbol{B}$ to denote $\boldsymbol{A} - \boldsymbol{B} \succeq 0$.

Definition 2 *Consider a function $h : [0, \infty) \to [0, \infty)$.*

(a) The function $h(\cdot)$ is operator monotone if for all $\boldsymbol{A}, \boldsymbol{B} \succeq 0$:

$$\boldsymbol{A} \succeq \boldsymbol{B} \Longrightarrow h(\boldsymbol{A}) \succeq h(\boldsymbol{B}).$$

(b) The function $h(\cdot)$ is operator concave if for all $\boldsymbol{A}, \boldsymbol{B} \succeq 0$ and $\lambda \in [0, 1]$:

$$h\left((1-\lambda)\boldsymbol{A} + \lambda\boldsymbol{B}\right) \succeq (1-\lambda)h(\boldsymbol{A}) + \lambda h(\boldsymbol{B}).$$

An example of a matrix function that is both operator monotone and operator concave is the square root function.

Theorem 3.1.3 *The function $h(t) = t^{1/2}$ is both operator monotone and operator concave.*

This theorem is a special case of the Löwner-Heinz Theorem (see Löwner 1934 and Heinz 1951). Before we introduce a key result of this section, we recall the definition of *strong convexity*.

Definition 3 *A function $V(\boldsymbol{x}) : \mathcal{D} \to \mathbb{R}$ where $\mathcal{D}$ is a convex subset of $\mathbb{R}^n$ is strongly convex if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}$ and $\lambda \in (0, 1)$, there exists a constant $m > 0$ such that*

$$V(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda V(\boldsymbol{x}) + (1-\lambda)V(\boldsymbol{y}) - \frac{m}{2}\lambda(1-\lambda)\|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

This brings us to the following theorem for the objective function in formulation (3.3), the proof of which would be given in section 3.5.

Theorem 3.1.4 *The function* $V(\boldsymbol{x}) = -trace\left(\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{S}(\boldsymbol{x})\boldsymbol{\Sigma}^{1/2}\right)^{1/2}\right)$ *defined on the unit simplex* $\boldsymbol{x} \in \Delta_{n-1}$ *is strongly convex for* $\boldsymbol{\Sigma} \succ 0$.

3.1.3 Optimality Conditions and Their Implications

One of the key advantages in developing the representative agent formulation for the CMM model is that it transforms a semidefinite program to a nonlinear strongly concave maximization problem over the unit simplex. In this section, we provide a characterization of the directional derivatives of the objective function and prove optimality conditions for the model. In addition, we show that as we approach the boundary of the feasible region from its interior, the (projected) gradient of the objective function blows to infinity. These results have important implications to the algorithm we shall propose in Section 3.2.

Projected Gradient of the Objective Function

Consider a vector $\boldsymbol{x}$ in the relative interior of the simplex and restrict the direction of the perturbation to be in the tangent space of $\boldsymbol{\Delta}_{n-1}$ defined as $\overline{\Delta}_{n-1} = \left\{\boldsymbol{v} \in \mathbb{R}_n \;\middle|\; \boldsymbol{e}^T\boldsymbol{v} = 0\right\}$. Let $||\boldsymbol{v}||_2 = 1$, then the directional derivative of $V(\boldsymbol{x})$ in the direction $\boldsymbol{v}$ is defined as:

$$\boldsymbol{\nabla}_{\boldsymbol{v}} V(\boldsymbol{x}) = \lim_{\epsilon \to 0} \frac{V(\boldsymbol{x} + \epsilon\boldsymbol{v}) - V(\boldsymbol{x})}{\epsilon}.$$

To compute the directional derivative, observe that:

$$\begin{aligned}
& V(\boldsymbol{x} + \epsilon\boldsymbol{v}) \\
= \; & -\text{trace}\left(\boldsymbol{\Sigma}^{1/2}\ \boldsymbol{S}(\boldsymbol{x} + \epsilon\boldsymbol{v})\ \boldsymbol{\Sigma}^{1/2}\right)^{1/2} \\
= \; & -\text{trace}\Big(\boldsymbol{\Sigma}^{1/2}\boldsymbol{S}(\boldsymbol{x})\boldsymbol{\Sigma}^{1/2} + \epsilon\boldsymbol{\Sigma}^{1/2}\left(\text{Diag}(\boldsymbol{v}) - \boldsymbol{x}\boldsymbol{v}^T - \boldsymbol{v}\boldsymbol{x}^T\right)\boldsymbol{\Sigma}^{1/2} \\
& -\epsilon^2\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{\Sigma}^{1/2}\Big)^{1/2} \\
= \; & -\text{trace}\left(\boldsymbol{T}(\boldsymbol{x}) + \boldsymbol{E}_{\boldsymbol{v}}(\epsilon, \boldsymbol{x})\right)^{1/2},
\end{aligned}$$

where:

$$\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{\Sigma}^{1/2}\boldsymbol{S}(\boldsymbol{x})\boldsymbol{\Sigma}^{1/2}$$

and

$$\boldsymbol{E_v}(\epsilon, \boldsymbol{x}) = \epsilon\boldsymbol{\Sigma}^{1/2}\left(\text{Diag}(\boldsymbol{v}) - \boldsymbol{x}\boldsymbol{v}^T - \boldsymbol{v}\boldsymbol{x}^T\right)\boldsymbol{\Sigma}^{1/2} - \epsilon^2\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{\Sigma}^{1/2}.$$

The next lemma provides a characterization of the null space of the matrix $\boldsymbol{T}(\boldsymbol{x})$ and its relation to the null space of the matrix $\boldsymbol{E_v}(\epsilon, \boldsymbol{x})$. This lemma is needed for the main result of this and the next section.

Lemma 3.1.5

(a) Let $\boldsymbol{x} = \binom{\mathbf{0}}{\underline{\boldsymbol{x}}} \in \Delta_{n-1}$, *where* $\mathbf{0}$ *is a vector of* r *zeros and* $\underline{\boldsymbol{x}} \in \mathbb{R}^{n-r}_{++}$ *is a strictly positive vector for some integer* r *such that* $0 \leq r \leq n-1$. *Then the null space of the matrix* $\boldsymbol{T}(\boldsymbol{x})$ *is given as:*

$$Null(\boldsymbol{T}(\boldsymbol{x})) = \left\{ k\boldsymbol{\Sigma}^{-1/2}\boldsymbol{z} \;\middle|\; \boldsymbol{z} = \begin{pmatrix} \boldsymbol{z_1} \\ \boldsymbol{e} \end{pmatrix},\ \boldsymbol{z_1} \in \mathbb{R}^r \right\},$$

where $\boldsymbol{e}$ *is a vector of ones of dimension* $n-r$ *and* $rank(\boldsymbol{T}(\boldsymbol{x})) = n-r-1$.

(b) Particularly, when $\boldsymbol{x}$ *is in the relative interior of the simplex denoted by* $int(\Delta_{n-1})$, *then* $rank(\boldsymbol{T}(\boldsymbol{x})) = n-1$ *and* $Null(\boldsymbol{T}(\boldsymbol{x})) \subseteq Null(\boldsymbol{E_v}(\epsilon, \boldsymbol{x}))$.

To prove the main theorem of this section, we make use of the Fréchet derivative of a matrix function which is defined as follows.

Definition 4 *The Fréchet derivative of a real matrix function* $g : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^{n\times n}$ *at* $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ *is a linear mapping* $L_g : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^{n\times n}$ *such that* $g(\boldsymbol{X}+\boldsymbol{E}) - g(\boldsymbol{X}) - L_g(\boldsymbol{X}, \boldsymbol{E}) = o(||\boldsymbol{E}||)$ *for all* $\boldsymbol{E} \in \mathbb{R}^{n\times n}$.

The Fréchet derivative, if exists, is known to be unique. The Fréchet derivative for the matrix square root function, which exists when $\boldsymbol{X}$ is positive definite, is the unique solution to the Sylvester equation (refer to Kenney and Laub 1989, Higham 2008):

$$\boldsymbol{X}^{1/2}L_{1/2}(\boldsymbol{X}, \boldsymbol{E}) + L_{1/2}(\boldsymbol{X}, \boldsymbol{E})\boldsymbol{X}^{1/2} = \boldsymbol{E}.$$

Next, we derive the first order derivative of $V(\cdot)$.

Theorem 3.1.6 *Define:*

$$\boldsymbol{g}(\boldsymbol{x}) \;=\; -\tfrac{1}{2}\,diag\left(\boldsymbol{\Sigma}^{1/2}(\boldsymbol{T}^{1/2}(\boldsymbol{x}))^{\dagger}\boldsymbol{\Sigma}^{1/2}\right) + \boldsymbol{\Sigma}^{1/2}(\boldsymbol{T}^{1/2}(\boldsymbol{x}))^{\dagger}\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}, \tag{3.10}$$

where $diag(\cdot)$ *is the column vector formed by the diagonal elements of the matrix and* $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{\Sigma}^{1/2}\boldsymbol{S}(\boldsymbol{x})\boldsymbol{\Sigma}^{1/2}$. *The directional derivative of* $V(\boldsymbol{x})$ *at* $\boldsymbol{x} \in int(\Delta_{n-1})$ *in the direction* $\boldsymbol{v} \in \overline{\Delta}_{n-1}$ *is* $\boldsymbol{\nabla}_{\boldsymbol{v}}V(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{x})^T\boldsymbol{v}$, *and its projected gradient on the tangent space is:*

$$\overline{\boldsymbol{\nabla}}V(\boldsymbol{x}) \;=\; \boldsymbol{g}(\boldsymbol{x}) - \tfrac{1}{n}\boldsymbol{e}^T\boldsymbol{g}(\boldsymbol{x})\boldsymbol{e}. \tag{3.11}$$

Optimality Conditions

The representative agent formulation for the CMM model is:

$$Z^* = \max\left\{ f(\boldsymbol{x}) \;\middle|\; \boldsymbol{x} \in \Delta_{n-1} \right\} \text{ where } f(\boldsymbol{x}) = \boldsymbol{\mu}^T\boldsymbol{x} - V(\boldsymbol{x}).$$

Since the objective function is strongly concave and the first-order derivatives of the objective function has been established in Theorem 3.1.6, we can now write down the first-order optimality conditions for the CMM model as follows:

$$\overline{\boldsymbol{\nabla}}f(\boldsymbol{x}) = \left(\boldsymbol{\mu} - \frac{1}{n}\boldsymbol{e}^T\boldsymbol{\mu}\boldsymbol{e}\right) - \left(\boldsymbol{g}(\boldsymbol{x}) - \frac{1}{n}\boldsymbol{e}^T\boldsymbol{g}(\boldsymbol{x})\boldsymbol{e}\right) = 0 \quad \text{and} \quad \boldsymbol{x} \in \Delta_{n-1}, \tag{3.12}$$

where $\overline{\boldsymbol{\nabla}}f(\boldsymbol{x})$ is the projected gradient of $f(\cdot)$ onto the tangent space of the feasible region. Next, we will discuss some of the implications of these optimality conditions.

Mapping between Mean Utilities and Choice Probabilities

In this subsubsection we show a one-to-one correspondence between the mean utility vector $\boldsymbol{\mu}$ under certain appropriate normalization and the choice probability vector $\boldsymbol{x}$ in the relative interior of the simplex in the CMM model. This is important from the modeling viewpoint since it shows that the CMM model is capable of generating all the choice probability vectors in the relative interior of the unit simplex. Furthermore, this is important in identification and estimation of demand parameters (see Berry 1994). We show that under mild assumptions on the covariance matrix, inverting the

choice probability vector in the CMM model is fairly easy. Towards this, we first prove the following lemma that characterizes the gradient of the objective function in the representative agent formulation of the CMM model near the relative boundary of the simplex.

Theorem 3.1.7 *Assume that* $\mathbf{\Sigma} \succ 0$. *As* $\boldsymbol{x}$ *approaches the relative boundary of the unit simplex, the projected gradient of* $V(\cdot)$ *blows up to* $+\infty$.

We are now ready to prove the main result of this section. Hofbauer and Sandholm (2002) have shown that given any joint distribution of the noise terms, the mapping from the deterministic components of the utilities $\boldsymbol{\mu}$ (under appropriate normalization) to the set of choice probabilities in the relative interior of the simplex is surjective, namely any vector of choice probabilities can be obtained by selecting suitable mean values. We show in the next theorem that under mild assumptions on the covariance matrix, there is a one-to-one correspondence between the mean utility vectors under the normalization condition $\mu_1 = 0$ and the choice probability vectors in the relative interior of the simplex for the CMM model.

Theorem 3.1.8 *Assume that* $\mathbf{\Sigma} \succ 0$. *Without loss of generality, set* $\mu_1 = 0$. *Let* $\boldsymbol{q} = P(\boldsymbol{\mu}) : \{0\} \times \mathbb{R}^{n-1} \to \Delta_{n-1}$ *be the mapping from the mean utilities to the choice probabilities in the CMM model. Then* $P(\cdot)$ *is a bijection between* $\{0\} \times \mathbb{R}^{n-1}$ *and the relative interior of the simplex* Δ_{n-1}, *namely there is a one-to-one correspondence between the mean utility vectors and the choice probability vectors.*

Proof:

(a) We first show that every mean vector in $\{0\} \times \mathbb{R}^{n-1}$ in the CMM model results in a unique vector of choice probabilites in the relative interior of the unit simplex. From the strong concavity of the objective function in the representative agent formulation of the CMM model (Theorems 3.1.1 and 3.1.4) and the observation that the gradient of the objective function blows up to infinity near the relative boundary of the simplex (Theorem 3.1.7), the choice probability vector in the CMM model lies strictly in the relative interior of the simplex and is unique.

(b) We next show that every choice probability vector in the relative interior of the simplex maps to a unique mean vector in $\{0\} \times \mathbb{R}^{n-1}$ in the CMM model. From the

optimality conditions in (3.12) and with $\mu_1 = 0$, by multiplying with the vector $\boldsymbol{e}_1$ we have:

$$\frac{1}{n}\boldsymbol{e}^T\boldsymbol{\mu} = \frac{1}{n}\boldsymbol{e}^T\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}_1(\boldsymbol{x}).$$

Plugging in back to the optimality conditions, we obtain the mean utilities from the choice probabilities as follows:

$$\boldsymbol{\mu} = \boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}_1(\boldsymbol{x})\boldsymbol{e}. \tag{3.13}$$

Put together, these results imply that there is a one-to-one correspondence between the set of deterministic utilities in $\{0\} \times \mathbb{R}^{n-1}$ and the set of choice probabilities in the relative interior of the unit simplex. □

For the MNL model, the mean utility vector is uniquely identified from the following simple formula:

$$\mu_i = \ln(q_i^{\text{mnl}}) - \ln(q_1^{\text{mnl}}), \quad \forall i \in \mathcal{N}.$$

A similar result exists for identifying the mean utility vector from the nested logit model (see Berry 1994). For the CMM model, the mean utility vector is uniquely identified from the simple calculation in (3.13) where

$$\boldsymbol{g}(\boldsymbol{x}) = -\frac{1}{2}\left(\text{diag}\left(\boldsymbol{\Sigma}^{1/2}(\boldsymbol{T}^{1/2}(\boldsymbol{x}))^\dagger\boldsymbol{\Sigma}^{1/2}\right) - 2\boldsymbol{\Sigma}^{1/2}(\boldsymbol{T}^{1/2}(\boldsymbol{x}))^\dagger\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\right).$$

To the best of our knowledge, no such easily computable formula is available for the MNP model.

3.2 Calculating the Choice Probabilities in the CMM model

3.2.1 The Gradient Ascent Algorithm

In this section we present a projected gradient ascent method [1] to calculate the choice probabilities in the CMM model. The algorithm is given in Algorithm 1 in which stepsizes are chosen according to the well-known Armijo's line search rule.

Algorithm 1: Projected gradient ascent algorithm with Armijo search

Input: $\boldsymbol{\mu}, \boldsymbol{\Sigma}$, starting point $\boldsymbol{x_0}$, initial step size $\alpha_0 \in (0,1]$, $\beta \in (0,1)$, $\tau \in (0,1)$, tolerance $\epsilon > 0$.

Output: Optimal solution $\boldsymbol{x}^*$.

Initialize stopping criteria: $criteria \leftarrow \epsilon + 1$;
while $criteria > \epsilon$ **do**
  $\alpha \leftarrow \alpha_0$
  $\boldsymbol{x} \leftarrow \boldsymbol{x}_0 + \alpha \overline{\boldsymbol{\nabla}} f(\boldsymbol{x}_0)$,
  while $\boldsymbol{x} \notin int(\Delta_{n-1})$ *or* $f(\boldsymbol{x}) < f(\boldsymbol{x}_0) + \tau\alpha \|\overline{\boldsymbol{\nabla}} f(\boldsymbol{x}_0)\|^2$ **do**
    $\alpha \leftarrow \beta\alpha$
    $\boldsymbol{x} \leftarrow \boldsymbol{x}_0 + \alpha \overline{\boldsymbol{\nabla}} f(\boldsymbol{x}_0)$,
  $\boldsymbol{x}_0 \leftarrow \boldsymbol{x}$
  $criteria \leftarrow \|\boldsymbol{x} - \boldsymbol{x_0}\|$.

From Theorem 3.1.7, we know that the choice probability vector lies in $\text{int}(\Delta_{n-1})$. The algorithm presented in Algorithm 1 converges to the optimal solution (see Iusem 2003). While the objective function has a nice curvature (it is strongly concave), it does not have a Lipschitz continuous gradient near the relative boundary. In fact, the function itself does not satisfy the Lipschitz continuity condition near the relative boundary (see Theorem 3.1.7). In the next section, we show that if $\boldsymbol{x}$ is sufficiently far away from the relative boundary of the feasible region, then the algorithm converges linearly for appropriately chosen parameters within a local neighborhood. As we show numerically in Section 5, this helps explain the good behavior of the algorithm in most cases while

[1] Note that the presentation here is slightly different than the one in classical references such as Section 2.3 of Bertsekas (1999), but the algorithm is the same since the projection is onto an affine subspace, i.e., $\text{Proj}_{\boldsymbol{Ax}=\boldsymbol{b}}(\hat{\boldsymbol{x}} + \nabla f) = \hat{\boldsymbol{x}} + \text{Proj}_{\boldsymbol{Ax}=\boldsymbol{0}}(\nabla f)$ for $\boldsymbol{A}\hat{\boldsymbol{x}} = \boldsymbol{b}$.

for some ill-conditioned problems where for example one of the choice probabilities is very low, the algorithm tends to be slower.

3.2.2 Local Linear Convergence of the Algorithm

We first show that with $\tau \in [0.5, 1)$, the distance between the solution at successive iterations and the optimal solution is non-increasing.

Lemma 3.2.1 *Let $d_k = ||\boldsymbol{x}^k - \boldsymbol{x}^*||$, where $\boldsymbol{x}^*$ is the optimal solution and $\boldsymbol{x}^k$ is the k-th iterate. Then $d_k \leq d_{k-1}$ for all $k > 0$ if $\tau \geq 0.5$.*

We are now ready to discuss the rate of convergence of the algorithm by choosing the parameter τ carefully.

Theorem 3.2.2 *If there exists a $\gamma \in (0, 1)$ such that $||\boldsymbol{x}^0 - \boldsymbol{x}^*|| \leq \gamma x^*_{\min}$ where $x^*_{\min} = \min_i x^*_i$ and if $\tau = 0.5$, then*

$$f(\boldsymbol{x}^*) - f(\boldsymbol{x}^k) \leq \theta^k \left(f(\boldsymbol{x}^*) - f(\boldsymbol{x}^0) \right),$$

where $\theta = 1 - \min\{m, \beta m/L\}$ with m defined as the strong convexity constant in the proof of Theorem 3.1.4 and

$$L = \frac{9n}{4(1-\gamma)x^*_{\min}\sigma_1} \left\| \boldsymbol{\Sigma}^{1/2} \right\|^4 + \frac{n}{((1-\gamma)x^*_{\min}\sigma_1)^{1/2}} \left\| \boldsymbol{\Sigma}^{1/2} \right\|^2,$$

where $\sigma_1 = \lambda_1(\boldsymbol{\Sigma})$.

Proof: From Lemma 3.2.1, we know that $||\boldsymbol{x}^k - \boldsymbol{x}^*|| \leq \gamma x^*_{\min}$ for all $k > 0$. So we can restrict the feasible region to $\mathcal{X} = \Delta_{n-1} \cap \{\boldsymbol{x} : ||\boldsymbol{x} - \boldsymbol{x}^*|| \leq \gamma x^*_{\min}\}$. It is easily seen that for all $\boldsymbol{x} \in \mathcal{X}$, $\min_i x_i \geq (1-\gamma)x^*_{\min}$ and, therefore, $\mathcal{X} \subset \text{int}(\Delta_{n-1})$. Applying Theorem 3.5.4 given in Section 3.5, $\overline{\boldsymbol{\nabla}} f(\boldsymbol{x}^k)$ is Lipschitz continuous over $\mathcal{X}$ with Lipschitz constant

$$L = \frac{9n}{4(1-\gamma)x^*_{\min}\sigma_1} \left\| \boldsymbol{\Sigma}^{1/2} \right\|^4 + \frac{n}{((1-\gamma)x^*_{\min}\sigma_1)^{1/2}} \left\| \boldsymbol{\Sigma}^{1/2} \right\|^2.$$

The result follows from the linear convergence rate result (see Boyd and Vandenberghe 2004) for $\tau \leq 0.5$, with $\theta = 1 - \min\{2m\tau, 2\beta\tau m/L\}$, where m is the strong convexity

constant. In our case, $\tau = 0.5$, so we have

$$f(\boldsymbol{x}^*) - f(\boldsymbol{x}^k) \leq \theta^k \left(f(\boldsymbol{x}^*) - f(\boldsymbol{x}^0) \right),$$

where $\theta = 1 - \min\{m, \beta m/L\}$ with m being the constant in the proof of Theorem 3.1.4. □

Since the algorithm converges globally (see Iusem 2003), Theorem 3.2.2 shows that there exists a large enough integer M, such that, after M iterations the algorithm converges linearly. The algorithm is thus locally linearly convergent. The typical behavior of the algorithm is presented in Figure 3.1. We provide a more detailed computational study regarding the convergence of the algorithm in the next section.

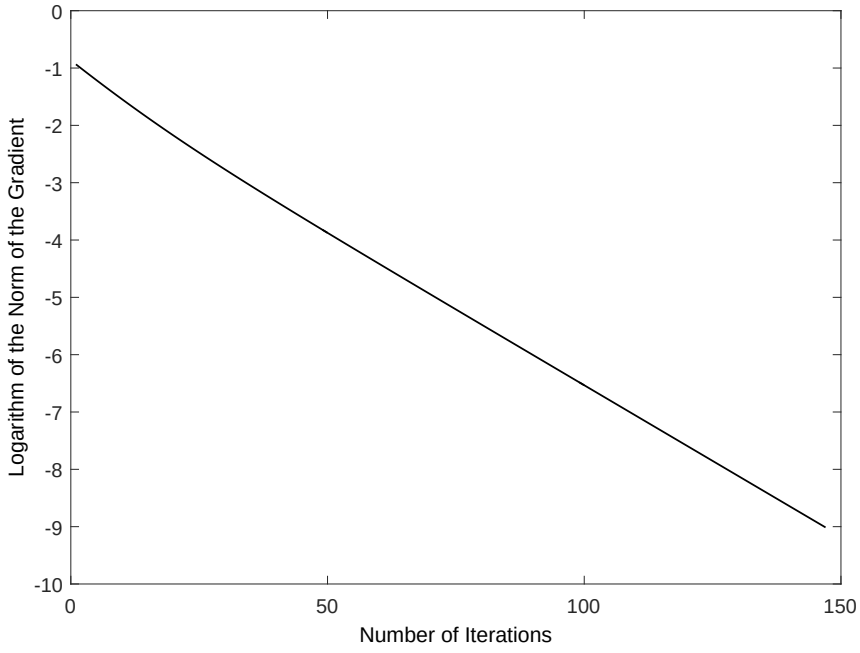

Figure 3.1: Local convergence of the algorithm for a random instance with $n = 100$.

3.3 Computational Results

In the first set of numerical experiments, we compare the computational times and accuracy of the gradient ascent method developed in this chapter for the CMM model with an SDP solver that is suitable for solving large scale SDPs. In the second set of experiments, we numerically test the convergence of the gradient method for the CMM model. In the third set of experiments, we compare the choice probabilities from the CMM and MNP model. The results help illustrate the efficacy of the model.

3.3.1 Comparison of SDP and the Gradient Method for the CMM Model

The SDP in (3.2) was solved using the SDPNAL+ version 0.3 (beta) while the code for the gradient ascent method was developed in MATLAB R2014[2]. The computational experiments were run on a laptop with an Intel(R) i7-5600U CPU processor (2.6 GHz) with 4GB RAM.

The number of alternatives n was varied in the set $\{100, 200, \ldots, 1000\}$. The mean of the utilities was randomly generated in $[0, 1]^n$. The covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{V}\text{Diag}(\boldsymbol{d})\boldsymbol{V}^T$ was randomly generated by choosing the eigenvalues in the vector $\boldsymbol{d}$ uniformly in $(0, 1]^n$ and the eigenvectors in $\boldsymbol{V}$ using an orthogonalization of a random n by n matrix with each entry in $[-1, 1]$. For each size n, 9 instances were randomly generated. In the computational experiments, we used the default settings for SDPNAL+ version 0.3. For the gradient method, the parameters were set as $\alpha_0 = 0.1$, $\tau = 0.5$, $\beta = 0.6$ with $\epsilon = 10^{-4}$. To compare the accuracy of the methods, we evaluated the error measured in L_2-distance between the choice probability vector obtained from the SDP solver and the gradient ascent method:

$$\text{error}_{\text{prob}} = \|\boldsymbol{x}^*_{sdp} - \boldsymbol{x}^*_{grd}\|_2,$$

where $\boldsymbol{x}^*_{sdp}$ and $\boldsymbol{x}^*_{grd}$ are the solutions obtained from the SDP solver and the gradient ascent method, respectively. We also evaluated the difference in the optimal objective

[2]The code for the gradient method and the test instances can be obtained from the webpages of the authors.

value as follows:

$$\text{error}_{\text{obj}} = |f(\boldsymbol{x}^*_{sdp}) - f(\boldsymbol{x}^*_{grd})|.$$

The results are provided in Table 3.1 which clearly indicates that both methods are very close in terms of choice probabilities and the objective value. In Table 3.2, the computational times for the two methods are provided which illustrate that the gradient method converges much faster for this set of instances in comparison to the SDP solver.

n	error	1	2	3	4	5	6	7	8	9
100	Prob	1.63e-5	2.46e-5	3.69e-5	0.743e-5	3.08e-5	0.84e-5	2.92e-5	2.04e-5	2.47e-5
	Obj	0.0056	0.0071	0.0055	0.0020	0.0054	0.0076	0.0067	0.0056	0.0057
200	Prob	2.67e-5	2.53e-5	8.76e-5	2.36e-4	2.109e-4	2.672e-4	1.867e-4	2.06e-5	3.32e-5
	Obj	0.0071	0.0077	0.0077	0.0060	0.0050	0.0067	0.0046	0.0075	0.0085
300	Prob	9.40e-5	1.53e-4	0.88e-5	1.14e-4	2.34e-4	1.19e-4	1.04e-4	1.53e-4	2.08e-4
	Obj	0.0033	0.0021	0.0065	0.0018	0.0016	0.0004	0.0028	0.0004	0.0052
400	Prob	1.73e-4	2.89e-4	7.14e-5	0.95e-5	4.17e-5	3.78e-5	5.23e-5	1.43e-4	3.75e-5
	Obj	0.0031	0.0028	0.0006	0.0055	0.0028	0.0041	0.0026	0.0039	0.0033
500	Prob	8.79e-5	5.30e-5	2.95e-5	1.80e-5	3.54e-4	3.01e-5	4.29e-4	3.28e-5	1.51e-4
	Obj	0.0027	0.0099	0.0026	0.0070	0.0086	0.0065	0.0104	0.0073	0.0024
600	Prob	2.14e-05	2.10e-05	2.21e-06	2.82e-05	2.52e-05	2.70e-03	2.13e-05	2.27e-05	3.70e-05
	Obj	0.0055	0.0042	0.0021	0.0190	0.0044	0.1066	0.0047	0.0046	0.0047
700	Prob	7.16e-05	4.61e-05	1.45e-04	3.99e-05	2.31e-05	3.63e-05	1.82e-04	2.60e-05	5.60e-05
	Obj	0.0015	0.0012	0.0016	0.0010	0.0066	0.0019	0.0051	0.0074	0.0093
800	Prob	4.49e-05	1.78e-04	3.01e-05	2.19e-04	4.32e-04	4.51e-05	2.84e-04	6.64e-04	5.08e-04
	Obj	0.0015	0.0000	0.0010	0.0058	0.0166	0.0017	0.0163	0.0013	0.0113
900	Prob	1.12e-04	4.53e-05	4.08e-05	7.41e-04	4.86e-05	3.72e-04	3.06e-05	2.79e-04	6.29e-05
	Obj	0.0137	0.0108	0.0055	0.0299	0.0120	0.0171	0.0009	0.0038	0.0008
1000	Prob	3.88e-05	3.60e-04	3.62e-05	2.53e-04	7.74e-05	3.45e-05	5.56e-05	1.16e-04	6.13e-05
	Obj	0.0019	0.0088	0.0084	0.0004	0.0045	0.0024	0.0158	0.0036	0.0098

Table 3.1: Comparison of SDPNAL+ and the gradient method in terms of accuracy for 9 instances for each n.

n	Time (s)	1	2	3	4	5	6	7	8	9
100	Grad	0.40	0.34	0.34	0.34	0.37	0.37	0.34	0.37	0.37
	SDP	11.82	6.78	8.23	65.49	6.22	7.84	6.35	8.22	5.85
200	Grad	1.43	1.17	1.23	1.38	1.27	1.35	1.29	1.26	1.21
	SDP	31.31	30.38	29.53	33.97	34.25	31.63	32.85	29.79	30.84
300	Grad	2.51	2.68	2.27	2.60	2.24	2.46	2.71	2.58	2.34
	SDP	114.45	99.74	101.27	114.67	116.59	121.58	113.42	111.21	120.35
400	Grad	4.35	3.60	3.47	3.86	3.83	3.49	3.65	5.24	4.21
	SDP	274.00	314.95	282.51	256.59	271.39	266.01	284.29	297.44	176.16
500	Grad	6.44	8.39	5.75	5.17	4.96	5.05	5.44	4.91	5.30
	SDP	617.63	548.72	527.94	477.75	585.20	467.12	521.08	462.04	499.21
600	Grad	13.61	14.24	12.62	12.94	13.35	13.35	13.49	14.22	19.07
	SDP	715.00	683.00	17903.00	2864.00	755.00	1829.00	655.00	649.00	891.00
700	Grad	23.16	23.49	20.87	19.92	21.38	20.51	19.84	20.73	20.87
	SDP	1220.30	1564.10	1692.80	1423.60	1163.10	1484.60	1264.40	1569.90	1186.40
800	Grad	33.08	33.83	33.86	38.14	39.65	43.21	39.65	29.87	33.80
	SDP	2304.20	1880.30	2330.00	2091.30	2812.10	2325.40	2717.60	2715.70	2600.10
900	Grad	49.18	52.07	53.41	48.36	52.50	48.95	53.83	48.90	47.92
	SDP	3665.10	3595.60	3663.40	3809.30	3728.10	3406.50	3571.90	3890.80	3233.50
1000	Grad	60.60	71.76	71.93	66.79	63.91	61.58	63.19	82.55	79.03
	SDP	4846.90	5220.90	5325.90	5398.60	4999.20	4840.20	4954.00	4815.50	5220.80

Table 3.2: Comparison of SDP solver SDPNAL+ and the gradient method in terms of computational times for 9 instances for each n.

3.3.2 Convergence of the Gradient Method for the CMM Model

In the second set of numerical experiments, we study the convergence behavior of the algorithm. Theorem 3.2.2 shows that the region of linear convergence for the algorithm depends on $x^*_{\min}$, which is the distance between the optimal solution and the relative boundary. To study the effect of $x^*_{\min}$, we plot the number of iterations versus the level of accuracy achieved by the algorithm within those iterations, i.e., the tolerance ϵ, in Figure 3.2. To plot Figure 3.2, we pick $n = 100$ and randomly generate a covariance matrix $\mathbf{\Sigma}$.

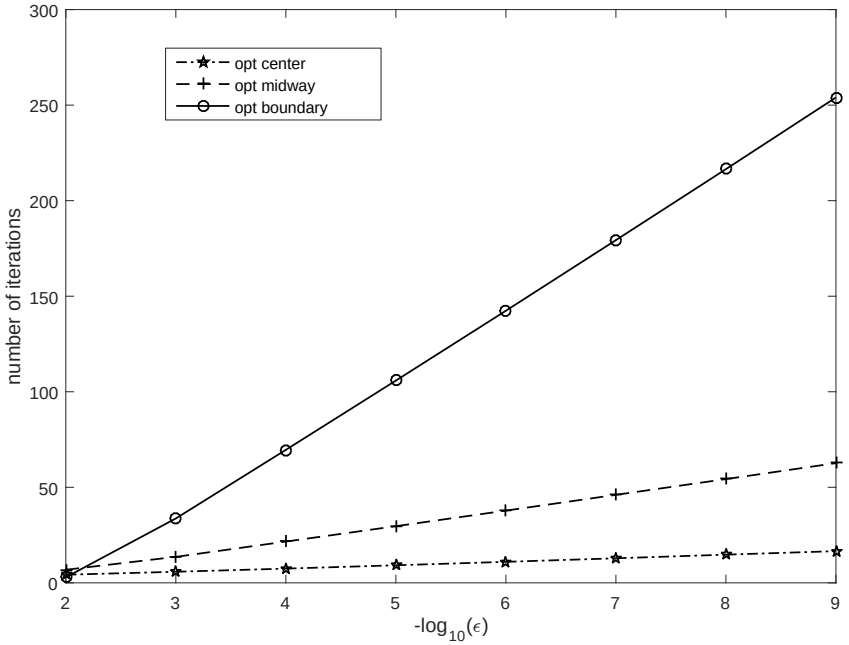

Figure 3.2: Average behavior of the algorithm with different $x^*_{\min}$.

Next, we choose a facet which is the convex combination of 10 randomly picked extreme points, and let the center of the facet be $\boldsymbol{x}^{\mathrm{bd}}$. We consider three scenarios: In the 'opt center' scenario, we let $\boldsymbol{x}^* = 0.9\boldsymbol{x}^{\mathrm{ct}} + 0.1\boldsymbol{x}^{\mathrm{bd}}$, where $\boldsymbol{x}^{\mathrm{ct}} = \{1/100, \ldots, 1/100\}$ is the center of the unit simplex; in the 'opt midway' scenario, we let $\boldsymbol{x}^* = 0.5\boldsymbol{x}^{\mathrm{ct}} + 0.5\boldsymbol{x}^{\mathrm{bd}}$; in the 'opt boundary' scenario, we let $\boldsymbol{x}^* = 0.1\boldsymbol{x}^{\mathrm{ct}} + 0.9\boldsymbol{x}^{\mathrm{bd}}$. Note that we can choose

$$\boldsymbol{\mu} = \boldsymbol{g}(\boldsymbol{x}^*) = -\frac{1}{2}\mathrm{diag}\left(\mathbf{\Sigma}^{1/2}(\boldsymbol{T}^{1/2}(\boldsymbol{x}^*))^\dagger\mathbf{\Sigma}^{1/2}\right) + \mathbf{\Sigma}^{1/2}(\boldsymbol{T}^{1/2}(\boldsymbol{x}^*))^\dagger\mathbf{\Sigma}^{1/2}\boldsymbol{x}^*,$$

to ensure that the optimal solution is $\boldsymbol{x}^*$. Clearly, $x^*_{\min} = 0.009$, 0.005 and 0.001 for these three scenarios. We randomly generate a starting point $\boldsymbol{x}^0$. We then vary ϵ from 10^{-2} to 10^{-9} and record the corresponding number of iterations. Figure 3.2 is obtained by averaging the results of 20 independent replications. From the plot, we can clearly observe the local linear convergence behavior for all three scenarios. As the optimal solution approaches to the boundary, the slope of the plot increases indicating that the constant in the linear convergence rate result increases as $x^*_{\min}$ decreases. This is also predicted by the theoretical results.

To study the influence of the starting point, we plot the average number of iterations and computation times versus the location of the starting point in the Figure 3.3. To

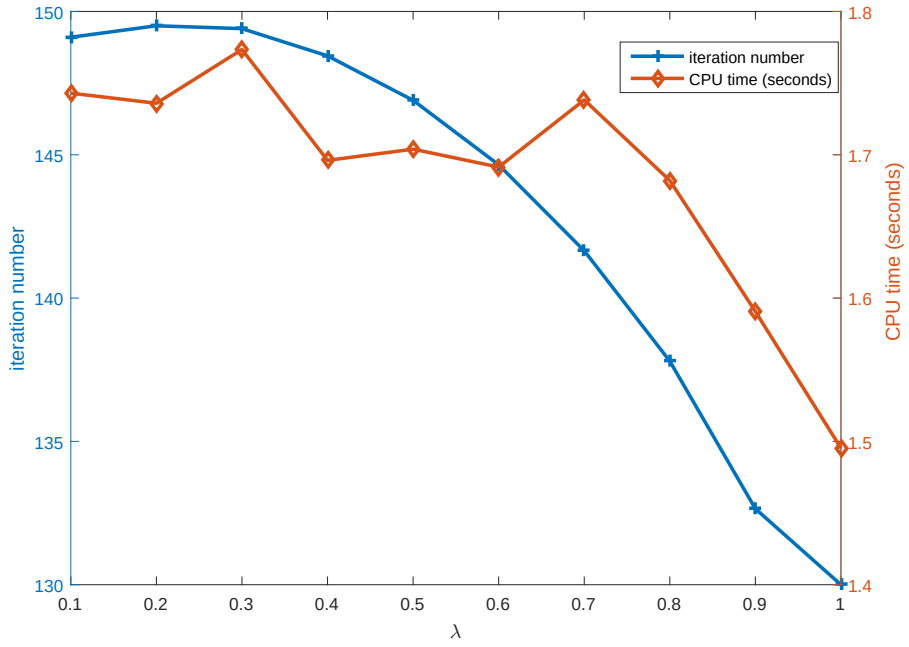

Figure 3.3: Average number of iterations and CPU time versus the location of the initial point.

obtain Figure 3.3, we adopt the settings as in Figure 3.2. Instead of varying the optimal solution, we randomly generate $\boldsymbol{\mu}$ once to use in all experiments but choose various starting points on a line segment between a boundary point and the center of the simplex by setting $\boldsymbol{x}^0 = \lambda \boldsymbol{x}^{\text{ct}} + (1-\lambda)\boldsymbol{x}^{\text{bd}}$ and varying the parameter λ. We fix the

tolerance to $\epsilon = 10^{-6}$. From the figure, we find that required number of iterations to achieve the fixed tolerance level and the corresponding CPU times do not change much with respect to the starting point. The figure indicates that the location of the optimal solution seems to play a more important role for convergence than the location of the initial starting point.

3.3.3 Comparison of the CMM Model and MNP Model

In the last set of numerical experiments, we compare the choice probabilities for the MNP model obtained with the results obtained by simulation and for the CMM model obtained with the gradient ascent method.

Small size examples from Börsch-Supan and Hajivassiliou (1993)

We first provide a comparison of the MNP and CMM choice probabilities for small size examples taken from Börsch-Supan and Hajivassiliou (1993). A popular alternative to the simple frequency simulator for MNP is the GHK simulator (see Geweke 1989, Geweke 1992, Hajivassiliou and McFadden 1998, Keane 1990 and Keane 1994). The GHK simulator makes use of draws from truncated univariate normal distributions and requires evaluation of univariate integrals. Börsch-Supan and Hajivassiliou (1993) have provided four examples with 5 alternatives to show that the GHK simulator produces probability estimates with substantially smaller variances than the simple frequency simulator. The details of the examples are provided in Table 3.3. Example 1 involves mild correlations and has a small choice probability for the first alternative; Example 2 has slightly higher correlations; Example 3 has some large correlation coefficients while Example 4 has a choice probability close to 0.5 with mild correlations. Comparison of the choice probabilities obtained from the GHK simulator for the MNP model and the gradient ascent method for the CMM model are provided in Table 3.3. The results indicate that the choice probability estimate for alternative 1 from the two models are fairly close to each other though developed under different assumptions on the utilities. For Examples 1 and 2, where the choice probability of alternative 1 is small, the CMM model gives a higher choice probability for alternative 1 to be the most preferred one in comparison with the MNP model. On the other hand, for Examples 3 and 4, where the choice probability of alternative 1 is larger, the CMM model gives a slightly lower

chance for alternative 1 to be the most preferred one in comparison with the MNP model.

Example	Parameters	MNP	CMM
1	$\Delta\boldsymbol{\mu} = \begin{pmatrix} -1.00 \\ -0.75 \\ -0.50 \\ -0.20 \end{pmatrix}, \Delta\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.2 & 0.3 & 0.1 \\ 0.2 & 1 & 0.4 & 0.3 \\ 0.3 & 0.4 & 1 & 0.5 \\ 0.1 & 0.3 & 0.5 & 1 \end{pmatrix}$	0.02409 (0.00068)	0.05366
2	$\Delta\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \Delta\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.2 & 0.2 & 0.2 \\ 0.2 & 1 & 0.4 & 0.4 \\ 0.2 & 0.4 & 1 & 0.6 \\ 0.2 & 0.4 & 0.6 & 1 \end{pmatrix}$	0.15037 (0.00444)	0.15668
3	$\Delta\boldsymbol{\mu} = \begin{pmatrix} 1.0 \\ 1.0 \\ 1.0 \\ 1.0 \end{pmatrix}, \Delta\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.9 & 0 & 0 \\ 0.9 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.95 \\ 0 & 0 & 0.95 & 1 \end{pmatrix}$	0.64773 (0.00773)	0.63789
4	$\Delta\boldsymbol{\mu} = \begin{pmatrix} 1.50 \\ 0.75 \\ 0.50 \\ 0.75 \end{pmatrix}, \Delta\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 & 0.2 & 0.1 \\ 0.5 & 1 & 0.5 & 0.2 \\ 0.2 & 0.5 & 1 & 0.5 \\ 0.1 & 0.2 & 0.5 & 1 \end{pmatrix}$	0.49716 (0.01394)	0.47787

Table 3.3: Comparison of the choice probability for alternative 1 between the MNP and CMM model. $\boldsymbol{\Delta\mu}$ and $\boldsymbol{\Delta\Sigma}$ are the mean and the covariance matrix for the utilities $(u_1 - u_2, u_1 - u_3, u_1 - u_4, u_1 - u_5)^T$. The number in parenthesis indicates the standard deviation of the estimator.

Larger example from Jester rating dataset

In this example, we compare the choice probabilities from the CMM and the MNP model where data is available regarding the utilities of a large number of alternatives. We use the rating dataset from the Jester Online Joke Recommender System, in particular Dataset 2+[3]. The data set consists of more than 2 million continuous ratings for 150 jokes collected from over 50000 individuals. Each individual provides ratings between -10 and 10 for a subset of the jokes, 10 of the jokes have never been rated and therefore excluded from the dataset.

To generate the utility parameter, we capture the data in a matrix of size 50000 by 140, whose the $(i, j)^{th}$ entry corresponds to the rating of individual i for joke j, if it exists. For the ratings that are incomplete, we use the standard Collaborative Filtering (CF) method, which is widely used in recommendation engines. The user-based version of CF estimates a missing rating from individual i for joke j based on existing ratings for joke j from a set of individuals who are similar to individual i. Alternatively, the item-based version of CF uses the existing ratings of individual i for other items. We use

[3]http://eigentaste.berkeley.edu/dataset/

the item-based CF in our application, since it is more suitable in situations where the number of items is significantly smaller than the number of individuals (see Ekstrand et al. 2010 for a recent extensive survey on the topic).

To begin with, let us provide the details of the item-based CF. Let r_j denote the j^{th} column of the data matrix and $r(i,j)$ the existing rating from individual i for joke j. We calculate the *estimated* rating $\hat{r}(i,j)$ for individual i and joke j as follows:

$$\hat{r}(i,j) = \frac{\sum_{k \in J_i} w(j,k) r(i,k)}{\sum_{k \in J_i} |w(j,k)|},$$

where J_i is the set of jokes that have been rated by individual i and $w(j,k)$ is a measure of similarity between jokes j and k. Although there are other similarity measures in the literature, we use the cosine similarity, defined as $w(j,k) = \frac{r_j . r_k}{\|r_j\|_2 \|r_k\|_2}$, for its simplicity, popularity, and good predictive properties. Similarly, the ratings can be estimated with alternative methods as well, nevertheless, the weighted average approach is a popular choice.

We use the *completed* data matrix to estimate the mean ratings $\boldsymbol{\mu} \in \mathbb{R}^{140}$ over all users and the corresponding covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{140 \times 140}$. Using these two parameter values, we calculate the choice probabilities, i.e., the probability that a joke j is the most prefered among all jokes,

1. Using the CMM and the gradient ascent algorithm developed in this chapter with tolerance level $\epsilon = 10^{-3}$ and the rest of the parameters as in the previous section.

2. Using the MNP model and the GHK simulator described above with 50000 samples.

We also calculate a basic in-sample statistic corresponding to the number of times a joke has the highest rating divided by the number of individuals. (Whenever there is a tie between l jokes for the highest rating, the count is incremented by $1/l$ instead of 1.) The choice probabilities from the CMM and MNP models together with the in-sample probabilities are provided in Figure 3.4. From the figure, we observe that the alternatives with very small choice probabilities in MNP take on higher choice probability values in the CMM model. On the other hand, the alternatives with larger choice probabilities in MNP take on smaller choice probability values in the CMM model. These results mirror

the observations from the previous section. A possible explanation for this observation is that the distribution of the random utilities that maximizes expected agent utility in the CMM model is a mixture of multivariate normal distributions. The mixture of normals is a fat-tailed distribution and tends to give higher probabilities to the events that are low probability events in the standard normal distribution. In terms of the trend, however the results clearly indicate that the alternatives that are more preferred in one model are also more preferred in the other model.

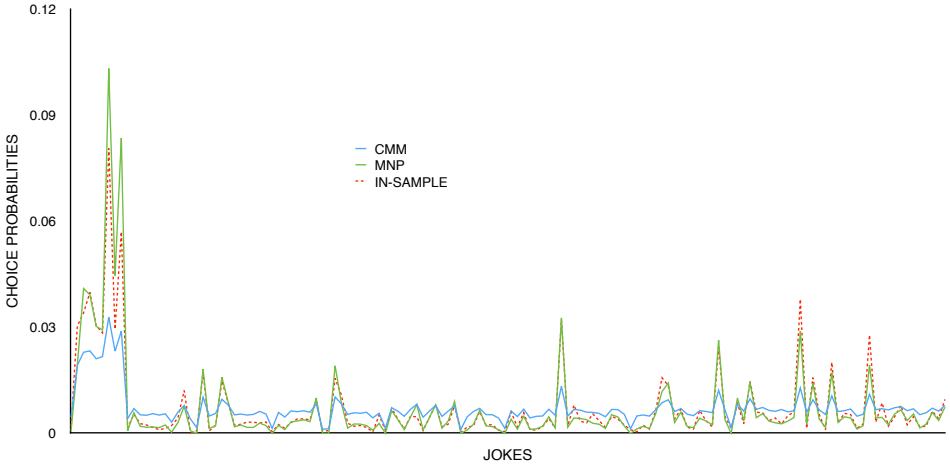

Figure 3.4: Choice Probabilities from CMM, MNP, and In-Sample for the Jester Dataset

3.4 Conclusion

In this chapter, we described a convex optimization approach to compute choice probabilities with correlated utilities. The choice model was derived for the joint distribution of the random utilities that maximizes expected agent utility given only the mean, variance and covariance information. Unlike MNP, the assumption of normality was dropped in this model. In contrast to MNP models where the choice probabilities are evaluated through simulation, we used a simple gradient ascent method to find the

choice probabilities. The biggest advantage of the convex optimization approach is that one can compute choice probabilities for many alternatives with correlated utilities in a reasonable amount of time. In our time, consumers have increasingly more alternative to choose from and the have an increasing amount of information available. Therefore, the method proposed in this chapter provide a viable approach to computing choice probabilities that scales well with size. The next research question is to develop efficient inference techniques for the CMM model.

3.5 Technical Proofs

Proof [Equivalence of formulations (3.1) and (3.2)]

Step 1: To show that optimal value of (3.1) $\leq$ optimal value of (3.2).
Consider an optimal solution to the semidefinite program in (3.1) denoted by $\{\boldsymbol{W}_i^*, \boldsymbol{v}_i^*, x_i^*\}$ for $i \in \mathcal{N}$. We consider the case with all the x_i^* values being strictly positive. Let $x_i = x_i^*$ and $\boldsymbol{Y}^T \boldsymbol{e}_i = \boldsymbol{v}_i^*$ for all i. We start by verifying that the following matrix in (3.2) is positive semdefinite:

$$\begin{pmatrix} \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T & \boldsymbol{Y}^T & \boldsymbol{\mu} \\ \boldsymbol{Y} & \text{Diag}(\boldsymbol{x}) & \boldsymbol{x} \\ \boldsymbol{\mu}^T & \boldsymbol{x}^T & 1 \end{pmatrix} \succeq 0,$$

To see this, observe that:

$$\begin{aligned} & \begin{pmatrix} \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T & \boldsymbol{\mu} \\ \boldsymbol{\mu}^T & 1 \end{pmatrix} - \begin{pmatrix} \boldsymbol{Y}^T \\ \boldsymbol{x}^T \end{pmatrix} \text{Diag}(\boldsymbol{x})^{-1} \begin{pmatrix} \boldsymbol{Y} & \boldsymbol{x} \end{pmatrix} \\ = & \begin{pmatrix} \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T - \boldsymbol{Y}^T \text{Diag}(\boldsymbol{x})^{-1} \boldsymbol{Y} & \boldsymbol{\mu} - \boldsymbol{Y}^T \boldsymbol{e} \\ \boldsymbol{\mu}^T - \boldsymbol{e}^T \boldsymbol{Y} & 1 - \boldsymbol{e}^T \boldsymbol{x} \end{pmatrix} \\ = & \begin{pmatrix} \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T - \sum_i \frac{\boldsymbol{v}_i^* \boldsymbol{v}_i^{*T}}{x_i^*} & \boldsymbol{\mu} - \sum_i \boldsymbol{v}_i^* \\ \boldsymbol{\mu}^T - \sum_i \boldsymbol{v}_i^{*T} & 1 - \sum_i x_i^* \end{pmatrix} \\ = & \begin{pmatrix} \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T - \sum_i \frac{\boldsymbol{v}_i^* \boldsymbol{v}_i^{*T}}{x_i} & \boldsymbol{0} \\ \boldsymbol{0}^T & 0 \end{pmatrix} \succeq 0, \end{aligned}$$

where the third equality comes from the feasibility condition in (3.1) and the positive semidefiniteness of the matrix in the last step follows from Schur's lemma since

$\boldsymbol{W}_i^* \succeq \boldsymbol{v}_i^* \boldsymbol{v}_i^{*T}/x_i$ for $i \in \mathcal{N}$. Thus the solution $\{\boldsymbol{Y}, \boldsymbol{x}\}$ is feasible to the semidefinite program (3.2) with the same objective value. The case with some of the variables $x_i^* = 0$ is handled similarly by dropping the rows and columns corresponding to the zero entries.

Step 2: To show that optimal value of (3.1) $\geq$ optimal value of (3.2).
Consider an optimal solution to the semidefinite program in (3.2) denoted by $\{\boldsymbol{Y^*}, \boldsymbol{x^*}\}$. Consider the case where all the x_i^* values are strictly positive. From Schur's lemma, the positive semidefiniteness of the matrix in (3.2) is equivalent to the following two conditions:

$$\begin{aligned} \boldsymbol{\Lambda} &= \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T - \boldsymbol{Y^*}^T \mathrm{Diag}(\boldsymbol{x^*})^{-1} \boldsymbol{Y^*} \succeq 0, \\ \boldsymbol{Y^*}^T \boldsymbol{e} &= \boldsymbol{\mu}. \end{aligned}$$

Define:

$$\begin{pmatrix} \boldsymbol{W_i} & \boldsymbol{y_i} \\ \boldsymbol{y}_i^T & x_i \end{pmatrix} = \begin{pmatrix} \boldsymbol{Y^*}^T \boldsymbol{e}_i \boldsymbol{e}_i^T \boldsymbol{Y^*}/x_i^* + \boldsymbol{\Lambda}/n & \boldsymbol{Y^*}^T \boldsymbol{e}_i \\ \boldsymbol{e}_i^T \boldsymbol{Y^*} & x_i^* \end{pmatrix}, \quad \forall i \in \mathcal{N}.$$

This is a feasible solution to the semidefinite program (3.1) with the same objective value. As before, the case with some of the $x_i^* = 0$ can be handled by dropping the rows and columns corresponding to the zero entries. Coming together, we obtain the desired result. □

Proof [Optimality of (3.8) for formulation (3.7)]

The dual formulation for the semidefinite program (3.7) is given as:

$$\begin{aligned} V_D^*(\boldsymbol{x}) = \min \quad & \boldsymbol{\Sigma} \cdot \boldsymbol{Y_1} + \boldsymbol{S}(\boldsymbol{x}) \cdot \boldsymbol{Y_2} \\ \text{s.t.} \quad & \begin{pmatrix} \boldsymbol{Y_1} & -\boldsymbol{I}/2 \\ -\boldsymbol{I}/2 & \boldsymbol{Y_2} \end{pmatrix} \succeq 0, \end{aligned}$$

where $\boldsymbol{I}$ is an identity matrix of size $n \times n$. The optimality conditions for the primal and dual semidefinite programs are given as:

1. Primal feasibility:
$$\begin{pmatrix} \boldsymbol{\Sigma} & \hat{\boldsymbol{Y}}^{*T} \\ \hat{\boldsymbol{Y}}^{*} & \boldsymbol{S}(\boldsymbol{x}) \end{pmatrix} \succeq 0.$$

2. Dual feasibility:
$$\begin{pmatrix} \boldsymbol{Y}_{\boldsymbol{1}}^{*} & -\boldsymbol{I}/2 \\ -\boldsymbol{I}/2 & \boldsymbol{Y}_{\boldsymbol{2}}^{*} \end{pmatrix} \succeq 0.$$

3. Complementary slackness:
$$\begin{pmatrix} \boldsymbol{\Sigma} & \hat{\boldsymbol{Y}}^{*T} \\ \hat{\boldsymbol{Y}}^{*} & \boldsymbol{S}(\boldsymbol{x}) \end{pmatrix} \begin{pmatrix} \boldsymbol{Y}_{\boldsymbol{1}}^{*} & -\boldsymbol{I}/2 \\ -\boldsymbol{I}/2 & \boldsymbol{Y}_{\boldsymbol{2}}^{*} \end{pmatrix} = 0.$$

Expanding the complementary slackness condition, we get

$$\begin{aligned} &\text{(a)} \quad \boldsymbol{\Sigma}\boldsymbol{Y}_{\boldsymbol{1}}^{*} - \hat{\boldsymbol{Y}}^{*T}/2 = 0 \\ &\text{(b)} \quad -\boldsymbol{\Sigma}/2 + \hat{\boldsymbol{Y}}^{*T}\boldsymbol{Y}_{\boldsymbol{2}}^{*} = 0 \\ &\text{(c)} \quad \hat{\boldsymbol{Y}}^{*}\boldsymbol{Y}_{\boldsymbol{1}}^{*} - \boldsymbol{S}(\boldsymbol{x})/2 = 0 \\ &\text{(d)} \quad -\hat{\boldsymbol{Y}}^{*}/2 + \boldsymbol{S}(\boldsymbol{x})\boldsymbol{Y}_{\boldsymbol{2}}^{*} = 0. \end{aligned}$$

From conditions (a) and (b), we get the equality:

$$\boldsymbol{\Sigma}\boldsymbol{Y}_{\boldsymbol{1}}^{*}\boldsymbol{Y}_{\boldsymbol{2}}^{*} = \boldsymbol{\Sigma}/4.$$

The matrices $\boldsymbol{Y}_{\boldsymbol{1}}^{*}$ and $\boldsymbol{Y}_{\boldsymbol{2}}^{*}$ are hence nonsingular, related through an inverse: $\boldsymbol{Y}_{\boldsymbol{1}}^{*} = \boldsymbol{Y}_{\boldsymbol{2}}^{*-1}/4$. Condition (a) implies that $\boldsymbol{Y}_{\boldsymbol{1}}^{*} = \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{Y}}^{*T}/2$. Using condition (c), we obtain the matrix equality:

$$\hat{\boldsymbol{Y}}^{*}\boldsymbol{Y}_{\boldsymbol{1}}^{*} = \boldsymbol{S}(\boldsymbol{x})/2 = \hat{\boldsymbol{Y}}^{*}\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{Y}}^{*T}/2.$$

The solution to this quadratic matrix equation is given as:

$$\hat{\boldsymbol{Y}}^{*T} = \boldsymbol{\Sigma}\boldsymbol{S}(\boldsymbol{x})^{1/2}\left(\left(\boldsymbol{S}(\boldsymbol{x})^{1/2}\boldsymbol{\Sigma}\boldsymbol{S}(\boldsymbol{x})^{1/2}\right)^{1/2}\right)^{\dagger}\boldsymbol{S}(\boldsymbol{x})^{1/2}.$$

Also,

$$\boldsymbol{Y}_{\boldsymbol{1}}^{*} = \boldsymbol{Y}_{\boldsymbol{2}}^{*-1}/4 = \frac{1}{2}\boldsymbol{S}(\boldsymbol{x})^{1/2}\left(\left(\boldsymbol{S}(\boldsymbol{x})^{1/2}\boldsymbol{\Sigma}\boldsymbol{S}(\boldsymbol{x})^{1/2}\right)^{1/2}\right)^{\dagger}\boldsymbol{S}(\boldsymbol{x})^{1/2}.$$

□

Proof of Theorem 3.1.4

For all $\boldsymbol{x}, \boldsymbol{y} \in \Delta_{n-1}$ with $\boldsymbol{x} \neq \boldsymbol{y}$ and $\lambda \in (0,1)$, we have:

$$\begin{aligned}
& \boldsymbol{S}(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \\
= \;& \text{Diag}\left(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}\right) - \left(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}\right)\left(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}\right)^T \\
= \;& \lambda \text{Diag}\left(\boldsymbol{x}\right) + (1-\lambda)\text{Diag}\left(\boldsymbol{y}\right) - \lambda^2 \boldsymbol{x}\boldsymbol{x}^T - (1-\lambda)^2 \boldsymbol{y}\boldsymbol{y}^T - \lambda(1-\lambda)(\boldsymbol{x}\boldsymbol{y}^T + \boldsymbol{y}\boldsymbol{x}^T) \\
= \;& \lambda\left(\text{Diag}\left(\boldsymbol{x}\right) - \boldsymbol{x}\boldsymbol{x}^T\right) + (1-\lambda)\left(\text{Diag}\left(\boldsymbol{y}\right) - \boldsymbol{y}\boldsymbol{y}^T\right) + \lambda(1-\lambda)(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^T \\
= \;& \lambda \boldsymbol{S}(\boldsymbol{x}) + (1-\lambda)\boldsymbol{S}(\boldsymbol{y}) + \lambda(1-\lambda)(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^T.
\end{aligned}$$

Pre-multiplying and post-multiplying by $\boldsymbol{\Sigma}^{1/2}$ implies that:

$$\begin{aligned}
& \boldsymbol{\Sigma}^{1/2}\boldsymbol{S}(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y})\boldsymbol{\Sigma}^{1/2} \\
= \;& \boldsymbol{\Sigma}^{1/2}\left(\lambda \boldsymbol{S}(\boldsymbol{x}) + (1-\lambda)\boldsymbol{S}(\boldsymbol{y}) + \lambda(1-\lambda)(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^T\right)\boldsymbol{\Sigma}^{1/2}.
\end{aligned} \tag{3.14}$$

Now let $\boldsymbol{A} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{S}(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y})\boldsymbol{\Sigma}^{1/2}$, $\boldsymbol{B} = \lambda \boldsymbol{\Sigma}^{1/2}\boldsymbol{S}(\boldsymbol{x})\boldsymbol{\Sigma}^{1/2} + (1-\lambda)\boldsymbol{\Sigma}^{1/2}\boldsymbol{S}(\boldsymbol{y})\boldsymbol{\Sigma}^{1/2}$, $\rho = \lambda(1-\lambda)$ and $\boldsymbol{w} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{x}-\boldsymbol{y})$. Using this notation, we can rewrite the equation (3.14) as:

$$\boldsymbol{A} = \boldsymbol{B} + \rho \boldsymbol{w}\boldsymbol{w}^T.$$

Let $\lambda_1(\boldsymbol{A}) \leq \lambda_2(\boldsymbol{A}) \cdots \leq \lambda_n(\boldsymbol{A})$ denote the eigenvalues of $\boldsymbol{A}$ (and respectively $\lambda_i(\boldsymbol{B})$ for $\boldsymbol{B}$). For $\rho > 0$ with a rank one perturbation, Bunch et al. (1978) have shown that:

$$\lambda_i(\boldsymbol{A}) \geq \lambda_i(\boldsymbol{B}), \quad \forall i \in \mathcal{N}.$$

Let $a = \rho \boldsymbol{w}^T\boldsymbol{w}$. Then there exists a vector $\boldsymbol{\beta} \geq 0$ with $\sum_i \beta_i = a$ such that

$$\lambda_i(\boldsymbol{A}) = \lambda_i(\boldsymbol{B}) + \beta_i, \quad \forall i \in \mathcal{N}.$$

Hence a lower bound on the sum of the square roots of the eigenvalues of the matrix $\boldsymbol{A}$ is obtained by solving the optimization problem:

$$\begin{aligned}\sum_{i\in\mathcal{N}}\lambda_i(\boldsymbol{A})^{1/2} \geq \quad & \min_{\boldsymbol{\beta}} \quad \sum_{i\in\mathcal{N}}(\lambda_i(\boldsymbol{B})+\beta_i)^{1/2} \\ & \text{s.t.} \quad \sum_{i\in\mathcal{N}}\beta_i = a, \\ & \qquad \beta_i \geq 0, \qquad \forall i \in \mathcal{N}.\end{aligned}$$

The right hand side of the above inequality corresponds to minimizing a concave function over a simplex. Therefore the minimizer must be attained by at least one of the vertices of the simplex. This gives:

$$\begin{aligned}\sum_{i\in\mathcal{N}}\lambda_i(\boldsymbol{A})^{1/2} &\geq \min_{j\in\mathcal{N}}\left\{\sum_{i\neq j}\lambda_i(\boldsymbol{B})^{1/2}+(\lambda_j(\boldsymbol{B})+a)^{1/2}\right\} \\ &\geq \min_{j\in\mathcal{N}}\left\{\sum_{i\in\mathcal{N}}\lambda_i(\boldsymbol{B})^{1/2}+\frac{a}{2\sqrt{\lambda_j(\boldsymbol{B})+a}}\right\} \\ &= \sum_{i\in\mathcal{N}}\lambda_i(\boldsymbol{B})^{1/2}+\frac{a}{2\sqrt{\lambda_n(\boldsymbol{B})+a}},\end{aligned}$$

where the second inequality is from the supergradient inequality for the concave square root function. Clearly there exists a positive number M_1, such that $a = \lambda(1-\lambda)(\boldsymbol{x}-\boldsymbol{y})^T\boldsymbol{\Sigma}(\boldsymbol{x}-\boldsymbol{y}) \leq M_1$ for all $\lambda \in (0,1)$ and $\boldsymbol{x},\boldsymbol{y} \in \Delta$. Similarly, there exists a positive number M_2, such that $\lambda_n(\boldsymbol{B}) = \max_i \lambda_i(\boldsymbol{B}) \leq M_2$, for all $\lambda \in (0,1)$ and $\boldsymbol{x},\boldsymbol{y} \in \Delta$. Letting $\alpha = \frac{1}{\sqrt{M_1+M_2}}$, we have

$$\sum_{i\in\mathcal{N}}\lambda_i(\boldsymbol{A})^{1/2} \geq \sum_{i\in\mathcal{N}}\lambda_i(\boldsymbol{B})^{1/2}+\frac{\alpha}{2}\lambda(1-\lambda)(\boldsymbol{x}-\boldsymbol{y})^T\boldsymbol{\Sigma}(\boldsymbol{x}-\boldsymbol{y}). \tag{3.15}$$

Since $\text{trace}(\boldsymbol{A}^{1/2}) = \sum_{i\in\mathcal{N}} \lambda_i(\boldsymbol{A})^{1/2}$, by (3.15) we obtain

$$\begin{aligned}
& \text{trace}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{S}(\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y})\boldsymbol{\Sigma}^{1/2}\right)^{1/2} \\
\geq\ & \text{trace}\left(\lambda\boldsymbol{\Sigma}^{1/2}\boldsymbol{S}(\boldsymbol{x})\boldsymbol{\Sigma}^{1/2}+(1-\lambda)\boldsymbol{\Sigma}^{1/2}\boldsymbol{S}(\boldsymbol{y})\boldsymbol{\Sigma}^{1/2}\right)^{1/2} \\
& +\frac{\alpha}{2}\lambda(1-\lambda)(\boldsymbol{x}-\boldsymbol{y})^T\boldsymbol{\Sigma}(\boldsymbol{x}-\boldsymbol{y}) \\
\geq\ & \lambda\text{trace}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{S}(\boldsymbol{x})\boldsymbol{\Sigma}^{1/2}\right)^{1/2}+(1-\lambda)\text{trace}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{S}(\boldsymbol{y})\boldsymbol{\Sigma}^{1/2}\right)^{1/2} \\
& +\frac{\alpha}{2}\lambda(1-\lambda)(\boldsymbol{x}-\boldsymbol{y})^T\boldsymbol{\Sigma}(\boldsymbol{x}-\boldsymbol{y}),
\end{aligned}$$

where the last inequality is from the concavity of the square root function. Let $\lambda_{min}(\boldsymbol{\Sigma})$ be the smallest eigenvalue of $\boldsymbol{\Sigma}$. Since $\boldsymbol{\Sigma}$ is positive definite, then $\lambda_{min}(\boldsymbol{\Sigma}) > 0$ and

$$\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{x}-\boldsymbol{y})\|^2 \geq \lambda_{min}(\boldsymbol{\Sigma})\|\boldsymbol{x}-\boldsymbol{y}\|^2.$$

Then by the definition of the function $V(\cdot)$, we obtain

$$V(\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y}) \leq \lambda V(\boldsymbol{x})+(1-\lambda)V(\boldsymbol{y})-\frac{\alpha}{2}\lambda_{min}(\boldsymbol{\Sigma})\lambda(1-\lambda)\|\boldsymbol{x}-\boldsymbol{y}\|^2,$$

and therefore the function $V(\boldsymbol{x})$ is strongly convex on its domain for $\boldsymbol{\Sigma} \succ 0$, where the strong convexity parameter depends on the matrix $\boldsymbol{\Sigma}$. □

Proof of Lemma 3.1.5

(a) Let $\boldsymbol{A}\circ\boldsymbol{B}$ denote the Hadamard product of two matrices of the same dimension. Any vector $\boldsymbol{z} \in \mathbb{R}^n$ can be expressed as $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}\binom{\boldsymbol{z_1}}{\boldsymbol{z_2}}$, for some $\boldsymbol{z_1} \in \mathbb{R}^r$ and $\boldsymbol{z_2} \in \mathbb{R}^{n-r}$. Then:

$$\begin{aligned}
\boldsymbol{T}(\boldsymbol{x})\boldsymbol{z} &= \boldsymbol{\Sigma}^{1/2}\left(\text{Diag}(\boldsymbol{x})\boldsymbol{\Sigma}^{1/2}\boldsymbol{z}-\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{z}\right) \\
&= \boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\circ\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{z}-\boldsymbol{x}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{z}\boldsymbol{e}\right),
\end{aligned}$$

where the last equality comes from the observation that $\text{Diag}(\boldsymbol{x})(\boldsymbol{\Sigma}^{1/2}\boldsymbol{z}) = \boldsymbol{x} \circ (\boldsymbol{\Sigma}^{1/2}\boldsymbol{z})$ and $\boldsymbol{x}(\boldsymbol{x}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{z}) = \boldsymbol{x} \circ (\boldsymbol{x}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{z})\boldsymbol{e}$. This is equivalent to

$$\boldsymbol{T}(\boldsymbol{x})\boldsymbol{z} \;=\; \boldsymbol{\Sigma}^{1/2}\begin{pmatrix}\boldsymbol{0}\\ \underline{\boldsymbol{x}}\end{pmatrix} \circ \left(\begin{pmatrix}\boldsymbol{z_1}\\ \boldsymbol{z_2}\end{pmatrix} - \begin{pmatrix}\boldsymbol{0}\\ \underline{\boldsymbol{x}}\end{pmatrix}^T \begin{pmatrix}\boldsymbol{z_1}\\ \boldsymbol{z_2}\end{pmatrix} \boldsymbol{e} \right).$$

Since $\underline{\boldsymbol{x}} > \boldsymbol{0}$, $\boldsymbol{T}(\boldsymbol{x})\boldsymbol{z} = \boldsymbol{0}$ implies that:

$$\boldsymbol{z_2} - (\underline{\boldsymbol{x}}^T\boldsymbol{z_2})\boldsymbol{e} = \boldsymbol{0}.$$

Solving this equation gives $\boldsymbol{z} \in \text{Null}(\boldsymbol{T}(\boldsymbol{x})) = \{k\boldsymbol{\Sigma}^{-1/2}\boldsymbol{z} \mid \boldsymbol{z} = \binom{\boldsymbol{z_1}}{\boldsymbol{e}}\}$ where $\boldsymbol{z_1} \in \mathbb{R}_r$. Therefore, $\text{rank}(\text{Null}(\boldsymbol{T}(\boldsymbol{x}))) = r + 1$ and the rank-nullity theorem implies that $\text{rank}(\boldsymbol{T}(\boldsymbol{x})) = n - r - 1$.

(b) For $\boldsymbol{x} \in \text{int}(\Delta_{n-1})$, all the entries are nonzero. To show that $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}$ lies in the null space of the matrix $\boldsymbol{E_v}(\epsilon, \boldsymbol{x})$, observe that:

$$\begin{aligned}
& \boldsymbol{E_v}(\epsilon, \boldsymbol{x})\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e} \\
= \;& \epsilon\boldsymbol{\Sigma}^{1/2}\left(\text{Diag}(\boldsymbol{v}) - \boldsymbol{x}\boldsymbol{v}^T - \boldsymbol{v}\boldsymbol{x}^T\right)\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e} - \epsilon^2\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e} \\
= \;& \epsilon\boldsymbol{\Sigma}^{1/2}\left(\text{Diag}(\boldsymbol{v})\boldsymbol{e} - \boldsymbol{x}\boldsymbol{v}^T\boldsymbol{e} - \boldsymbol{v}\boldsymbol{x}^T\boldsymbol{e}\right) - \epsilon^2\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{e} \\
= \;& \boldsymbol{0},
\end{aligned}$$

where the final equality comes from the observation that $\boldsymbol{e}^T\boldsymbol{x} = 1$ and $\boldsymbol{v}^T\boldsymbol{e} = 0$. Hence, $\text{Null}(\boldsymbol{T}(\boldsymbol{x})) \subseteq \boldsymbol{E_v}(\epsilon, \boldsymbol{x})$. □

Proof of Theorem 3.1.6

Lemma 3.1.5 implies that for the given symmetric matrices $\boldsymbol{T}(\boldsymbol{x})$ and $\boldsymbol{E_v}(\epsilon, \boldsymbol{x})$, there exists an orthogonal matrix $\boldsymbol{P}$ with $\boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}$ such that

$$\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{P}\begin{pmatrix} 0 & \boldsymbol{0}^T \\ \boldsymbol{0} & \overline{\boldsymbol{\Lambda}}(\boldsymbol{x}) \end{pmatrix}\boldsymbol{P}^T \text{ and } \boldsymbol{E_v}(\epsilon, \boldsymbol{x}) = \boldsymbol{P}\begin{pmatrix} 0 & \boldsymbol{0}^T \\ \boldsymbol{0} & \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon, \boldsymbol{x}) \end{pmatrix}\boldsymbol{P}^T,$$

where $\overline{\boldsymbol{\Lambda}}(\boldsymbol{x})$ is a diagonal matrix of size $(n-1) \times (n-1)$ containing the non-zero eigenvalues of $\boldsymbol{T}(\boldsymbol{x})$ and $\boldsymbol{P}$ is the matrix of eigenvectors of matrix $\boldsymbol{T}(\boldsymbol{x})$ with the first

eigenvector equal to $\frac{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}}{||\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}||_2}$. The matrix $\overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon, \boldsymbol{x})$ is however not necessarily diagonal. Thus, we obtain:

$$
\begin{aligned}
& V(\boldsymbol{x}+\epsilon\boldsymbol{v}) - V(\boldsymbol{x}) \\
= \; & -\text{trace}\left((\boldsymbol{T}(\boldsymbol{x}) + \boldsymbol{E}_{\boldsymbol{v}}(\epsilon, \boldsymbol{x}))^{1/2}\right) + \text{trace}\left(\boldsymbol{T}(\boldsymbol{x})^{1/2}\right) \\
= \; & -\text{trace}\left(\boldsymbol{P}\begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \overline{\boldsymbol{\Lambda}}(\boldsymbol{x}) + \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon, \boldsymbol{x}) \end{pmatrix}\boldsymbol{P}^T\right)^{1/2} + \text{trace}\left(\boldsymbol{P}\begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \overline{\boldsymbol{\Lambda}}(\boldsymbol{x}) \end{pmatrix}\boldsymbol{P}^T\right)^{1/2} \\
= \; & -\text{trace}\begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \overline{\boldsymbol{\Lambda}}(\boldsymbol{x}) + \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon, \boldsymbol{x}) \end{pmatrix}^{1/2} + \text{trace}\begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \overline{\boldsymbol{\Lambda}}(\boldsymbol{x}) \end{pmatrix}^{1/2} \\
= \; & -\text{trace}\left(\left(\overline{\boldsymbol{\Lambda}}(\boldsymbol{x}) + \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon, \boldsymbol{x})\right)^{1/2} - \overline{\boldsymbol{\Lambda}}^{1/2}(\boldsymbol{x})\right).
\end{aligned}
$$

To evaluate the last expression, let $L_{1/2}(\overline{\boldsymbol{\Lambda}}(\boldsymbol{x}), \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon, \boldsymbol{x}))$ denote the Fréchet derivative for the matrix square root which is the unique solution to the Sylvester equation:

$$
\overline{\boldsymbol{\Lambda}}^{1/2}(\boldsymbol{x})L_{1/2}(\overline{\boldsymbol{\Lambda}}(\boldsymbol{x}), \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon, \boldsymbol{x})) + L_{1/2}(\overline{\boldsymbol{\Lambda}}(\boldsymbol{x}), \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon, \boldsymbol{x}))\overline{\boldsymbol{\Lambda}}^{1/2}(\boldsymbol{x}) = \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon, \boldsymbol{x}).
$$

The existence of solution is guaranteed since $\overline{\boldsymbol{\Lambda}}(\boldsymbol{x}) \succ 0$. The Sylvester equation can then be expressed as:

$$
L_{1/2}(\overline{\boldsymbol{\Lambda}}(\boldsymbol{x}), \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon, \boldsymbol{x})) + \overline{\boldsymbol{\Lambda}}^{-1/2}(\boldsymbol{x})L_{1/2}(\overline{\boldsymbol{\Lambda}}(\boldsymbol{x}), \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon, \boldsymbol{x}))\overline{\boldsymbol{\Lambda}}(\boldsymbol{x})^{1/2} = \overline{\boldsymbol{\Lambda}}^{-1/2}(\boldsymbol{x})\overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon, \boldsymbol{x}).
$$

Hence:

$$
\text{trace}\left(L_{1/2}(\overline{\boldsymbol{\Lambda}}(\boldsymbol{x}), \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon, \boldsymbol{x}))\right) = \frac{1}{2}\text{trace}\left(\overline{\boldsymbol{\Lambda}}^{-1/2}(\boldsymbol{x})\overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon, \boldsymbol{x})\right).
$$

Using the definition of the Fréchet derivative we have:

$$
\begin{aligned}
& V(\boldsymbol{x}+\epsilon\boldsymbol{v}) - V(\boldsymbol{x}) \\
= \; & -\frac{1}{2}\text{trace}\left(\overline{\boldsymbol{\Lambda}}^{-1/2}(\boldsymbol{x})\overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})\right) + o\left(||\overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})||\right) \\
= \; & -\frac{1}{2}\text{trace}\left(\begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \overline{\boldsymbol{\Lambda}}^{-1/2}(\boldsymbol{x}) \end{pmatrix}\begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x}) \end{pmatrix}\right) + o\left(||\overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})||\right) \\
= \; & -\frac{1}{2}\text{trace}\left(\begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \overline{\boldsymbol{\Lambda}}^{-1/2}(\boldsymbol{x}) \end{pmatrix}\boldsymbol{P}^T\boldsymbol{E}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})\boldsymbol{P}\right) + o\left(||\overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})||\right) \\
= \; & -\frac{1}{2}\text{trace}\left(\boldsymbol{P}\begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \overline{\boldsymbol{\Lambda}}^{-1/2}(\boldsymbol{x}) \end{pmatrix}\boldsymbol{P}^T\boldsymbol{E}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})\right) + o\left(||\overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})||\right) \\
= \; & -\frac{\epsilon}{2}\text{trace}\left((\boldsymbol{T}^{1/2}(\boldsymbol{x}))^{\dagger}\boldsymbol{\Sigma}^{1/2}\left(\text{Diag}(\boldsymbol{v}) - \boldsymbol{x}\boldsymbol{v}^T - \boldsymbol{v}\boldsymbol{x}^T\right)\boldsymbol{\Sigma}^{1/2}\right) + o\left(||\overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})||\right) \\
= \; & -\frac{\epsilon}{2}\left(\text{diag}\left(\boldsymbol{\Sigma}^{1/2}(\boldsymbol{T}^{1/2}(\boldsymbol{x}))^{\dagger}\boldsymbol{\Sigma}^{1/2}\right)^T - 2\boldsymbol{x}^T\boldsymbol{\Sigma}^{1/2}(\boldsymbol{T}^{1/2}(\boldsymbol{x}))^{\dagger}\boldsymbol{\Sigma}^{1/2}\right)\boldsymbol{v} \\
& + o\left(||\overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})||\right).
\end{aligned}
$$

Hence, we obtain the expression for the directional derivative in the direction $\boldsymbol{v} \in \overline{\Delta}_{n-1}$ as:

$$
\begin{aligned}
\boldsymbol{\nabla}_{\boldsymbol{v}}V(\boldsymbol{x}) &= \lim_{\epsilon\to 0}\frac{V(\boldsymbol{x}+\epsilon\boldsymbol{v}) - V(\boldsymbol{x})}{\epsilon} \\
&= -\frac{1}{2}\left(\text{diag}\left(\boldsymbol{\Sigma}^{1/2}(\boldsymbol{T}^{1/2}(\boldsymbol{x}))^{\dagger}\boldsymbol{\Sigma}^{1/2}\right)^T - 2\boldsymbol{x}^T\boldsymbol{\Sigma}^{1/2}(\boldsymbol{T}^{1/2}(\boldsymbol{x}))^{\dagger}\boldsymbol{\Sigma}^{1/2}\right)\boldsymbol{v} \\
&= \boldsymbol{g}(\boldsymbol{x})^T\boldsymbol{v},
\end{aligned}
$$

where $\boldsymbol{g}(\boldsymbol{x}) = -\frac{1}{2}\left(\text{diag}\left(\boldsymbol{\Sigma}^{1/2}(\boldsymbol{T}^{1/2}(\boldsymbol{x}))^{\dagger}\boldsymbol{\Sigma}^{1/2}\right) - 2\boldsymbol{\Sigma}^{1/2}(\boldsymbol{T}^{1/2}(\boldsymbol{x}))^{\dagger}\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\right)$. Since this is true for all $\boldsymbol{v} \in \overline{\Delta}_{n-1}$, we obtain $\overline{\boldsymbol{\nabla}}V(\boldsymbol{x})$ by projecting $\boldsymbol{g}(\boldsymbol{x})$ onto the tangent space $\overline{\Delta}_{n-1}$. [4] □

[4] We abuse the notation slightly here since the gradient of $V(\boldsymbol{x})$ does not exist outside the feasible region. For all theoretical and algorithmic purposes, the mathematical quantity that we calculate, i.e., $\overline{\boldsymbol{\nabla}}V(\boldsymbol{x})$, behaves as the projected gradient. To achieve mathematical rigor, one would embed the function into the affine subspace $\boldsymbol{e}^T\boldsymbol{x} = 1$ by substituting for one of the decision variables, but this approach is notationally cumbersome in exposition.

Proof of Theorem 3.1.7

Suppose that the sequence of interior points $\{\boldsymbol{x}_k\}_{k=1,\ldots\infty}$ approaches a point $\hat{\boldsymbol{x}}$ on the relative boundary of the unit simplex, along the direction $-\boldsymbol{z} \in \mathbb{R}^n$. Assume that $\hat{\boldsymbol{x}}$ has exactly m zeros. Any such $\boldsymbol{z}$ must satisfy $\boldsymbol{e}^T\boldsymbol{z} = 0$ and $\hat{x}_i = 0 \Rightarrow z_i > 0$. We first prove that:

$$\lim_{t\to 0+} \frac{V(\hat{\boldsymbol{x}}) - V(\hat{\boldsymbol{x}} + t\boldsymbol{z})}{t} = +\infty.$$

Let $z_0 = \min_{i:\hat{x}_i=0} z_i$ and $\sigma_1 = \lambda_1(\boldsymbol{\Sigma})$. We have:

$$\begin{aligned} \boldsymbol{T}(\hat{\boldsymbol{x}} + t\boldsymbol{z}) &= \boldsymbol{\Sigma}^{1/2}\boldsymbol{S}(\hat{\boldsymbol{x}} + t\boldsymbol{z})\boldsymbol{\Sigma}^{1/2} \\ &= \boldsymbol{\Sigma}^{1/2}\mathrm{Diag}(\hat{\boldsymbol{x}} + t\boldsymbol{z})\boldsymbol{\Sigma}^{1/2} - \boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{x}} + t\boldsymbol{z})(\hat{\boldsymbol{x}} + t\boldsymbol{z})^T\boldsymbol{\Sigma}^{1/2}. \end{aligned}$$

It is clear that $\min_i\{\hat{x}_i + tz_i\} = \min_{i:\hat{x}_i=0} tz_i = tz_0$ if t is sufficiently small. From Lemmas 3.5.2 and 3.5.3, both given in the section 3.5, this implies that

$$\lambda_2(\boldsymbol{\Sigma}^{1/2}\boldsymbol{S}(\hat{\boldsymbol{x}} + t\boldsymbol{z})\boldsymbol{\Sigma}^{1/2}) \geq \lambda_1(\boldsymbol{\Sigma}^{1/2}\mathrm{Diag}(\hat{\boldsymbol{x}} + t\boldsymbol{z})\boldsymbol{\Sigma}^{1/2}) \geq t\sigma_1 z_0 > 0$$

since $\boldsymbol{\Sigma}^{1/2}\boldsymbol{S}(\hat{\boldsymbol{x}} + t\boldsymbol{z})\boldsymbol{\Sigma}^{1/2}$ is a rank one update of $\boldsymbol{\Sigma}^{1/2}\mathrm{Diag}(\hat{\boldsymbol{x}} + t\boldsymbol{z})\boldsymbol{\Sigma}^{1/2}$. Therefore, together with Lemma 3.1.5, we have:

$$\lambda_1(\boldsymbol{T}(\hat{\boldsymbol{x}} + t\boldsymbol{z})) = 0,\ \lambda_2(\boldsymbol{T}(\hat{\boldsymbol{x}} + t\boldsymbol{z})) \geq t\sigma_1 z_0$$

and

$$\lambda_3(\boldsymbol{T}(\hat{\boldsymbol{x}} + t\boldsymbol{z})), \ldots, \lambda_n(\boldsymbol{T}(\hat{\boldsymbol{x}} + t\boldsymbol{z})) > 0.$$

Recall also that Lemma 3.1.5 gives:

$$\lambda_1(\boldsymbol{T}(\hat{\boldsymbol{x}})) = \cdots = \lambda_{m+1}(\boldsymbol{T}(\hat{\boldsymbol{x}})) = 0 \text{ and } \lambda_{m+2}(\boldsymbol{T}(\hat{\boldsymbol{x}})), \ldots, \lambda_n(\boldsymbol{T}(\hat{\boldsymbol{x}})) > 0.$$

Furthermore, from standard results for the continuity of the matrix eigenvalue function (see Golub and Loan 1996), we know that $\|\lambda_j(\boldsymbol{T}(\hat{\boldsymbol{x}} + t\boldsymbol{z})) - \lambda_j(\boldsymbol{T}(\hat{\boldsymbol{x}}))\| \leq O(t)$, for all

j. Combining above facts,

$$\begin{aligned} & \lim_{t\to 0_+} \frac{V(\hat{\boldsymbol{x}}) - V(\hat{\boldsymbol{x}} + t\boldsymbol{z})}{t} \\ = & \lim_{t\to 0_+} \frac{\sum_{j=1}^{n} \left(\sqrt{\lambda_j(\boldsymbol{T}(\hat{\boldsymbol{x}} + t\boldsymbol{z}))} - \sqrt{\lambda_j(\boldsymbol{T}(\hat{\boldsymbol{x}}))} \right)}{t} \\ \geq & \lim_{t\to 0_+} \frac{\sqrt{t\sigma_1 z_0} + \sum_{j=m+2}^{n} \left(\sqrt{\lambda_j(\boldsymbol{T}(\hat{\boldsymbol{x}})) + O(t)} - \sqrt{\lambda_j(\boldsymbol{T}(\hat{\boldsymbol{x}}))} \right)}{t} \\ \to & +\infty, \end{aligned}$$

since the ratio $\sqrt{t}/t$ diverges to $+\infty$ as t approaches 0 and $\lim_{t\to 0_+} O(t)/t = 0$.

We next show that

$$\lim_{t\to 0+} |\boldsymbol{\nabla}_{\boldsymbol{z}} V(\hat{\boldsymbol{x}} + t\boldsymbol{z})| = \lim_{t\to 0+} \lim_{s\to 0^+} \left(\frac{V(\hat{\boldsymbol{x}} + t\boldsymbol{z}) - V(\hat{\boldsymbol{x}} + (t+s)\boldsymbol{z})}{s} \right) = +\infty.$$

Suppose this is not the case. Since $V(\hat{\boldsymbol{x}})$ is convex, $\boldsymbol{\nabla}_{\boldsymbol{z}} V(\hat{\boldsymbol{x}} + t\boldsymbol{z})$ is monotone in t, which implies that

$$\begin{aligned} & \liminf_{t\to 0+} \lim_{s\to 0^+} \left(\tfrac{V(\hat{\boldsymbol{x}}+t\boldsymbol{z}) - V(\hat{\boldsymbol{x}}+(t+s)\boldsymbol{z})}{s} \right) \\ = & \limsup_{t\to 0+} \lim_{s\to 0^+} \left(\tfrac{V(\hat{\boldsymbol{x}}+t\boldsymbol{z}) - V(\hat{\boldsymbol{x}}+(t+s)\boldsymbol{z})}{s} \right). \end{aligned}$$

Therefore, there must exists a finite M such that

$$\lim_{s\to 0^+} \left(\frac{V(\hat{\boldsymbol{x}} + t\boldsymbol{z}) - V(\hat{\boldsymbol{x}} + (t+s)\boldsymbol{z})}{s} \right) \leq M, \quad \forall t > 0 \text{ with } \hat{\boldsymbol{x}} + t\boldsymbol{z} \in \text{ int}(\Delta_{n-1}). \quad (3.16)$$

However, since $\lim_{t\to 0+} \frac{V(\hat{\boldsymbol{x}}) - V(\hat{\boldsymbol{x}}+t\boldsymbol{z})}{t} = +\infty$, there exists $t_0 > 0$ such that

$$\frac{V(\hat{\boldsymbol{x}}) - V(\hat{\boldsymbol{x}} + t_0\boldsymbol{z})}{t_0} > 2M.$$

In addition, since V is continuous on Δ_{n-1}, it is also uniformly continuous on Δ_{n-1}. Therefore, there exists $\delta > 0$ such that $|V(\hat{\boldsymbol{x}}) - V(\hat{\boldsymbol{x}} + s\boldsymbol{z})| < t_0 M$ for all $s \in [0, \delta)$. It

is without loss of generality to assume that $\delta < t_0$. Therefore,

$$\begin{aligned}
\lim_{s\to 0^+}\left(\frac{V(\hat{\boldsymbol{x}}+\frac{\delta}{2}\boldsymbol{z})-V(\hat{\boldsymbol{x}}+(\frac{\delta}{2}+s)\boldsymbol{z})}{s}\right) &\geq \left(\frac{V(\hat{\boldsymbol{x}}+\frac{\delta}{2}\boldsymbol{z})-V(\hat{\boldsymbol{x}}+t_0\boldsymbol{z})}{t_0-\frac{\delta}{2}}\right)\\
&\geq \left(\frac{V(\hat{\boldsymbol{x}}+\frac{\delta}{2}\boldsymbol{z})-V(\hat{\boldsymbol{x}}+t_0\boldsymbol{z})}{t_0}\right)\\
&> \left(\frac{V(\hat{\boldsymbol{x}})-t_0M-V(\hat{\boldsymbol{x}}+t_0\boldsymbol{z})}{t_0}\right)\\
&> M.
\end{aligned}$$

This is in contradiction with equation (3.16). Therefore, the theorem is proven. □

Proof of Lemma 3.2.1

From the definition of d_k, we have

$$\begin{aligned}
d_{k+1}^2 &= ||\boldsymbol{x}^{k+1}-\boldsymbol{x}^*||^2\\
&= ||\boldsymbol{x}^k+\alpha_k\overline{\boldsymbol{\nabla}}f(\boldsymbol{x}^k)-\boldsymbol{x}^*||^2\\
&= d_k^2+\alpha_k^2||\overline{\boldsymbol{\nabla}}f(\boldsymbol{x}^k)||^2-2\alpha_k\overline{\boldsymbol{\nabla}}f(\boldsymbol{x}^k)^T(\boldsymbol{x}^*-\boldsymbol{x}^k).
\end{aligned} \tag{3.17}$$

Since $f(\cdot)$ is concave, we have

$$\overline{\boldsymbol{\nabla}}f(\boldsymbol{x}^k)^T(\boldsymbol{x}^*-\boldsymbol{x}^k)\geq f(\boldsymbol{x}^*)-f(\boldsymbol{x}^k)=\epsilon_k.$$

Note that $\epsilon_k \geq 0$ by definition. Combined with inequality (3.17), we have

$$d_{k+1}^2 \leq d_k^2-\alpha_k(2\epsilon_k-\alpha_k||\overline{\boldsymbol{\nabla}}f(\boldsymbol{x}^k)||^2).$$

By the result of the Armijo's rule, we have

$$f(\boldsymbol{x}^{k+1})\geq f(\boldsymbol{x}^k)+\tau\alpha_k\|\overline{\boldsymbol{\nabla}}f(\boldsymbol{x}^k)\|^2.$$

Therefore,

$$\alpha_k\|\overline{\boldsymbol{\nabla}}f(\boldsymbol{x}^k)\|^2\leq\frac{1}{\tau}(f(\boldsymbol{x}^{k+1})-f(\boldsymbol{x}^k))=\frac{1}{\tau}(\epsilon_k-\epsilon_{k+1}).$$

Hence, we have

$$\begin{aligned} d_{k+1}^2 &\leq d_k^2 - \alpha_k(2\epsilon_k - \alpha_k||\overline{\boldsymbol{\nabla}} f(\boldsymbol{x}^k)||^2) \\ &\leq d_k^2 - \alpha_k \left(2\epsilon_k - \frac{1}{\tau}(\epsilon_k - \epsilon_{k+1})\right) \\ &\leq d_k^2 - \alpha_k \left(2\epsilon_k - 2(\epsilon_k - \epsilon_{k+1})\right) \\ &\leq d_k^2 - 2\alpha_k \epsilon_{k+1} \\ &\leq d_k^2, \end{aligned}$$

where the third inequality uses the fact that $\tau \geq 0.5$. □

Note that the above proof is very similar to the proof of Proposition 9.1.2 in Ben-Tal and Nemirovski (2001) for the unconstrained convex case.

Lemmas for Sections 3.1.3 and 3.2

Lemma 3.5.1 *Suppose* $\boldsymbol{x} > \boldsymbol{0}$ *and let* $x_{\min} = \min_i x_i$, $x_{\max} = \max_i x_i$. *Then:*

$$||L_{1/2}(Diag(\boldsymbol{x}), \boldsymbol{E})|| \leq \frac{||\boldsymbol{E}||}{2x_{\min}^{1/2}},$$

and

$$||L_{-1/2}(Diag(\boldsymbol{x}), \boldsymbol{E})|| \leq \frac{n||\boldsymbol{E}||}{2x_{\min}^{3/2}}.$$

Proof: The Fréchet derivative for the matrix inverse function is given as:

$$L_{-1}(\boldsymbol{X}, \boldsymbol{E}) = -\boldsymbol{X}^{-1}\boldsymbol{E}\boldsymbol{X}^{-1}.$$

The Fréchet derivative for the matrix square root function, which exists when $\boldsymbol{X}$ is positive definite, is the unique solution to the Sylvester equation (refer to Kenney and Laub 1989, Higham 2008):

$$\boldsymbol{X}^{1/2}L_{1/2}(\boldsymbol{X}, \boldsymbol{E}) + L_{1/2}(\boldsymbol{X}, \boldsymbol{E})\boldsymbol{X}^{1/2} = \boldsymbol{E}. \tag{3.18}$$

Following the chain rule (Theorem 3.4 in Higham Higham (2008)), we have

$$L_{-1/2}(\boldsymbol{X}, \boldsymbol{E}) = L_{1/2}(\boldsymbol{X}^{-1}, L_{-1}(\boldsymbol{X}, \boldsymbol{E})) = L_{1/2}(\boldsymbol{X}^{-1}, -\boldsymbol{X}^{-1}\boldsymbol{E}\boldsymbol{X}^{-1}),$$

and therefore,

$$\boldsymbol{X}^{-1/2}L_{-1/2}(\boldsymbol{X}, \boldsymbol{E}) + L_{-1/2}(\boldsymbol{X}, \boldsymbol{E})\boldsymbol{X}^{-1/2} = -\boldsymbol{X}^{-1}\boldsymbol{E}\boldsymbol{X}^{-1}. \tag{3.19}$$

Combining equations (3.18) and (3.19), we have

$$L_{-1/2}(\boldsymbol{X}, \boldsymbol{E}) = -\boldsymbol{X}^{-1/2}L_{1/2}(\boldsymbol{X}, \boldsymbol{E})\boldsymbol{X}^{-1/2}. \tag{3.20}$$

Define $\boldsymbol{L} = L_{1/2}(\text{Diag}(\boldsymbol{x}), \boldsymbol{E})$. From equation (3.18), we have:

$$\text{Diag}(\boldsymbol{x})^{1/2}\boldsymbol{L} + \boldsymbol{L}\text{Diag}(\boldsymbol{x})^{1/2} = \boldsymbol{E},$$

which implies that

$$L_{i,j} = \frac{E_{i,j}}{x_i^{1/2} + x_j^{1/2}} \leq \frac{E_{i,j}}{2x_{\min}^{1/2}}.$$

Therefore, we have

$$||L|| = \sqrt{\sum_{i,j} L_{i,j}^2} \leq \sqrt{\sum_{i,j} \frac{E_{i,j}^2}{4x_{\min}}} = \frac{||\boldsymbol{E}||}{2x_{\min}^{1/2}}.$$

In addition, by equation (3.20), we have

$$\begin{aligned} ||L_{-1/2}(\text{Diag}(\boldsymbol{x}), \boldsymbol{E})|| &= ||\text{Diag}(\boldsymbol{x})^{-1/2}\boldsymbol{L}\text{Diag}(\boldsymbol{x})^{-1/2}|| \\ &\leq ||\text{Diag}(\boldsymbol{x})^{-1/2}||^2||\boldsymbol{V}L|| \\ &= \frac{n||\boldsymbol{E}||}{2x_{\min}^{3/2}}. \end{aligned}$$

□

Lemma 3.5.2 *Let* $\boldsymbol{B} = \boldsymbol{A} - \boldsymbol{u}\boldsymbol{u}^T$. *Then* $\lambda_1(\boldsymbol{B}) \leq \lambda_1(\boldsymbol{A})$, *and*

$$\lambda_{i-1}(\boldsymbol{A}) \leq \lambda_i(\boldsymbol{B}) \leq \lambda_i(\boldsymbol{A}), \quad \forall i = 2, \ldots, n.$$

Proof: The proof can be found on page 97-98 of Wilkinson (1965).

Lemma 3.5.3 *Let* $\boldsymbol{D} = Diag(d_1, ..., d_n)$ *be a diagonal matrix with* $d_i > 0, \forall i \in \mathcal{N}$. *Let* $\boldsymbol{\Sigma}$ *be a positive definite matrix. Then* $\lambda_1(\boldsymbol{\Sigma}^{1/2}\boldsymbol{D}\boldsymbol{\Sigma}^{1/2}) \geq \lambda_1(\boldsymbol{\Sigma}) \min_i\{d_i\}$.

Proof: Since the set of eigenvalue of AA^T coincides with that of A^TA, we have

$$\lambda_1(\boldsymbol{\Sigma}^{1/2}\boldsymbol{D}\boldsymbol{\Sigma}^{1/2}) = \lambda_1(\boldsymbol{D}^{1/2}\boldsymbol{\Sigma}\boldsymbol{D}^{1/2}).$$

But

$$\boldsymbol{D}^{1/2}\boldsymbol{\Sigma}\boldsymbol{D}^{1/2} \succeq \lambda_1(\boldsymbol{\Sigma})\boldsymbol{D}^{1/2}\boldsymbol{I}\boldsymbol{D}^{1/2} = \lambda_1(\boldsymbol{\Sigma})\boldsymbol{D}.$$

Therefore, for all $\boldsymbol{v}$ with $||\boldsymbol{v}|| = 1$, we have

$$\boldsymbol{v}^T\boldsymbol{D}^{1/2}\boldsymbol{\Sigma}\boldsymbol{D}^{1/2}\boldsymbol{v} \geq \boldsymbol{v}^T\lambda_1(\boldsymbol{\Sigma})\boldsymbol{D}\boldsymbol{v} \geq \lambda_1(\boldsymbol{\Sigma}) \min_i\{d_i\}.$$

□

Theorem 3.5.4 *Assume that* $\boldsymbol{\Sigma} \succ 0$. *For any feasible direction* $\boldsymbol{v} \in \overline{\Delta}_{n-1}$ *with* $\|\boldsymbol{v}\| = 1$ *and* $\boldsymbol{x} \in int(\Delta_{n-1})$,

$$|f''_{\boldsymbol{x},\boldsymbol{v}}(0)| \leq \frac{9n}{4x_{\min}\sigma_1}\left\|\boldsymbol{\Sigma}^{1/2}\right\|^4 + \frac{n}{(x_{\min}\sigma_1)^{1/2}}\left\|\boldsymbol{\Sigma}^{1/2}\right\|^2,$$

where $f_{\boldsymbol{x},\boldsymbol{v}}(t) = V(\boldsymbol{x} + t\boldsymbol{v}) > 0$, $\sigma_1 = \lambda_1(\boldsymbol{\Sigma})$ *and* $x_{\min} = \min_i\ x_i$.

Proof: Let $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{g}_1(\boldsymbol{x}) + \boldsymbol{g}_2(\boldsymbol{x})$ where

$$\boldsymbol{g}_1(\boldsymbol{x}) = -\frac{1}{2}\text{diag}\left(\boldsymbol{\Sigma}^{1/2}(\boldsymbol{T}^{1/2}(\boldsymbol{x}))^\dagger\boldsymbol{\Sigma}^{1/2}\right), \quad \boldsymbol{g}_2(\boldsymbol{x}) = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{T}^{1/2}(\boldsymbol{x}))^\dagger\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}.$$

From the proof of Theorem 3.1.6, we know that:

$$
\begin{aligned}
& \boldsymbol{g}_1(\boldsymbol{x}+\epsilon\boldsymbol{v})^T\boldsymbol{v} - \boldsymbol{g}_1(\boldsymbol{x})^T\boldsymbol{v} \\
= \; & -\frac{1}{2}\boldsymbol{v}^T \mathrm{diag}\left(\boldsymbol{\Sigma}^{1/2}(\boldsymbol{T}^{1/2}(\boldsymbol{x}+\epsilon\boldsymbol{v}))^\dagger\boldsymbol{\Sigma}^{1/2}\right) + \frac{1}{2}\boldsymbol{v}^T \mathrm{diag}\left(\boldsymbol{\Sigma}^{1/2}(\boldsymbol{T}^{1/2}(\boldsymbol{x}))^\dagger\boldsymbol{\Sigma}^{1/2}\right) \\
= \; & -\frac{1}{2}\boldsymbol{v}^T \mathrm{diag}\left(\boldsymbol{\Sigma}^{1/2}\left((\boldsymbol{T}^{1/2}(\boldsymbol{x}+\epsilon\boldsymbol{v}))^\dagger - (\boldsymbol{T}^{1/2}(\boldsymbol{x}))^\dagger\right)\boldsymbol{\Sigma}^{1/2}\right) \\
= \; & -\frac{1}{2}\boldsymbol{v}^T \mathrm{diag}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{P}\begin{pmatrix} 0 & \boldsymbol{0}^T \\ \boldsymbol{0} & \left(\overline{\boldsymbol{\Lambda}}(\boldsymbol{x}) + \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})\right)^{-1/2} - \overline{\boldsymbol{\Lambda}}(\boldsymbol{x})^{-1/2} \end{pmatrix}\boldsymbol{P}^T\boldsymbol{\Sigma}^{1/2}\right) \\
= \; & -\frac{1}{2}\boldsymbol{v}^T \mathrm{diag}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{P}\begin{pmatrix} 0 & \boldsymbol{0}^T \\ \boldsymbol{0} & L_{-1/2}\left(\overline{\boldsymbol{\Lambda}}(\boldsymbol{x}), \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})\right) + o\left(||\overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})||\right) \end{pmatrix}\boldsymbol{P}^T\boldsymbol{\Sigma}^{1/2}\right) \\
= \; & -\frac{1}{2}\boldsymbol{v}^T \mathrm{diag}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{P}\begin{pmatrix} 0 & \boldsymbol{0}^T \\ \boldsymbol{0} & L_{-1/2}\left(\overline{\boldsymbol{\Lambda}}(\boldsymbol{x}), \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})\right) \end{pmatrix}\boldsymbol{P}^T\boldsymbol{\Sigma}^{1/2}\right) + o(\epsilon).
\end{aligned}
$$

Therefore, we have:

$$
\begin{aligned}
& ||\boldsymbol{g}_1(\boldsymbol{x}+\epsilon\boldsymbol{v})^T\boldsymbol{v} - \boldsymbol{g}_1(\boldsymbol{x})^T\boldsymbol{v}|| \\
= \;& || -\frac{1}{2}\boldsymbol{v}^T \mathrm{diag}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{P}\begin{pmatrix} 0 & \boldsymbol{0}^T \\ \boldsymbol{0} & L_{-1/2}\left(\overline{\boldsymbol{\Lambda}}(\boldsymbol{x}), \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})\right)\end{pmatrix}\boldsymbol{P}^T\boldsymbol{\Sigma}^{1/2}\right) + o(\epsilon)|| \\
\leq \;& \frac{1}{2}||\boldsymbol{v}|| \cdot \left\|\boldsymbol{\Sigma}^{1/2}\right\|^2 \cdot \left\|\boldsymbol{P}\begin{pmatrix} 0 & \boldsymbol{0}^T \\ \boldsymbol{0} & L_{-1/2}\left(\overline{\boldsymbol{\Lambda}}(\boldsymbol{x}), \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})\right)\end{pmatrix}\boldsymbol{P}^T\right\| + o(\epsilon) \\
\leq \;& \frac{1}{2}\left\|\boldsymbol{\Sigma}^{1/2}\right\|^2 \cdot \left\|\boldsymbol{P}\begin{pmatrix} 0 & \boldsymbol{0}^T \\ \boldsymbol{0} & L_{-1/2}\left(\overline{\boldsymbol{\Lambda}}(\boldsymbol{x}), \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})\right)\end{pmatrix}\boldsymbol{P}^T\right\| + o(\epsilon) \\
\leq \;& \frac{1}{2}\left\|\boldsymbol{\Sigma}^{1/2}\right\|^2 \cdot \left\|L_{-1/2}\left(\overline{\boldsymbol{\Lambda}}(\boldsymbol{x}), \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})\right)\right\| + o(\epsilon) \\
\leq \;& \frac{1}{2}\left\|\boldsymbol{\Sigma}^{1/2}\right\|^2 \cdot \frac{n||\overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})||}{2\lambda_2^{3/2}(\boldsymbol{T}(\boldsymbol{x}))} + o(\epsilon) \\
= \;& \frac{1}{2}\left\|\boldsymbol{\Sigma}^{1/2}\right\|^2 \cdot \frac{n||\boldsymbol{E}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})||}{2\lambda_2^{3/2}(\boldsymbol{T}(\boldsymbol{x}))} + o(\epsilon) \\
= \;& \frac{1}{2}\left\|\boldsymbol{\Sigma}^{1/2}\right\|^2 \cdot \frac{n||\epsilon\boldsymbol{\Sigma}^{1/2}\left(\mathrm{Diag}(\boldsymbol{v}) - \boldsymbol{x}\boldsymbol{v}^T - \boldsymbol{v}\boldsymbol{x}^T\right)\boldsymbol{\Sigma}^{1/2}||}{2\lambda_2^{3/2}(\boldsymbol{T}(\boldsymbol{x}))} + o(\epsilon) \\
\leq \;& \frac{\epsilon}{2}\left\|\boldsymbol{\Sigma}^{1/2}\right\|^4 \cdot \frac{n||\boldsymbol{v}||(1+2||\boldsymbol{x}||)}{2\lambda_2^{3/2}(\boldsymbol{T}(\boldsymbol{x}))} + o(\epsilon) \\
\leq \;& \frac{\epsilon}{2}\left\|\boldsymbol{\Sigma}^{1/2}\right\|^4 \cdot \frac{3n}{2\lambda_2^{3/2}(\boldsymbol{T}(\boldsymbol{x}))} + o(\epsilon).
\end{aligned}
$$

Note that the last inequality holds since $||\boldsymbol{x}|| \leq 1$ for all $\boldsymbol{x} \in \Delta_{n-1}$. On the other hand,

$$\begin{aligned}
& \boldsymbol{g}_2(\boldsymbol{x}+\epsilon\boldsymbol{v})^T\boldsymbol{v} - \boldsymbol{g}_2(\boldsymbol{x})^T\boldsymbol{v} \\
= \;& \boldsymbol{v}^T\boldsymbol{\Sigma}^{1/2}(\boldsymbol{T}^{1/2}(\boldsymbol{x}+\epsilon\boldsymbol{v}))^\dagger\boldsymbol{\Sigma}^{1/2}(\boldsymbol{x}+\epsilon\boldsymbol{v}) - \boldsymbol{v}^T\boldsymbol{\Sigma}^{1/2}(\boldsymbol{T}^{1/2}(\boldsymbol{x}))^\dagger\boldsymbol{\Sigma}^{1/2}\boldsymbol{x} \\
= \;& \boldsymbol{v}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{P}\begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \left(\overline{\boldsymbol{\Lambda}}(\boldsymbol{x}) + \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})\right)^{-1/2} \end{pmatrix}\boldsymbol{P}^T\boldsymbol{\Sigma}^{1/2}(\boldsymbol{x}+\epsilon\boldsymbol{v}) \\
& -\boldsymbol{v}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{P}\begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \overline{\boldsymbol{\Lambda}}(\boldsymbol{x})^{-1/2} \end{pmatrix}\boldsymbol{P}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{x} \\
= \;& \boldsymbol{v}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{P}\begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & L_{-1/2}\left(\overline{\boldsymbol{\Lambda}}(\boldsymbol{x}), \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})\right) \end{pmatrix}\boldsymbol{P}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{x} \\
& +\epsilon\boldsymbol{v}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{P}\begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \overline{\boldsymbol{\Lambda}}(\boldsymbol{x})^{-1/2} \end{pmatrix}\boldsymbol{P}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{v} + o(\epsilon).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
& \left\|\boldsymbol{g}_2(\boldsymbol{x}+\epsilon\boldsymbol{v})^T\boldsymbol{v} - \boldsymbol{g}_2(\boldsymbol{x})^T\boldsymbol{v}\right\| \\
\leq \;& \left\|\boldsymbol{v}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{P}\begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & L_{-1/2}\left(\overline{\boldsymbol{\Lambda}}(\boldsymbol{x}), \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})\right) \end{pmatrix}\boldsymbol{P}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\right\| \\
& +\epsilon\left\|\boldsymbol{v}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{P}\begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \overline{\boldsymbol{\Lambda}}(\boldsymbol{x})^{-1/2} \end{pmatrix}\boldsymbol{P}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\right\| + o(\epsilon) \\
\leq \;& \|\boldsymbol{x}\|\left\|\boldsymbol{\Sigma}^{1/2}\right\|^2\left\|L_{-1/2}\left(\overline{\boldsymbol{\Lambda}}(\boldsymbol{x}), \overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})\right)\right\| + \epsilon\left\|\boldsymbol{\Sigma}^{1/2}\right\|^2\left\|\overline{\boldsymbol{\Lambda}}(\boldsymbol{x})^{-1/2}\right\| + o(\epsilon) \\
\leq \;& \left\|\boldsymbol{\Sigma}^{1/2}\right\|^2 \cdot \frac{n||\overline{\boldsymbol{E}}_{\boldsymbol{v}}(\epsilon,\boldsymbol{x})||}{2\lambda_2^{3/2}(\boldsymbol{T}(\boldsymbol{x}))} + \epsilon\left\|\boldsymbol{\Sigma}^{1/2}\right\|^2\frac{n}{\lambda_2(\boldsymbol{T}(\boldsymbol{x}))^{1/2}} + o(\epsilon) \\
\leq \;& \epsilon\left\|\boldsymbol{\Sigma}^{1/2}\right\|^4 \cdot \frac{3n}{2\lambda_2^{3/2}(\boldsymbol{T}(\boldsymbol{x}))} + \epsilon\left\|\boldsymbol{\Sigma}^{1/2}\right\|^2\frac{n}{\lambda_2(\boldsymbol{T}(\boldsymbol{x}))^{1/2}} + o(\epsilon).
\end{aligned}$$

In addition, from Lemma 3.5.2 and Lemma 3.5.3, we have

$$\lambda_2(\boldsymbol{T}(\boldsymbol{x})) = \lambda_2(\boldsymbol{\Sigma}^{1/2}\boldsymbol{S}(\boldsymbol{x})\boldsymbol{\Sigma}^{1/2}) \geq \lambda_1(\boldsymbol{\Sigma}^{1/2}\mathrm{Diag}(\boldsymbol{x})\boldsymbol{\Sigma}^{1/2}) \geq x_{\min}\sigma_1 > 0.$$

Therefore, we have

$$\begin{aligned}
\left|f''_{\boldsymbol{x},\boldsymbol{v}}(0)\right| &= \left|\lim_{\epsilon\to 0^+} \frac{\boldsymbol{g}_1(\boldsymbol{x}+\epsilon\boldsymbol{v})^T\boldsymbol{v} - \boldsymbol{g}_1(\boldsymbol{x})^T\boldsymbol{v} + \boldsymbol{g}_2(\boldsymbol{x}+\epsilon\boldsymbol{v})^T\boldsymbol{v} - \boldsymbol{g}_2(\boldsymbol{x})^T\boldsymbol{v}}{\epsilon}\right| \\
&\leq \left|\lim_{\epsilon\to 0^+} \frac{\boldsymbol{g}_1(\boldsymbol{x}+\epsilon\boldsymbol{v})^T\boldsymbol{v} - \boldsymbol{g}_1(\boldsymbol{x})^T\boldsymbol{v}}{\epsilon}\right| + \left|\lim_{\epsilon\to 0^+} \frac{\boldsymbol{g}_2(\boldsymbol{x}+\epsilon\boldsymbol{v})^T\boldsymbol{v} - \boldsymbol{g}_2(\boldsymbol{x})^T\boldsymbol{v}}{\epsilon}\right| \\
&\leq \frac{9n}{4\lambda_2^{3/2}(\boldsymbol{T}(\boldsymbol{x}))}\left\|\boldsymbol{\Sigma}^{1/2}\right\|^4 + \frac{n}{\lambda_2(\boldsymbol{T}(\boldsymbol{x}))^{1/2}}\left\|\boldsymbol{\Sigma}^{1/2}\right\|^2 \\
&\leq \frac{9n}{4(x_{\min}\sigma_1)^{3/2}}\left\|\boldsymbol{\Sigma}^{1/2}\right\|^4 + \frac{n}{(x_{\min}\sigma_1)^{1/2}}\left\|\boldsymbol{\Sigma}^{1/2}\right\|^2.
\end{aligned}$$

□

Remark 3.5.5 *Theorem 3.5.4 develops an upper bound for the absolute value of the second order derivatives in direction* $\boldsymbol{v} \in \overline{\Delta}_{n-1}$. *The significance of the theorem is that the second order derivative of* $V(\boldsymbol{x})$ *is bounded for all points* $\boldsymbol{x} \in int(\Delta_{n-1})$, *with the bound associated with the minimum components of* $\boldsymbol{x}$.

Chapter 4

Analysis of Discrete Choice Models: A Welfare-Based Approach

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 perspective to view choice models by treating the welfare function as the primitive. We call the resulting choice model the *welfare-based choice model.* The welfare-based choice model is meaningful on its own by providing an alternative way to construct choice models. Moreover, it provides considerable convenience in analyzing the connections among existing choice models. By using convex analysis theory, we prove that the welfare-based choice model is equivalent to the representative agent choice model and the semi-parametric choice model, thus establishing the equivalence of the latter two as well. We show that these three models are all strictly more general than the random utility model. In case when there are only two alternatives, those four models are equivalent. Moreover, the welfare-based choice model subsumes the nested logit model with positive dissimilarity parameters. We then introduce a new notion for choice modeling: substitutability/complementarity between alternatives. We show that the random utility model only allows substitutability between different alternatives, while the welfare-based choice model allows more flexible substitutability/complementarity

patterns. We argue that such flexibility is desirable in capturing certain practical choice patterns, such as the halo effects. We also present ways to construct new choice models using our approach.

4.1 Research Questions and Main Contribution

In the previous chapter, we established a closed-form representative agent model for the CMM model. Together with the MDM and MMM model proposed in Natarajan et al. (2009), we understand that most of the existing semi-parametric choice models can be reformulated as representative agent models. This motivates us to study the relations between these two models. In particular, we try to answer the following questions in this chapter:

1. What is the relation between the representative agent model and the semi-parametric model? It has been shown that for several special cases, the semi-parametric model can be represented as a representative agent model. However, it is unknown whether this is generally true or not.

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 chapter, we present precise answers to the above questions. We view from another perspective of choice models and consider a *welfare-based approach*. The welfare-based 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*. We show that the welfare-based choice model is not only meaningful on its own, but also offer considerable convenience in establishing connections among existing choice models.

First, by using the welfare-based choice model as an intermediate model, we show that 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 appears to have originated from very different sources. In addition, our proof of the equivalence of these three models is constructive, therefore, it gives a way to convert one model to another constructively, potentially alleviating the pain of establishing correspondences in a case by case manner as is done in the literature.

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.

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. However, in certain applications, it might be desirable to allow some alternatives to exhibit complementarity in certain range, in order to explain certain phenomenon observed in practice, such as the halo effect (or the synergistic effect). Therefore, we derive conditions under which a choice model exhibits substitutable/complementary properties. In addition, we show a few examples of choice models that allow complementarity between alternatives (in a certain range)

and propose a few ways to construct choice models with complementary patterns. As far as we know, this is the first formal study of 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.

The remainder of this chapter is organized as follows. In Section 4.2, we define the welfare-based choice model and study its relation with other choice models. In Section 4.3, we study the relation between the welfare-based choice model and the random utility model. In Section 4.4, we propose the concept of substitutability and complementarity between choice alternatives and derive conditions under which each model exhibits such properties. We show some examples of non-substitutable choice models as well as ways to construct them in Section 4.5. We conclude the chapter in Section 4.6. All the proofs in this chapter are relegated to Section 4.7.

4.2 Welfare-Based Choice Model

In this section, we propose an approach to unify the various choice models reviewed in Chapter 2. We first notice that although the choice models reviewed in Chapter 2 are based on different ideas, they are all essentially functions from a vector of utilities $\boldsymbol{\mu}$ to a vector of choice probabilities $\boldsymbol{q}(\boldsymbol{\mu})$. Moreover, each of these models allows a welfare function $w(\boldsymbol{\mu})$ 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 $w(\boldsymbol{\mu})$ with respect to $\boldsymbol{\mu}$. Our proposed approach is based on these observations. We start by making the following definition:

Definition 5 (Choice Welfare Function) *Let $w(\boldsymbol{\mu})$ be a mapping from $\mathbb{R}^n$ to $\bar{\mathbb{R}}$. We call $w(\boldsymbol{\mu})$ a choice welfare function if $w(\boldsymbol{\mu})$ satisfies the following properties:*

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

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

Here we make a few comments on the three conditions in Definition 5. 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. We will see later 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 obtain from 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 a convex one.

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 4.2.1 *The following statements are equivalent:*

1. *$w(\boldsymbol{\mu})$ is a choice welfare function;*

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

$$w(\boldsymbol{\mu}) = \max \left\{ \boldsymbol{\mu}^T \boldsymbol{x} - V(\boldsymbol{x}) \middle| \, \boldsymbol{x} \in \Delta_{n-1} \right\}; \tag{4.1}$$

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

$$w(\boldsymbol{\mu}) = \sup_{\theta \in \Theta} \ \mathbb{E}_{\boldsymbol{\epsilon} \sim \theta} \left[\max_{i \in \mathcal{N}} \ \mu_i + \epsilon_i \right]. \tag{4.2}$$

The proof of Theorem 4.2.1 uses several results in convex analysis and optimization. In the following, we establish its implication to discrete choice models. In this chapter, we refer to discrete choice models as the entire set of functions $\boldsymbol{q}(\boldsymbol{\mu}) : \mathbb{R}^n \mapsto \Delta_{n-1}$,

mapping a utility vector to a choice probability vector. We first propose the following choice model based on the choice welfare function:

Definition 6 (Welfare-based Choice Model) *Suppose $w(\boldsymbol{\mu})$ is a differentiable choice welfare function. Then the welfare-based choice model derived from $w(\boldsymbol{\mu})$ is defined by*

$$\boldsymbol{q}(\boldsymbol{\mu}) = \nabla w(\boldsymbol{\mu}). \tag{4.3}$$

Note that when $w(\cdot)$ is differentiable, we have $\nabla w(\boldsymbol{\mu}) \in \Delta_{n-1}$ by the translation invariance property of $w(\boldsymbol{\mu})$. Therefore $\boldsymbol{q}(\boldsymbol{\mu})$ defined by (4.3) 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 7 (Proper Function) *A function $f : X \mapsto \bar{\mathbb{R}}$ is proper if $f(\boldsymbol{x}) < \infty$ for at least one $\boldsymbol{x} \in X$ and $f(\boldsymbol{x}) > -\infty$ for all $\boldsymbol{x} \in X$.*

Definition 8 (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*

$$dom(\partial f) = \left\{ \boldsymbol{x} \;\middle|\; \partial f(\boldsymbol{x}) \neq \phi \right\},$$

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

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

Theorem 4.2.2 *For a choice model $\boldsymbol{q} : \mathbb{R}^n \mapsto \Delta_{n-1}$, the following statements are equivalent:*

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

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

3. There exists a distribution set Θ such that

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

The next corollary follows immediately from Theorem 4.2.2 and Propositions 2.2.1 and 2.2.2.

Corollary 4.2.3 *Let $\boldsymbol{q}(\boldsymbol{\mu})$ be a random utility model with absolutely continuous distribution θ and $w(\boldsymbol{\mu})$ be the corresponding expected utility an individual can get under this model. Then $w(\boldsymbol{\mu})$ is a differentiable choice welfare function, and $\boldsymbol{q}(\boldsymbol{\mu}) = \nabla w(\boldsymbol{\mu})$. Moreover, the reverse statement is not true, i.e., there exists a differentiable choice welfare function $w(\boldsymbol{\mu})$ such that there is no random utility model that gives the choice probability $\boldsymbol{q}(\boldsymbol{\mu}) = \nabla w(\boldsymbol{\mu})$.*

In Theorems 4.2.1 and 4.2.2, with the help of the welfare-based choice model, 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.

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 start with. For the welfare-based choice model, one needs to choose a choice welfare function $w(\boldsymbol{\mu})$ 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 4.2.3, 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 4.3 and 4.4.

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 9 (superlinear choice welfare function) *A differentiable choice welfare function $w(\boldsymbol{\mu})$ is called superlinear if there exist $b_i, i = 1, ..., n$, such that for any $\boldsymbol{\mu} \in \mathbb{R}^n$:*

$$w(\boldsymbol{\mu}) \geq \mu_i + b_i, \ \forall \ i = 1, ..., n.$$

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 4.2.4 *For a choice model $\boldsymbol{q} : \mathbb{R}^n \mapsto \Delta_{n-1}$, the following statements are equivalent:*

1. *There exists a superlinear differentiable choice welfare function $w(\boldsymbol{\mu})$ such that $\boldsymbol{q}(\boldsymbol{\mu}) = \nabla w(\boldsymbol{\mu})$;*

2. *There exists an essentially strictly convex function $V(\boldsymbol{x})$ that is upper bounded on Δ_{n-1} such that*

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

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*

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

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

Theorem 4.2.4 further develops the equivalence of choice models obtained in Theorem 4.2.2 by narrowing down the discussion to welfare-based choice models with the superlinear property. In particular, we find that a superlinear differentiable choice welfare function has a semi-parametric representation, of which the distribution set only contains distributions with finite expectation, a property that is desirable in practice. The last statement that $\boldsymbol{q}(\boldsymbol{\mu})$ 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 $\boldsymbol{q}(\boldsymbol{\mu})$ defined from the RUM or the MDM can span the whole simplex. Theorem 4.2.4 extends these results to more general conditions.

4.3 Relation to the Random Utility Model

In the last section, we proposed the welfare-based choice model. In particular, by Corollary 4.2.3, 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 4.3.1 *For any differentiable choice welfare function $w(\mu_1, \mu_2)$, there exists a distribution θ of $\{\epsilon_1, \epsilon_2\}$ such that:*

$$w(\mu_1, \mu_2) = \mathbb{E}_\theta[\max\{\mu_1 + \epsilon_1, \mu_2 + \epsilon_2\}]. \tag{4.4}$$

In addition, if $w(\mu_1, \mu_2)$ 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 (4.4).

By Proposition 2.2.2, 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

4.5.4, 4.5.6 and 4.5.7 in Section 4.5), 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 exactly the 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, as we shall proceed next. The following result is a direct consequence of the result in McFadden (1980) (as well as Williams 1977 and Daly and Zachary 1978):

Proposition 4.3.2 *Let $w(\boldsymbol{\mu}) : \mathbb{R}^n \mapsto \mathbb{R}$ be a differentiable function. Then $\nabla w(\boldsymbol{\mu})$ 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,*

$$(-1)^k \frac{\partial^k w(\boldsymbol{\mu})}{\partial \mu_{i_1} \cdots \partial \mu_{i_k}} \leq 0.$$

Proposition 4.3.2 is also known as the Williams-Daly-Zachary theorem in the literature. By Proposition 4.3.2 and the above discussions, 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 4.2.2) lies in the requirement on the higher-order derivatives of $w(\cdot)$. In particular, a random utility model requires that the higher-order cross-partial derivatives of $w(\cdot)$ have alternating signs, while the welfare-based choice model only requires that the Hessian matrix of $w(\cdot)$ be positive semidefinite, and there is no requirement on other higher-order derivatives. This characteristic helps us better understand the difference between those models and later construct choice models with new properties.

Remark. Note that when $n = 2$, for any welfare-based choice model with choice welfare function $w(\cdot)$, we have $\frac{\partial w(\boldsymbol{\mu})}{\partial \mu_1} + \frac{\partial w(\boldsymbol{\mu})}{\partial \mu_2} = q_1(\boldsymbol{\mu}) + q_2(\boldsymbol{\mu}) = 1$. By taking another derivative with respect to μ_1, we have $\frac{\partial^2 w(\boldsymbol{\mu})}{\partial \mu_1^2} + \frac{\partial^2 w(\boldsymbol{\mu})}{\partial \mu_1 \partial \mu_2} = 0$. Since $\frac{\partial^2 w(\boldsymbol{\mu})}{\partial \mu_1^2} \geq 0$ by the convexity of $w(\cdot)$, we get $\frac{\partial^2 w(\boldsymbol{\mu})}{\partial \mu_1 \partial \mu_2} \leq 0$ when $n = 2$. Thus by Proposition 4.3.2, any welfare-based choice model is also a random utility model when $n = 2$, which is consistent with Theorem 4.3.1.

4.4 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 the practical meaning of such properties. We introduce two concepts, which we call the *substitutability* and *complementarity* of choices. We show that if a choice model is derived from a random utility model, then the alternatives can only exhibit substitutability. However, the welfare-based choice model allows for more flexible substitutability or complementarity patterns. We also show how this property can be reflected through the choice welfare function in a welfare-based choice model or through the regularization term in a representative agent model. Before we formally introduce these two notions, we first introduce the definition of local monotonicity:

Definition 10 (local monotonicity) *A function $f(x) : \mathbb{R} \mapsto \mathbb{R}$ is locally increasing at x if there exists $\delta > 0$ such that*

$$f(x-h) \leq f(x) \leq f(x+h),\ \forall\ 0 < h < \delta.$$

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

$$f(x-h) \geq f(x) \geq f(x+h),\ \forall\ 0 < h < \delta.$$

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

Definition 11 *Consider a choice model $\boldsymbol{q}(\boldsymbol{\mu}) : \mathbb{R}^n \mapsto \Delta_{n-1}$. For any fixed $\boldsymbol{\mu}$ and $i, j \in \mathcal{N}$:*

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

2. *(Complementarity) If $q_j(\boldsymbol{\mu})$ is locally increasing in μ_i at $\boldsymbol{\mu}$, then we say alternative i is complementary to alternative j at $\boldsymbol{\mu}$. Furthermore, if $q_j(\boldsymbol{\mu})$ is locally increasing*

in μ_i for all $\boldsymbol{\mu}$, then we say alternative i is complementary to alternative j.

3. *(Substitutable and Non-Substitutable Choice Model) If alternative i is substitutable to alternative j for all $i \neq j$, then we say $\boldsymbol{q}(\boldsymbol{\mu})$ is a substitutable choice model. Otherwise, we say $\boldsymbol{q}(\boldsymbol{\mu})$ is a non-substitutable choice model.*

We note that the complementarity property is closely related to the halo effect, which is first conceptualized in Thorndike (1920). The halo effect is a cognitive bias in which an observer's overall impression of a person, a company, a brand or a product influences the observer's feelings and thoughts about that entity's character or properties (see McShane and Von Glinow 2015). For a comprehensive review and discussion about the halo effect, we refer the readers to Rosenzweig (2014). In the context of consumer theory and marketing, the halo effect is the phenomenon that the choice probabilities of certain existing products increase after a new product (usually of the same brand) is introduced.[1] In a choice model that maps a vector of utilities to a vector of choice probabilities, introducing a new product can be viewed as increasing the utility of that product from $-\infty$ to some finite value. Therefore, the notion of complementarity defined in Definition 11 provides an alternative characterization of the halo effect in the context of choice modeling. We believe that this notion could provide new insights on such phenomena.

In the following, we investigate some basic facts about substitutability and complementarity.

Proposition 4.4.1 *Consider a choice model $\boldsymbol{q}(\boldsymbol{\mu}) : \mathbb{R}^n \mapsto \Delta_{n-1}$ that is derived from a differentiable choice welfare function $w(\boldsymbol{\mu})$. For any i, alternative i must be complementary to itself. Furthermore, if $w(\boldsymbol{\mu})$ is second-order continuously differentiable and alternative i is substitutable (complementary, resp.) to alternative j at $\boldsymbol{\mu}$, then alternative j must be substitutable (complementary, resp.) to alternative i at $\boldsymbol{\mu}$.*

Proposition 4.4.1 implies that when $w(\boldsymbol{\mu})$ 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 subsequent discussions.

[1] Such phenomenon is also called the synergistic effect, see Davis et al. (2014).

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

Theorem 4.4.2 *Any random utility model $\boldsymbol{q}(\boldsymbol{\mu})$ is a substitutable choice model.*

Theorem 4.4.2 directly follows from Proposition 4.3.2. 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 plausible intuitively, especially if $\boldsymbol{\mu}$ 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 interpretations of the utility $\boldsymbol{\mu}$.

Example 4.4.3 *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 review scores for 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 $\boldsymbol{q}$ is a function of $\boldsymbol{v} = (v_1, v_2, v_3)$. 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 question 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 (this scenario can be viewed as a case of the halo effect). □

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 probability of choosing the individual product given that the nest is chosen (which will be lower). However, we note that the nested logit model with dissimilarity parameters within $(0, 1]$ is essentially a random utility model (with the randomness $\boldsymbol{\epsilon}$ chosen to be an extreme value distribution, see, e.g., Anderson et al. 1992). Therefore, it is impossible to capture complementarity between alternatives through such a nested logit model. In Section 6, we show that complementarity of alternatives can be captured through a general nested logit model, in which the dissimilarity parameters are allowed to be greater than one. We also show other ways to construct choice models with complementarity property through our welfare-based approach.

Before we end this section, we study conditions for a choice model to be substitutable or non-substitutable. In the following discussion, we only consider choice models $\boldsymbol{q}(\boldsymbol{\mu})$ that are derived from differentiable choice welfare functions $w(\boldsymbol{\mu})$. We provide necessary (sufficient, resp.) conditions for a choice model to be substitutable, and consequently also obtain sufficient (necessary, resp.) conditions for a choice model to be non-substitutable. We first review the concepts of supermodularity and submodularity:

Definition 12 (Supermodularity and Submodularity) *A function $f : \mathbb{R}^n \mapsto \mathbb{R} \cup \{\infty\}$ is called supermodular if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, $f(\boldsymbol{x} \vee \boldsymbol{y}) + f(\boldsymbol{x} \wedge \boldsymbol{y}) \geq f(\boldsymbol{x}) + f(\boldsymbol{y})$, where $\boldsymbol{x} \vee \boldsymbol{y}$ and $\boldsymbol{x} \wedge \boldsymbol{y}$ denote the componentwise maximum and minimum of $\boldsymbol{x}$ and $\boldsymbol{y}$, 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 4.4.4 *Consider a choice model $\boldsymbol{q}(\boldsymbol{\mu}) : \mathbb{R}^n \mapsto \Delta_{n-1}$ that is derived from a differentiable choice welfare function $w(\boldsymbol{\mu})$. Then*

1. *$\boldsymbol{q}(\boldsymbol{\mu})$ is a substitutable choice model if and only if $w(\boldsymbol{\mu})$ is submodular.*
2. *If $\boldsymbol{q}(\boldsymbol{\mu})$ 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*

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

where

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

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

Theorem 4.4.4 provides some sufficient and necessary conditions for $\boldsymbol{q}(\boldsymbol{\mu})$ to be substitutable. We note that the supermodularity of $\bar{V}_i(\cdot)$ has nothing to do with the supermodularity of $V(\cdot)$. In fact, since $V(\boldsymbol{x})$ is only defined on Δ_{n-1}, it can always be extended to a supermodular function in $\mathbb{R}^n$ by defining $V(\boldsymbol{x}) = +\infty$ for all $\boldsymbol{x} \notin \Delta_{n-1}$. The definition of $\bar{V}_i(\cdot)$ reduces a redundant variable in $V(\cdot)$, making the operations "$\boldsymbol{x} \vee \boldsymbol{y}$" and "$\boldsymbol{x} \wedge \boldsymbol{y}$" meaningful.

Next we provide an easy-to-check sufficient condition for a choice model to be substitutable. We note that in the MDM and the MMM introduced in Propositions 2.3.1 and 2.3.2, the corresponding $V(\cdot)$s are separable. The following theorem shows that choice models derived from such $V(\cdot)$s are always substitutable:

Theorem 4.4.5 *If* $V(\boldsymbol{x}) = \sum_{i \in \mathcal{N}} V_i(x_i)$ *on* Δ_{n-1} *where* $V_i(x_i) : [0, 1] \mapsto \mathbb{R}$ *is a strictly convex function for all* $i \in \mathcal{N}$. *Then* $\boldsymbol{q}(\boldsymbol{\mu})$ *defined by*

$$\boldsymbol{q}(\boldsymbol{\mu}) = \arg\max \left\{ \boldsymbol{\mu}^T\boldsymbol{x} - V(\boldsymbol{x}) \middle| \boldsymbol{x} \in \Delta_{n-1} \right\} \tag{4.5}$$

is a substitutable choice model.

Now we have obtained some general conditions for a choice model to be substitutable/non-substitutable. In the next section, we will discuss several concrete examples of non-substitutable choice models as well as ways to construct them. As we will see, such choice models have the potential to capture more flexible choice patterns, thus may be of interest in practice.

4.5 Examples and Constructions of Non-Substitutable Choice Models

In this section, we show different ways of constructing non-substitutable choice models. Some specific models are presented as examples to show how they could be used to explain certain choice scenarios in practice.

4.5.1 General Nested Logit model

The nested logit model, first proposed in Ben-Akiva (1973), is perhaps the most widely used choice model other than the MNL model. In this model, it is assumed that the set of alternatives is partitioned into K subsets (*nests*) denoted by $B_1, B_2, ..., B_K$. The probability of choosing alternative i, given that $i \in B_k$ is

$$q_i^{\mathrm{nl}}(\boldsymbol{\mu}) = \frac{\exp(\mu_i/\lambda_k)\left(\sum_{j\in B_k}\exp(\mu_j/\lambda_k)\right)^{\lambda_k-1}}{\sum_{l=1}^{K}\left(\sum_{j\in B_l}\exp(\mu_j/\lambda_l)\right)^{\lambda_l}}, \tag{4.6}$$

where λ_k is called the *dissimilarity parameter* for the k-th nest. Train et al. (1987) interpret the dissimilarity parameters as a measure of substitutability among alternatives: if $\lambda_k \in (0,1)$, then the substitution is greater within nests than across nests, while if $\lambda_k > 1$, then the substitution is greater across nests than within nests. In addition, in the nested logit model, the independent of irrelevant alternatives (IIA) property holds within each nest, but not across nests. The following result by McFadden (1977) states that λ_k must lie in the unit interval in order for the nested logit model to be consistent with the RUM.

Lemma 4.5.1 (McFadden 1977) *The nested logit model $\boldsymbol{q}^{\mathrm{nl}}(\boldsymbol{\mu})$ is consistent with the RUM for all $\boldsymbol{\mu} \in \mathbb{R}^n$ if and only if $\lambda_k \in (0,1]$ for all $k = 1, ..., K$.*

There have been many studies on the case where λ_k is greater than one for some k. Most of those studies focus on how to relate this case with the RUM. For example, Börsch-Supan (1990) show that $\boldsymbol{q}^{\mathrm{nl}}(\boldsymbol{\mu})$ is consistent with the RUM for certain ranges of $\boldsymbol{\mu}$. Kling and Herriges (1995) and Herriges and Kling (1996) provide tests for consistency of the nested logit model with utility maximization. Train et al. (1987), Lee (1999),

Tiwari and Hasegawa (2004) and Yates and Mackay (2006) fit the nested logit model to consumer data and show that λ_k could be indeed greater than one in real data set.

When $\lambda_k > 1$ for some k, the nested logit model can possess some interesting properties. In particular, when a new product is introduced to a nest k with $\lambda_k > 1$, the probability of choosing certain existing products in that nest may increase in some circumstances, thus certain pairs of products may exhibit complementarity relationship (see Davis et al. 2014). In fact, as the next proposition shows, complementarity property exists in any nested logit model with certain dissimilarity parameter greater than one.

Proposition 4.5.2 *Consider a nested logit model with at least two nests. For any nest k with dissimilarity parameter $\lambda_k > 1$ and any two distinct alternatives i and j in that nest, there always exists $\boldsymbol{\mu} \in \mathbb{R}^n$ such that $\frac{\partial q_i^{\mathrm{nl}}(\boldsymbol{\mu})}{\partial \mu_j} > 0$.*

One implication of Proposition 4.5.2 is that it provides an alternative proof that the nested logit model is not consistent with the RUM if some dissimilarity parameters exceed one. Nevertheless, we show next that the nested logit model is always consistent with the welfare-based choice model. We have the following result:

Proposition 4.5.3 *The nested logit model with $\lambda_k > 0$ for all $k = 1, ..., K$ is a welfare-based choice model with the choice welfare function defined as*

$$w^{nl}(\boldsymbol{\mu}) = \log\left(\sum_{k=1}^{K}\left(\sum_{i\in B_k}\exp(\mu_i/\lambda_k)\right)^{\lambda_k}\right).$$

In addition, it can be expressed as a representative agent model with

$$V^{nl}(\boldsymbol{x}) = \sum_{k=1}^{K}\left((1-\lambda_k)\sum_{i\in B_k}x_i\log\left(\sum_{j\in B_k}x_j\right) + \lambda_k\sum_{i\in B_k}x_i\log x_i\right).$$

We note that the function $V^{nl}(\boldsymbol{x})$ appears in Verboven (1996) when the author studies the nested logit model with dissimilarity parameters less than one. However, the discussion does not cover whether the convexity of $V^{nl}(\boldsymbol{x})$ still holds when some dissimilarity parameters exceed one. Therefore, it was not clear whether $V^{nl}(\boldsymbol{x})$ continues to hold for the representative agent form for that case. Proposition 4.5.3 gives an

affirmative answer to this question, thus providing a complete characterization of the nested logit model for all positive dissimilarity parameters.

Before we end this section, we show an example of a nested logit choice model with some dissimilarity parameters greater than one. In particular, we illustrate the complementarity property in such a model as well as the different representations based on Proposition 4.5.3.

Example 4.5.4 *We consider a nested logit model with three products and two nests. The first nest, with dissimilarity parameter* $\lambda_1 = 2$*, consists of products* 1 *and* 2*; while the second nest, with dissimilarity parameter* $\lambda_2 = 1$*, only consists of product* 3*. From Proposition 4.5.3, the choice welfare function for this model is* $w^{nl}(\boldsymbol{\mu}) = \log\Big((\exp(\mu_1/2) + \exp(\mu_2/2))^2 + \exp(\mu_3)\Big)$*, and the corresponding regularization term in the representative agent model is* $V^{nl}(\boldsymbol{x}) = -(x_1 + x_2)\log(x_1 + x_2) + 2(x_1 \log x_1 + x_2 \log x_2) + x_3 \log x_3$*. The choice probability of product* 1 *is given by:*

$$q_1(\boldsymbol{\mu}) = \frac{\exp(\mu_1/2)(\exp(\mu_1/2) + \exp(\mu_2/2))}{(\exp(\mu_1/2) + \exp(\mu_2/2))^2 + \exp(\mu_3)}.$$

We note that $\frac{\partial q_1(\boldsymbol{\mu})}{\partial \mu_2} > 0$ *is equivalent to* $\exp(\mu_1/2) + \exp(\mu_2/2) < \exp(\mu_3/2)$*. In Figure 4.1, fixing* $\mu_3 = 3$*, we show the regions in which* $\frac{\partial q_1(\boldsymbol{\mu})}{\partial \mu_2} > 0$ *(products 1 and 2 are complementary) or* $\frac{\partial q_1(\boldsymbol{\mu})}{\partial \mu_2} < 0$ *(products 1 and 2 are substitutable). Now we relate this model to Example 4.4.3 and provide some explanations. Consider products 1 and 2 in this model to be the Canon-A and Canon-B models in Example 4.4.3 respectively, and product 3 to be the Sony-C model. Then this model will capture a scenario in which when the reviews of the Canon brand are generally low (*$\exp(\mu_1/2) + \exp(\mu_2/2) < \exp(\mu_3/2)$*), increasing the review score of one Canon model will have significant impact on its brand image, thus will increase the demand of the other Canon model. However, when some Canon models have already received high review score (*$\exp(\mu_1/2) + \exp(\mu_2/2) > \exp(\mu_3/2)$*), this brand effect dwindles, and the two products become substitutable to each other. We believe this model provides some plausible explanation to the situation described in Example 4.4.3.* □

In addition, Example 4.5.4 shows that even when $n = 3$, there exist welfare-based choice models that do not have a random utility representation (remember all random

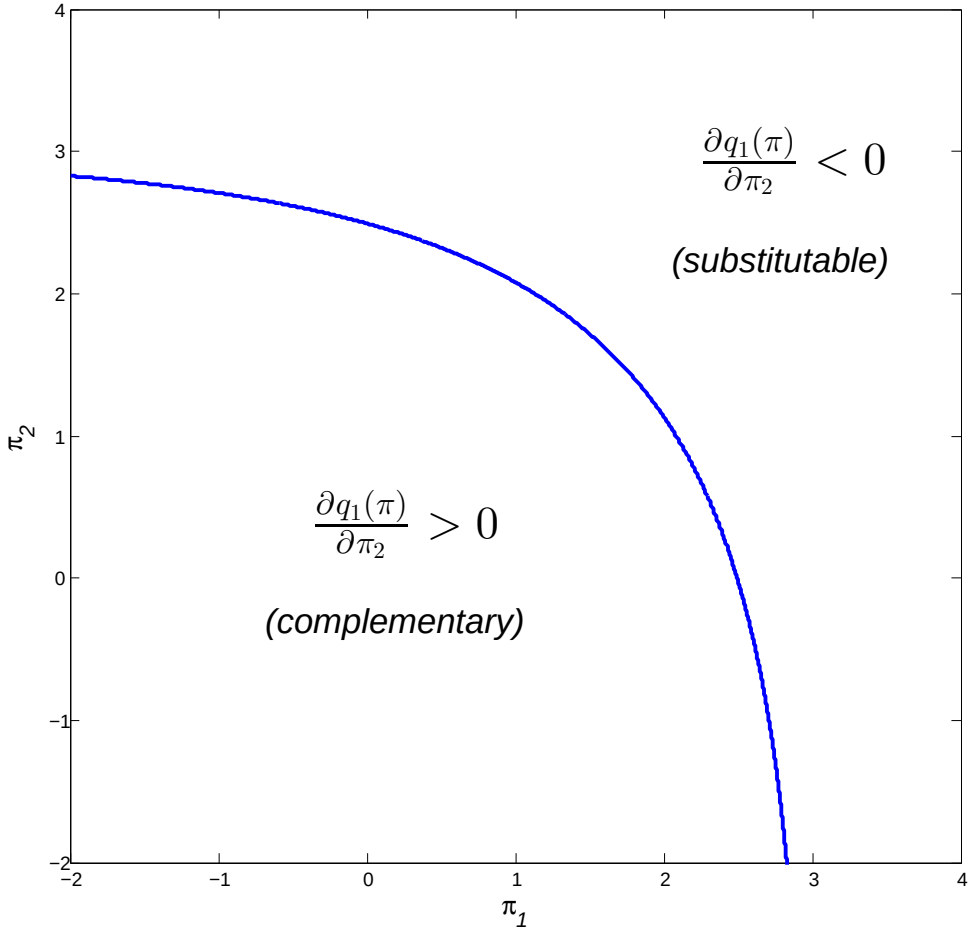

Figure 4.1: Substitutability/Complementarity for Different Values of (μ_1, μ_2) in Example 4.5.4 ($\mu_3 = 3$)

utility models are substitutable choice models). Therefore, the welfare-based choice model (thus also the representative agent model and the semi-parametric choice model) 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 $n \geq 4$.

Finally we note that the results on the one level nested logit model could be extended to the case of d-level nested logit model. We omit the formal discussion here and leave it for future research.

4.5.2 Quadratic Regularization

Another way to generate non-substitutable choice models is to start from the representative agent model and to choose $V(\cdot)$ as a convex quadratic function. Remember that $V(\cdot)$ has to be a convex function in the representative agent model. Thus, a convex quadratic function could be used as an approximation. We have the following proposition about the substitutability and complementarity in such models.

Proposition 4.5.5 *Consider a choice model* $\boldsymbol{q}(\boldsymbol{\mu}) = \arg\max\left\{\boldsymbol{\mu}^T\boldsymbol{x} - V(\boldsymbol{x})\middle|\boldsymbol{x} \in \Delta_{n-1}\right\}$ *with* $V(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ *strictly convex with* $\boldsymbol{A} \succ \boldsymbol{0}$*. If the choice model is substitutable, then* $A_{jk} - A_{ik} - A_{ij} + A_{ii} \geq 0$ *for all distinct* $i, j, k \in \mathcal{N}$*, where* A_{ij} *is the* (i, j)*-th entry of* $\boldsymbol{A}$*. Furthermore, the reverse is true if* $n = 3$*.*

By Proposition 4.5.5, we know that when $n = 3$ and $V(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ with $\boldsymbol{A} \succ \boldsymbol{0}$, the choice model defined by $\boldsymbol{q}(\boldsymbol{\mu}) = \arg\max\left\{\boldsymbol{\mu}^T\boldsymbol{x} - V(\boldsymbol{x})\middle|\boldsymbol{x} \in \Delta_{n-1}\right\}$ is substitutable if and only if

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

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

Example 4.5.6 *Consider* $\boldsymbol{q}(\boldsymbol{\mu}) = \arg\max\left\{\boldsymbol{\mu}^T\boldsymbol{x} - V(\boldsymbol{x})\middle|\boldsymbol{x} \in \Delta_{n-1}\right\}$*, where* $V(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ *with*

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

It is easy to see that $V(\boldsymbol{x})$ *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*

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

which is not supermodular.

Therefore $\boldsymbol{q}(\boldsymbol{\mu})$ is not a substitutable choice model by Theorem 4.4.4. In fact, when we fix $\mu_2 = \mu_3 = 0$ and plot the choice probabilities against μ_1 in the range of values $[-2, 2]$ as shown in Figure 4.2, it is observed that q_3 increases in μ_1 in the range of $[-1.5, -1]$, i.e., alternative 3 is complementary to alternative 1 in that interval. □

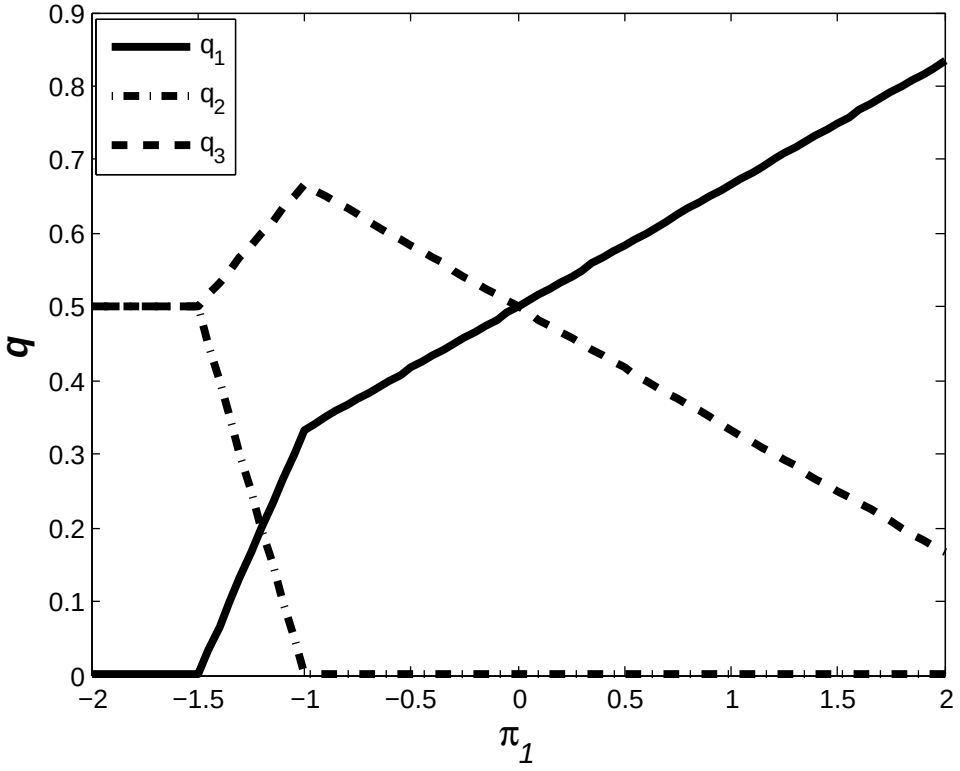

Figure 4.2: Choice Probabilities in Example 4.5.6 with $\mu_2 = \mu_3 = 0$

4.5.3 Crossing Transformation

Next, we provide a systematic way to generate non-substitutable choice models from existing substitutable choice models by using the welfare-based approach. Let $\boldsymbol{A}$ be an $m \times n$ matrix with $A_{ij} \geq 0$ and $\boldsymbol{A}\boldsymbol{e}_n = \boldsymbol{e}_m$, where $\boldsymbol{e}_\ell$ refers to an ℓ-dimensional column vector of ones. Given an existing choice welfare function $\bar{w}(\cdot) : \mathbb{R}^m \mapsto \bar{\mathbb{R}}$ and its choice probabilities $\bar{\boldsymbol{q}}(\cdot)$, we can easily verify that

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

is still a choice welfare function that maps $\mathbb{R}^n$ to $\bar{\mathbb{R}}$ and the corresponding welfare-based choice model is

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

By some calculation, we have:

$$\nabla^2 w(\boldsymbol{\mu}) = \boldsymbol{A}^T \nabla^2 \bar{w}(\boldsymbol{A}\boldsymbol{\mu})\boldsymbol{A}.$$

Even if $\bar{w}(\boldsymbol{\mu})$ is submodular, i.e., the off-diagonal entries of $\nabla^2 \bar{w}(\boldsymbol{\mu})$ are all negative, it is still possible to construct matrix $\boldsymbol{A}$ such that $\boldsymbol{A}^T \nabla^2 \bar{w}(\boldsymbol{A}\boldsymbol{\mu})\boldsymbol{A}$ has positive off-diagonal entries. Therefore, by choosing some proper matrix $\boldsymbol{A}$, we can construct non-substitutable choice model $w(\boldsymbol{\mu})$ from substitutable choice model $\bar{w}(\boldsymbol{\mu})$. We call this method the *crossing transformation* and the corresponding matrix $\boldsymbol{A}$ the crossing matrix.

In the following, we give an example of constructing a non-substitutable choice model from the MNL model using the crossing transformation.

Example 4.5.7 *Let* $\bar{w}(\boldsymbol{x}) = \log\left(e^{x_1} + e^{x_2} + e^{x_3} + e^{x_4}\right)$ *be the choice welfare function for an MNL model for* 4 *alternatives. Let the crossing matrix be*

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

Then the new welfare-based choice model becomes:

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

It is easy to see that $w(\boldsymbol{\mu})$ *is monotone, translation invariant and convex. Therefore it is a choice welfare function. Also, it is differentiable with the corresponding choice*

probability:

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

Furthermore, the second-order derivative of $w(\boldsymbol{\mu})$ *with respect to* μ_1 *and* μ_2 *is*

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

It is positive if and only if $e^{\mu_3} \geq 4e^{0.5\mu_1+0.5\mu_2} + e^{\mu_1} + e^{\mu_2}$. *When* $\mu_3 = 3$, *this inequality is satisfied in the region below the curve in Figure 4.3. Therefore, under this choice model, when both* μ_1 *and* μ_2 *are small enough (compared to* μ_3*), alternatives* 1 *and* 2 *are complementary. Otherwise, they are substitutable. Similar to the discussions in the Example 4.5.4, this model could also provide an explanation to Example 4.4.3, in which the substitutability and complementarity between products 1 and 2 depend on the current review level of the brand. However, this model is different from that in Example 4.5.4 (they have a different choice welfare function).*[2]

Finally, we note that this example shows that the RUM is not closed under the crossing transformation, i.e., even if $\bar{\boldsymbol{q}}(\cdot)$ has an RUM representation, $\boldsymbol{q}(\boldsymbol{\mu})$ may not. Thus, the crossing transformation also provides a way of generating choice models outside the random utility family.

4.6 Conclusion

In this chapter, we proposed a welfare-based approach to study discrete choice models. 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

[2]In this case, it is possible to construct a crossing matrix $\boldsymbol{A}$ to make the resulting choice model equivalent to that in Example 4.5.4, which is a nested logit model. However, we note that this is because the dissimilarity parameters in Example 4.5.4 are all integers. In general, it is not always possible to construct a nested logit model from an MNL model by only applying the crossing transformation.

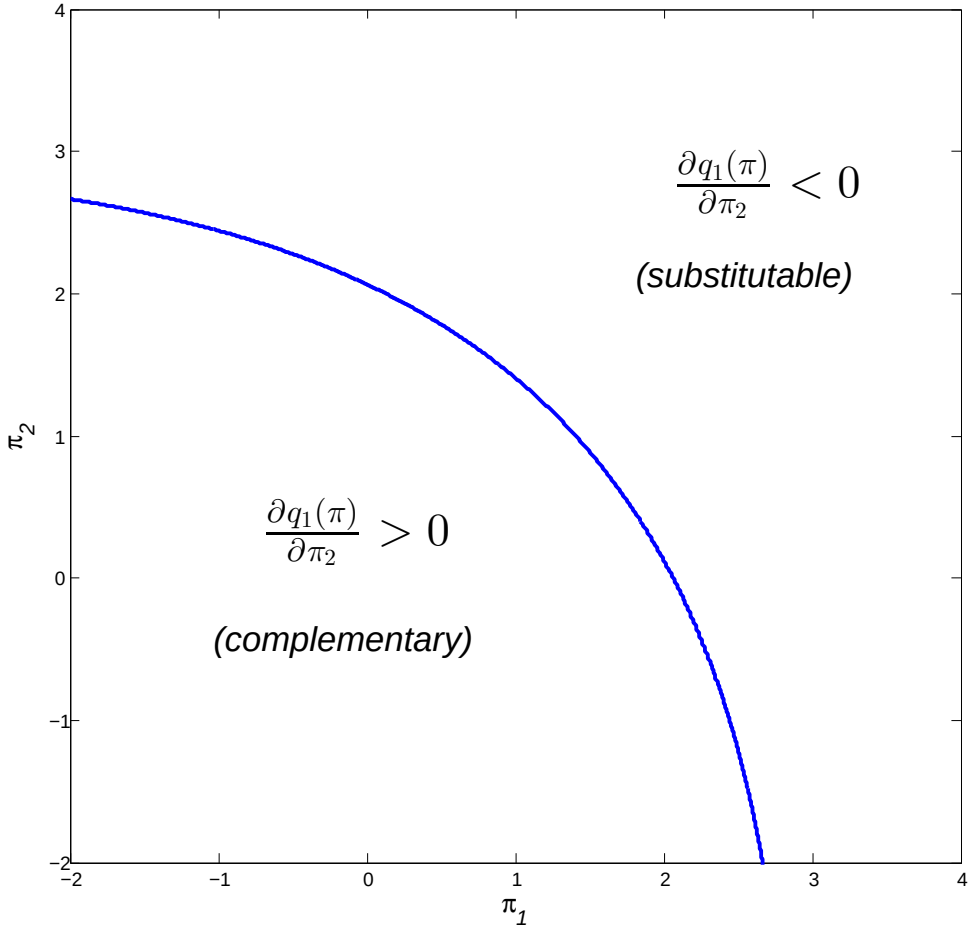

Figure 4.3: Substitutability/Complementarity for Different Values of (μ_1, μ_2) in Example 4.5.7 ($\mu_3 = 3$)

only two alternatives, the welfare-based choice model is equivalent to the random utility model. Furthermore, we introduced a new notion for choice models – substitutability and complementarity – and showed that with the help of the welfare-based choice model, we can construct choice models with complementary alternatives, thus enabling us to capture more flexible choice patterns. We believe that this approach provides useful insights for the study of choice models.

From an application perspective, we notice that the parameter estimation for discrete choice models is a very important issue, attracting much attention in the literature. In practice, with given data, it is likely that one will first choose a certain parametric model

and then the parameter estimation problem will become a regression problem. In our current work, the main focus is on the study of the theoretical relation between choice models and the inherent choice patterns. We leave the parameter estimation problem using the available data to be a future research topic.

4.7 Technical Proofs

Proof of Theorem 4.2.1: First we show that the $w(\boldsymbol{\mu})$ defined in (4.1) and (4.2) are choice welfare functions. To see this, we note that the monotonicity and translation invariance properties are immediate from (4.1) and (4.2). For the convexity, we note that $w(\boldsymbol{\mu})$ defined in (4.1) is the supremum of linear functions of $\boldsymbol{\mu}$ thus is convex in $\boldsymbol{\mu}$. In (4.2), for each $\boldsymbol{\epsilon}$, $\max_{i\in\mathcal{N}}\{\mu_i+\epsilon_i\}$ is a convex function in $\boldsymbol{\mu}$, and so is the expectation. Therefore, if $w(\boldsymbol{\mu})$ is defined by (4.1) or (4.2), then it must be a choice welfare function.

Next we show the other direction. That is, if $w(\boldsymbol{\mu})$ is a choice welfare function, then it can be represented in the form of (4.1) and (4.2). First we note that if a choice welfare function $w(\boldsymbol{\mu}) = +\infty$ for some $\boldsymbol{\mu}$, then for any $\boldsymbol{\mu}'$, we have $w(\boldsymbol{\mu}') \geq w(\boldsymbol{\mu}+\min_i(\mu_i'-\mu_i)\mathbf{e}) = w(\boldsymbol{\mu})+\min_i(\mu_i'-\mu_i) = +\infty$, where the first inequality uses the monotonicity property and the first equality uses the translation invariance property. Thus $w(\boldsymbol{\mu}) = +\infty$ for all $\boldsymbol{\mu}$. In that case, we can choose $V(\boldsymbol{x}) = -\infty$ and $\Theta = \{\theta_\infty\}$ where θ_∞ is a singleton distribution taking value on $(\infty, ..., \infty)$. Therefore, $w(\boldsymbol{\mu})$ can be represented by (4.1) and (4.2) in that case. Similarly, if $w(\boldsymbol{\mu}) = -\infty$ for some $\boldsymbol{\mu}$, then it must be that $w(\boldsymbol{\mu}) = -\infty$ for all $\boldsymbol{\mu}$, and we can take $V(\boldsymbol{x}) = \infty$ and $\Theta = \{\theta_{-\infty}\}$, where $\theta_{-\infty}$ is a singleton distribution on $(-\infty, ..., -\infty)$. Therefore, $w(\boldsymbol{\mu})$ can be represented in (4.1) and (4.2) in this case too.

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

1. We show that any choice welfare function $w(\boldsymbol{\mu})$ can be represented by (4.1). Since $w(\boldsymbol{\mu})$ is monotone and translation invariant, the following holds:

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

Here the first equality holds since for any $\boldsymbol{y}$,

$$w(\boldsymbol{\mu}) = w(\boldsymbol{\mu} - \max_i \{\mu_i - y_i\}\, \mathbf{e}) + \max_i \{\mu_i - y_i\}$$

by the translation invariance property. Furthermore, by the monotonicity property,

$$w(\boldsymbol{\mu} - \max_i \{\mu_i - y_i\}\, \mathbf{e}) \leq w(\boldsymbol{y})$$

and the equality holds when $\boldsymbol{y} = \boldsymbol{\mu}$.

Next we define $L(\boldsymbol{x}, \boldsymbol{y}) = w(\boldsymbol{y}) + (\boldsymbol{\mu} - \boldsymbol{y})^T \boldsymbol{x}$. We have for fixed $\boldsymbol{x}$, $L(\boldsymbol{x}, \cdot)$ is convex in $\boldsymbol{y}$ (by the convexity of $w(\cdot)$); and for fixed $\boldsymbol{y}$, $L(\cdot, \boldsymbol{y})$ is convex and closed in $\boldsymbol{x}$. Furthermore,

$$\inf_{\boldsymbol{y}} \max_{\boldsymbol{x} \in \Delta_{n-1}} L(\boldsymbol{x}, \boldsymbol{y}) = w(\boldsymbol{\mu}) < \infty$$

and the function $q(\boldsymbol{u}) = \inf_{\boldsymbol{y}} \max_{\boldsymbol{x} \in \Delta_{n-1}} \left\{ L(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{u}^T \boldsymbol{x} \right\} = w(\boldsymbol{\mu} - \boldsymbol{u})$ is continuous. Therefore, by Proposition 2.6.2 of Bertsekas (2003), the minimax equality holds, i.e.,

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

Therefore, we have:

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

where

$$V(\boldsymbol{x}) = \sup_{\boldsymbol{y}} \{ \boldsymbol{y}^T \boldsymbol{x} - w(\boldsymbol{y}) \}$$

is a convex function.

2. Next we show that any choice welfare function can be represented by (4.2). Since $w(\boldsymbol{\mu})$ is convex, there exists a subgradient for any $\boldsymbol{\mu}$. We denote the subgradient vector by $\boldsymbol{d}(\boldsymbol{\mu}) = (d_1(\boldsymbol{\mu}), \ldots, d_n(\boldsymbol{\mu}))^T$. Here it is possible that the choice of $\boldsymbol{d}(\boldsymbol{\mu})$ is not unique, when we can choose any one of them. 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 $\boldsymbol{d}(\boldsymbol{\mu})$, it must hold that $\mathbf{e}^T \boldsymbol{d}(\boldsymbol{\mu}) = 1$. Similarly, by the monotonicity property of $w(\boldsymbol{\mu})$, we must have

$\boldsymbol{d}(\boldsymbol{\mu}) \geq \mathbf{0}$. By the definition of subgradient and the convexity of $w(\boldsymbol{\mu})$, we must have:

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

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

$$w(\boldsymbol{\mu}) = \sup_{\boldsymbol{z}} \{\boldsymbol{\mu}^T \boldsymbol{d}(\boldsymbol{z}) + l(\boldsymbol{z})\}. \tag{4.7}$$

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

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

where

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

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

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

Then by (4.7), we have

$$w(\boldsymbol{\mu}) = \sup_{\boldsymbol{z}} \ \{\boldsymbol{\mu}^T \boldsymbol{d}(\boldsymbol{z}) + l(\boldsymbol{z})\} = \sup_{\boldsymbol{z}} \ \mathbb{E}_{\theta_{\boldsymbol{z}}}[\max_i \ \mu_i + \epsilon_i] = \sup_{\theta \in \Theta} \ \mathbb{E}_{\theta}[\max_i \ \mu_i + \epsilon_i].$$

Therefore, the theorem is proved. □

Proof of Theorem 4.2.2: The equivalence between 1 and 3 directly follows from Theorem 1. Next we show that 1 $\Rightarrow$ 2. If $w(\boldsymbol{\mu})$ is a differentiable choice welfare function, by Theorem 4.2.1, we know that

$$w(\boldsymbol{\mu}) = \max \left\{\boldsymbol{\mu}^T \boldsymbol{x} - V(\boldsymbol{x}) \big| \boldsymbol{x} \in \Delta_{n-1}\right\},$$

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

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

where $\boldsymbol{x}^* = \operatorname{argmax}\left\{\boldsymbol{\mu}^T\boldsymbol{x} - V(\boldsymbol{x})\big|\boldsymbol{x} \in \Delta_{n-1}\right\}$. Therefore,

$$\boldsymbol{q}(\boldsymbol{\mu}) = \nabla w(\boldsymbol{\mu}) = \operatorname{argmax}\left\{\boldsymbol{\mu}^T\boldsymbol{x} - V(\boldsymbol{x})\big|\boldsymbol{x} \in \Delta_{n-1}\right\}.$$

Last, we show that $2 \Rightarrow 1$. Given an essentially strictly convex $V(\boldsymbol{x})$, by Theorem 4.2.1, we know that

$$w(\boldsymbol{\mu}) = \max\left\{\boldsymbol{\mu}^T\boldsymbol{x} - V(\boldsymbol{x})\big|\boldsymbol{x} \in \Delta_{n-1}\right\}$$

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

Proof of Theorem 4.2.4: First we show the equivalence between 1 and 2. Based on Theorem 4.2.2, it suffices to prove that $w(\boldsymbol{\mu})$ is superlinear if and only if $V(\boldsymbol{x})$ defined by $\max_{\boldsymbol{y}}\{\boldsymbol{y}^T\boldsymbol{x} - w(\boldsymbol{y})\}$ is upper bounded. If $w(\boldsymbol{\mu})$ is superlinear, we have, for any $\boldsymbol{x} \in \Delta_{n-1}$,

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

By reorganizing terms, we have

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

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

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

have

$$w(\boldsymbol{\mu}) \geq \max\left\{\boldsymbol{\mu}^T\boldsymbol{x} - u \middle| \boldsymbol{x} \in \Delta_{n-1}\right\} \geq \mu_i - u, \quad \forall i,$$

i.e., $w(\boldsymbol{\mu})$ 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 $w(\boldsymbol{\mu})$, we can find a distribution set Θ consisting of only distributions with finite expectation such that $w(\boldsymbol{\mu})$ can be represented as $w(\boldsymbol{\mu}) = \sup_{\theta\in\Theta}\ \mathbb{E}_\theta\left[\max_{i\in\mathcal{N}}\ \mu_i + \epsilon_i\right]$.

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

$$w(\boldsymbol{\mu}) = \sup_{\boldsymbol{z}} \quad \{\boldsymbol{\mu}^T\boldsymbol{q}(\boldsymbol{z}) + l(\boldsymbol{z})\}, \tag{4.8}$$

where $l(\boldsymbol{z}) = w(\boldsymbol{z}) - \boldsymbol{z}^T\boldsymbol{q}(\boldsymbol{z})$. Now we define a distribution set Θ that is slightly different from that of Theorem 4.2.1. Specifically, let $\Theta = \{\theta_{\boldsymbol{z}} \big| \boldsymbol{z} \in \mathbb{R}^n\}$, where $\theta_{\boldsymbol{z}}$ is an n-point distribution with $\mathbb{P}_{\theta_{\boldsymbol{z}}}\left(\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^i_{\boldsymbol{z}}\right) = q_i(\boldsymbol{z})$, $\forall i \in \mathcal{N}$(Note that by the monotonicity and the translation invariance properties, $\boldsymbol{q}(\boldsymbol{z}) = \nabla w(\boldsymbol{z})$ must satisfy $\boldsymbol{q}(\boldsymbol{z}) \geq \boldsymbol{0}$ and $\boldsymbol{e}^T\boldsymbol{q}(\boldsymbol{z}) = 1$). Here,

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

where

$$M(\boldsymbol{z}) = \max\left\{1 + \max_{i,j}\ \{z_i - z_j\},\ \frac{l(\boldsymbol{z}) - \min_i\ \{b_i\}}{t^*(\boldsymbol{z})}\right\}, \tag{4.9}$$

with

$$t^*(\boldsymbol{z}) = \min\{q_i(\boldsymbol{z}) | q_i(\boldsymbol{z}) > 0\}. \tag{4.10}$$

Since $M(\boldsymbol{z}) > z_i - z_j$, for all i, j, we have $i = \text{argmax}_j\ (z_j + \epsilon^i_{\boldsymbol{z}}(j))$. Therefore,

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

Next we show that:

$$\mathbb{E}_{\theta_{\boldsymbol{z}}}[\max_i\ \mu_i + \epsilon_i] \leq w(\boldsymbol{\mu}), \quad \forall \boldsymbol{\mu}.$$

For any given $\boldsymbol{\mu}$, define $k(i) \triangleq \text{argmax}_j\ (\mu_j + \epsilon^i_{\boldsymbol{z}}(j))$ (we break ties arbitrarily). There

are two cases:

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

$$\mathbb{E}_{\theta_{\boldsymbol{z}}}[\max_j \ \mu_j + \epsilon_j] = \sum_{i \in \mathcal{N}} q_i(\boldsymbol{z})(\mu_i + l(\boldsymbol{z})) = \boldsymbol{\mu}^T \boldsymbol{q}(\boldsymbol{z}) + l(\boldsymbol{z}) \leq w(\boldsymbol{\mu}),$$

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

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

$$\begin{aligned}
\mathbb{E}_{\theta_{\boldsymbol{z}}}[\max_j \ \mu_j + \epsilon_j] &= \sum_{i \in \mathcal{N}, q_i(\boldsymbol{z}) > 0} q_i(\boldsymbol{z})(\mu_{k(i)} + l(\boldsymbol{z}) - M(\boldsymbol{z}) \mathbb{I}_{\{k(i) \neq i\}}) \\
&\leq \max_i \{\mu_i\} + l(\boldsymbol{z}) - t^*(\boldsymbol{z}) M(\boldsymbol{z}) \\
&\leq \max_i \{\mu_i\} + \min_j \{b_j\} \\
&\leq \max_i \ \{\mu_i + b_i\} \\
&\leq w(\boldsymbol{\mu}),
\end{aligned}$$

   where the first inequality follows from the fact that $M(\boldsymbol{z}) > 0$ and

$$\sum_{i \in \mathcal{N}} q_i(\boldsymbol{z}) \mathbb{I}_{\{q_i(\boldsymbol{z}) > 0, k(i) \neq i\}} \geq t^*(\boldsymbol{z}),$$

   where the second inequality is because of the definition of $M(\boldsymbol{z})$ and the last inequality follows from the definition of superlinear function.

Based on the analysis of these two cases, we have

$$\mathbb{E}_{\boldsymbol{\epsilon} \sim \theta_{\boldsymbol{z}}}[\max_i \ \mu_i + \epsilon_i] \leq w(\boldsymbol{\mu}), \quad \forall \boldsymbol{\mu}.$$

Then by equation (4.8) we have

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

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 $\boldsymbol{\mu} \in \mathbb{R}^n$ we have

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

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

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

$$\boldsymbol{x} \in \text{int}(\Delta_{n-1}) \triangleq \{\boldsymbol{x} \,|\boldsymbol{e}^T\boldsymbol{x} = 1, x_i > 0, \forall i \in \mathcal{N}\},$$

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

$$V(\boldsymbol{x}) = \max_{\boldsymbol{\mu}} \{\boldsymbol{\mu}^T\boldsymbol{x} - w(\boldsymbol{\mu})\}. \tag{4.11}$$

Clearly, $V(\boldsymbol{x}) \geq -w(\mathbf{0})$, since $\boldsymbol{\mu} = \mathbf{0}$ is a feasible solution. Moreover, since $w(\boldsymbol{\mu})$ is translation invariant, we can restrict the feasible region of (4.11) to $\mathcal{L} \triangleq \{\boldsymbol{\mu}|\boldsymbol{e}^T\boldsymbol{\mu} = 0\}$. For all $\boldsymbol{\mu} \in \mathcal{L}$, we have $\mu_j \leq 0$ for some $j \in \mathcal{N}$. Thus

$$\boldsymbol{\mu}^T\boldsymbol{x} \leq \sum_{i\neq j} \mu_i x_i \leq \sum_{i\neq j} x_i \max_k\{\mu_k\} \leq (1 - \min_i\{x_i\}) \max_k\{\mu_k\}.$$

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

$$w(\boldsymbol{\mu}) \geq \max_k\{\mu_k + b_k\} \geq \max_k\{\mu_k\} + \min_k\{b_k\}.$$

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

$$\boldsymbol{\mu}^T\boldsymbol{x} - w(\boldsymbol{\mu}) \leq -\min_i\{x_i\} \max_k\{\mu_k\} - \min_k\{b_k\}.$$

Let $K = \frac{w(\mathbf{0}) - \min_k\{b_k\}}{\min_i\{x_i\}}$. In order for $\boldsymbol{\mu}$ to be optimal to (4.11), by the above arguments, we would have $\mu_i \leq K$ for all i. Thus we can further restrict the feasible set of (4.11) to $\{\boldsymbol{\mu}|\boldsymbol{e}^T\boldsymbol{\mu} = 0, \mu_i \leq K \ \forall i \in \mathcal{N}\}$, which is a compact set. Since $w(\boldsymbol{\mu})$ is continuous, there exists $\boldsymbol{\mu}_{\boldsymbol{x}} \in \{\boldsymbol{\mu}|\boldsymbol{e}^T\boldsymbol{\mu} = 0, \mu_i \leq K \ \forall i \in \mathcal{N}\}$ that attains maximum in problem (4.11). By the first-order necessary condition, $\nabla w(\boldsymbol{\mu}_{\boldsymbol{x}}) = \boldsymbol{x}$. This concludes the proof. □

Proof of Theorem 4.3.1: 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 θ of $\{\epsilon_1,\epsilon_2\}$ as follows:

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

where $v_0 = v(0) = w(0,0)$ and ξ is a random variable with c.d.f. $F_\xi(x) = \mathbb{P}(\xi \le x) = v'(x)$. 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 (see Rockafellar 1974).

Now we compute $\mathbb{E}_\theta[\max\{\mu_1+\epsilon_1, \mu_2+\epsilon_2\}]$. We have

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

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

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

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

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

$$\begin{aligned}
\mathbb{E}_\theta[\max\{0, x+\xi\} - \max\{0,\xi\}] &= x\mathbb{P}(\xi > -x) + \mathbb{E}_\theta[-\xi \cdot \mathbb{I}(0 < \xi \leq -x)] \\
&= x\mathbb{P}(\xi > -x) - \int_0^{-x} \xi dv'(\xi) \\
&= x(1 - v'(-x)) - \xi v'(\xi) \mid_0^{-x} + \int_0^{-x} v'(\xi) d\xi \\
&= x - v_0 + v(-x).
\end{aligned}$$

Therefore, $\mathbb{E}_\theta[\max\{\mu_1+\epsilon_1, \mu_2+\epsilon_2\}] = \mu_1 + v_0 + (\mu_2 - \mu_1) - v_0 + v(\mu_1 - \mu_2) = w(\mu_1, \mu_2)$.

To prove the last statement, it suffices to show that both $\mathbb{E}_\theta[\max\{0,\xi\}]$ and $\mathbb{E}_\theta[\max\{0,-\xi\}]$ are finite if $w(\boldsymbol{\mu})$ 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:

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

and

$$\mathbb{E}_\theta[\max\{0,-\xi\}] = \int_0^{+\infty} \mathbb{P}_\theta\left(-\xi \geq t\right) dt = \int_0^{+\infty} v'(-t) dt = \int_{-\infty}^{0} v'(t) dt = v(0) - L_2.$$

Thus, the theorem is proved. □

Proof of Proposition 4.4.1: Since $w(\boldsymbol{\mu})$ is convex and differentiable, for any $\boldsymbol{\mu} \in \mathbb{R}^n$ and any $t > 0$, we have

$$\begin{aligned}
&w(\boldsymbol{\mu} + t\boldsymbol{e_i}) - w(\boldsymbol{\mu}) \geq t\boldsymbol{e_i^T}\nabla w(\boldsymbol{\mu}) = tq_i(\boldsymbol{\mu}), \\
&w(\boldsymbol{\mu}) - w(\boldsymbol{\mu} + t\boldsymbol{e_i}) \geq -t\boldsymbol{e_i^T}\nabla w(\boldsymbol{\mu} + t\boldsymbol{e_i}) = -tq_i(\boldsymbol{\mu} + t\boldsymbol{e_i}).
\end{aligned}$$

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

Furthermore, if $w(\boldsymbol{\mu})$ is second-order continuously differentiable, then we have

$$\frac{\partial q_i}{\partial \mu_j} = \frac{\partial^2 w}{\partial \mu_i \partial \mu_j} = \frac{\partial^2 w}{\partial \mu_j \partial \mu_i} = \frac{\partial q_j}{\partial \mu_i}.$$

Thus, if alternative i is substitutable (complementary, resp.) to alternative j at $\boldsymbol{\mu}$, then alternative j is substitutable (complementary, resp.) to alternative i at $\boldsymbol{\mu}$. □

Proof of Theorem 4.4.4: In this proof, we use the following lemma from Murota (2003).

Lemma 4.7.1 *Let $f : \mathbb{R}^n \mapsto \mathbb{R} \cup \{\infty\}$ be a function such that there exists at least one $\boldsymbol{\mu}$ such that $f(\boldsymbol{\mu}) < \infty$. Let $g(\boldsymbol{x}) = \max_{\boldsymbol{\mu}}\{\boldsymbol{\mu}^T\boldsymbol{x} - f(\boldsymbol{\mu})\}$ 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 $w(\boldsymbol{\mu})$ is submodular in $\boldsymbol{\mu}$ if and only if $\frac{\partial w(\boldsymbol{\mu})}{\partial \mu_i}$ is decreasing in μ_j for all $i \neq j$. By the definition of $\boldsymbol{q}(\boldsymbol{\mu}) = \nabla w(\boldsymbol{\mu})$, the result holds.

For the second part, let $V(\boldsymbol{x}) = \max_{\boldsymbol{\mu}}\{\boldsymbol{\mu}^T\boldsymbol{x} - w(\boldsymbol{\mu})\}$ be the convex conjugate of $w(\boldsymbol{\mu})$. From Theorem 4.2.2, $V(\boldsymbol{x})$ is essentially strictly convex and

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

For any $\boldsymbol{y} \in \mathbb{R}^{n-1}$ and $i \in \mathcal{N}$, define $f_i(\boldsymbol{y}) = w(y_1, y_2, ..., y_{i-1}, 0, y_i, ..., y_{n-1})$. Also define $\boldsymbol{\mu_{-i}} = (\mu_1, ..., \mu_{i-1}, \mu_{i+1}, ..., \mu_n)$, then we have

$$\begin{aligned}
\bar{V}_i(\boldsymbol{z}) &= \max_{\boldsymbol{\mu}}\{\boldsymbol{\mu}_{-i}^T\boldsymbol{z} + \mu_i(1 - \boldsymbol{e}^T\boldsymbol{z}) - w(\boldsymbol{\mu})\} \\
&= \max_{\boldsymbol{\mu}, \mu_i = 0}\{\boldsymbol{\mu}_{-i}^T\boldsymbol{z} + \mu_i(1 - \boldsymbol{e}^T\boldsymbol{z}) - w(\boldsymbol{\mu})\} \\
&= \max_{\boldsymbol{y}}\{\boldsymbol{y}^T\boldsymbol{z} - f_i(\boldsymbol{y})\},
\end{aligned}$$

where the second equality is due to the translation invariance property of $w(\boldsymbol{\mu})$. The submodularity of $w(\boldsymbol{\mu})$ implies the submodularity of $f_i(\boldsymbol{y})$ for all $i \in \mathcal{N}$. Thus $\bar{V}_i(\boldsymbol{z})$, as

the convex conjugate of $f_i(\boldsymbol{y})$, is supermodular by Lemma 4.7.1.

For the last statement, since $V(\cdot)$ is an essentially strictly convex function,

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

is well-defined. By Theorem 4.2.2, $\boldsymbol{q}(\boldsymbol{\mu}) = \nabla w(\boldsymbol{\mu})$ where

$$w(\boldsymbol{\mu}) = \sup\left\{\boldsymbol{\mu}^T\boldsymbol{x} - V(\boldsymbol{x})\big|\boldsymbol{x} \in \Delta_{n-1}\right\}.$$

For any $\boldsymbol{y} \in \mathbb{R}^{n-1}$ and $i \in \mathcal{N}$, define $f_i(\boldsymbol{y}) = w(y_1, y_2, ..., y_{i-1}, 0, y_i, ..., y_{n-1})$. Also define $\boldsymbol{x_{-i}} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)$, then we have

$$\begin{aligned} f_i(\boldsymbol{y}) &= \max_{\boldsymbol{x}\in\Delta_{n-1}} \left\{\boldsymbol{x}_{-i}^T\boldsymbol{y} + 0(1 - \boldsymbol{e}^T\boldsymbol{x}_{-i}) - V(\boldsymbol{x})\right\} \\ &= \max_{\boldsymbol{x}\in\Delta_{n-1}} \left\{\boldsymbol{x}_{-i}^T\boldsymbol{y} - \bar{V}_i(\boldsymbol{x}_{-i})\right\} \\ &= \max_{\boldsymbol{e}_{n-1}^T\boldsymbol{x}_{-i}\leq 1,\ \boldsymbol{x}_{-i}\geq 0} \left\{\boldsymbol{x}_{-i}^T\boldsymbol{y} - \bar{V}_i(\boldsymbol{x}_{-i})\right\} \\ &= \max_{\boldsymbol{z}} \left\{\boldsymbol{y}^T\boldsymbol{z} - \bar{V}_i(\boldsymbol{z})\right\}, \end{aligned}$$

where the last equality holds since $\bar{V}_i(\boldsymbol{z}) = +\infty$ for all $\boldsymbol{z} \notin \{\beta \in \mathbb{R}^{n-1} | \boldsymbol{e}_{n-1}^T\beta \leq 1,\ \beta \geq 0\}$. From Lemma 4.7.1, given that $n = 3$ and thus $\boldsymbol{y} \in \mathbb{R}^2$, $f_i(\boldsymbol{y})$ is submodular. It remains to show that $w(\boldsymbol{\mu})$ is also submodular. According to Theorem 4.4.4, it suffices to show that $q_i(\boldsymbol{\mu})$ is locally decreasing with μ_j for all $j \neq i$ for all $\boldsymbol{\mu}$. Fix i, j and let $k \neq i, j$. We assume $i > j$ without loss of generality. We have $q_i(\boldsymbol{\mu} - \mu_k\boldsymbol{e}) = q_i(\boldsymbol{\mu})$ from the translation invariance property. But $q_i(\boldsymbol{\mu} - \mu_k\boldsymbol{e}) = \frac{\partial f_k(\mu_i - \mu_k, \mu_j - \mu_k)}{\partial \mu_i}$ is non-decreasing with μ_j due to the submodularity of f_k. Thus $w(\boldsymbol{\mu})$ is submodular and $\boldsymbol{q}(\boldsymbol{\mu}) = \nabla w(\boldsymbol{\mu})$ is a substitutable choice model. □

Proof of Theorem 4.4.5: We first consider the case where $V_i(x_i)$ is differentiable for all $i \in \mathcal{N}$. Let $\lambda(\boldsymbol{\mu})$ be the Lagrangian multiplier of the constraint $\sum_i x_i = 1$. The KKT

conditions (see Bertsekas 2003) for problem (4.5) can be written as:

$$\begin{aligned}
\mu_i - V_i'(q_i(\boldsymbol{\mu})) - \lambda(\boldsymbol{\mu}) \leq 0, \quad & \forall i \in \mathcal{N}; \\
\mu_i - V_i'(q_i(\boldsymbol{\mu})) - \lambda(\boldsymbol{\mu}) = 0, \quad & \forall i \text{ s.t. } q_i(\boldsymbol{\mu}) \neq 0; \\
q_i(\boldsymbol{\mu}) \geq 0, \quad & \forall i \in \mathcal{N}; \\
\textstyle\sum_{i\in\mathcal{N}} q_i(\boldsymbol{\mu}) = 1. &
\end{aligned}$$

Now we consider any two points $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_0 + t\boldsymbol{e_i}$ where $\boldsymbol{e_i}$ is a unit vector along the i-th coordinate axis and $t > 0$. Suppose that there exists a $j \neq i$ such that $q_j(\boldsymbol{\mu}_0 + t\boldsymbol{e_i}) > q_j(\boldsymbol{\mu}_0)$. Since V_j is strictly convex, $V_j'(q_j(\boldsymbol{\mu}_0 + t\boldsymbol{e_i})) > V_j'(q_j(\boldsymbol{\mu}_0))$. There are two possible cases for $q_j(\boldsymbol{\mu}_0)$:

- $q_j(\boldsymbol{\mu}_0) > 0$: In this case, we have $\mu_j - V_j'(q_j(\boldsymbol{\mu}_0 + t\boldsymbol{e_i})) - \lambda(\boldsymbol{\mu}_0 + t\boldsymbol{e_i}) = 0$ and $\mu_j - V_j'(q_j(\boldsymbol{\mu}_0)) - \lambda(\boldsymbol{\mu}_0) = 0$, therefore, we have $\lambda(\boldsymbol{\mu}_0 + t\boldsymbol{e_i}) < \lambda(\boldsymbol{\mu}_0)$.

- $q_j(\boldsymbol{\mu}_0) = 0$: In this case, $\mu_j - V_j'(q_j(\boldsymbol{\mu}_0)) - \lambda(\boldsymbol{\mu}_0) \leq 0$, which implies that $\mu_j - V_j'(q_j(\boldsymbol{\mu}_0 + t\boldsymbol{e_i})) - \lambda(\boldsymbol{\mu}_0) < 0$. But $\mu_j - V_j'(q_j(\boldsymbol{\mu}_0 + t\boldsymbol{e_i})) - \lambda(\boldsymbol{\mu}_0 + t\boldsymbol{e_i}) = 0$, we have $\lambda(\boldsymbol{\mu}_0 + t\boldsymbol{e_i}) < \lambda(\boldsymbol{\mu}_0)$.

In both cases, $\lambda(\boldsymbol{\mu}_0 + t\boldsymbol{e_i}) < \lambda(\boldsymbol{\mu}_0)$. This implies that $q_j(\boldsymbol{\mu}_0 + t\boldsymbol{e_i}) \geq q_j(\boldsymbol{\mu}_0)$ for all $j \neq i$. Note that we also have $q_i(\boldsymbol{\mu}_0 + t\boldsymbol{e_i}) > q_i(\boldsymbol{\mu}_0)$ by Proposition 4.4.1. Therefore, we have $\sum_{j\in\mathcal{N}} q_j(\boldsymbol{\mu}_0 + t\boldsymbol{e_i}) > \sum_{j\in\mathcal{N}} q_j(\boldsymbol{\mu}_0) = 1$, which contradicts with that $\boldsymbol{q}(\boldsymbol{\mu}_0 + t\boldsymbol{e_i}) \in \Delta_{n-1}$. Thus we have $q_j(\boldsymbol{\mu}_0 + t\boldsymbol{e_i}) \leq q_j(\boldsymbol{\mu}_0)$ for all $j \neq i$. Since this is true for all $\boldsymbol{\mu}_0$ and $t > 0$, $\boldsymbol{q}$ 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 Proposition 4.5.2 : By simple algebra, we have:

$$\frac{\partial q_i^{\mathrm{nl}}(\boldsymbol{\mu})}{\partial \mu_j} = K(\boldsymbol{\mu}) \left(-\frac{1}{\lambda_k} \left(\sum_{s\in B_k} \exp(\mu_s/\lambda_k) \right)^{\lambda_k} + \frac{\lambda_k - 1}{\lambda_k} \sum_{l\neq k} \left(\sum_{s\in B_l} \exp(\mu_s/\lambda_l) \right)^{\lambda_l} \right),$$

where

$$K(\boldsymbol{\mu}) = \frac{\exp\left((\mu_i + \mu_j)/\lambda_k\right)\left(\sum_{s\in B_k}\exp(\mu_s/\lambda_k)\right)^{\lambda_k - 2}}{\left(\sum_{l=1}^{K}\left(\sum_{s\in B_l}\exp(\mu_s/\lambda_l)\right)^{\lambda_l}\right)^2} > 0.$$

Clearly,

$$-\frac{1}{\lambda_k}\left(\sum_{s\in B_k}\exp(\mu_s/\lambda_k)\right)^{\lambda_k} < 0$$

and

$$\frac{\lambda_k - 1}{\lambda_k}\sum_{l\neq k}\left(\sum_{s\in B_l}\exp(\mu_s/\lambda_l)\right)^{\lambda_l} > 0.$$

Therefore, when $\boldsymbol{\mu}$ is chosen such that

$$\left(\sum_{s\in B_k}\exp(\mu_s/\lambda_k)\right)^{\lambda_k} \leq (\lambda_k - 1)\sum_{l\neq k}\left(\sum_{s\in B_l}\exp(\mu_s/\lambda_l)\right)^{\lambda_l}$$

and

$$\frac{\partial q_i^{\mathrm{nl}}(\boldsymbol{\mu})}{\partial \mu_j} \geq 0.$$

Finally we note that one can always choose $\boldsymbol{\mu}$ such that the inequality holds. This is because we can choose μ_s large enough for some $s \in B_l$, $l \neq k$. □

Proof of Proposition 4.5.3: In this proof, we use the following Lemma (see Boyd and Vandenberghe 2004).

Lemma 4.7.2 *Define $f(\cdot)$ to be a log-convex function if $\log f(\cdot)$ is a convex function. Then $f(\boldsymbol{x}) = \sum_{i=1}^{n} f_i(\boldsymbol{x})$ is log-convex if $f_i(\boldsymbol{x})$ is log-convex for all $i = 1, ..., n$.*

Now we use the lemma to prove the proposition. We start by proving the first part. It can be easily verified that $w^{nl}(\boldsymbol{\mu})$ is differentiable and $\boldsymbol{q}^{nl}(\boldsymbol{\mu}) = \nabla_{\boldsymbol{\mu}} w^{nl}(\boldsymbol{\mu})$. Now we need to verify that $w^{nl}(\boldsymbol{\mu})$ satisfies the three properties in Definition 5. First, $w^{nl}(\boldsymbol{\mu})$ is clearly monotone since $\lambda_k > 0$ for all $k = 1, ..., K$. The translation invariance property

is also true since

$$\begin{aligned} w^{nl}(\boldsymbol{\mu}+t\boldsymbol{e}) &= \log\left(\sum_{k=1}^{K}\left(\sum_{j\in B_k}\exp((\mu_j+t)/\lambda_k)\right)^{\lambda_k}\right) \\ &= \log\left(\sum_{k=1}^{K}\left(\sum_{j\in B_k}\exp(\mu_j/\lambda_k)\exp(t/\lambda_k)\right)^{\lambda_k}\right) \\ &= \log\left(\sum_{k=1}^{K}\left(\sum_{j\in B_k}\exp(\mu_j/\lambda_k)\right)^{\lambda_k}\exp(t)\right) = w^{nl}(\boldsymbol{\mu})+t. \end{aligned}$$

Now it remains to show that $w^{nl}(\boldsymbol{\mu})$ is convex, or equivalently

$$\sum_{k=1}^{K}\left(\sum_{j\in B_k}\exp(\mu_j/\lambda_k)\right)^{\lambda_k}$$

is log-convex. Note that we have

$$\log\left(\left(\sum_{j\in B_k}\exp(\mu_j/\lambda_k)\right)^{\lambda_k}\right) = \lambda_k\log\left(\sum_{j\in B_k}\exp(\mu_j/\lambda_k)\right),$$

which is convex due to Lemma 4.7.2 since $\exp(\mu_j/\lambda_k)$ is log-convex. This in turn implies that $\left(\sum_{j\in B_k}\exp(\mu_j/\lambda_k)\right)^{\lambda_k}$ is log-convex for all k. Applying Lemma 4.7.2 again concludes the proof for the $w^{nl}(\cdot)$ part.

We next prove the second part. From Theorem 4.2.2, $\boldsymbol{q}^{nl}(\boldsymbol{\mu})$ is consistent with a representative agent model with

$$V^{nl}(\boldsymbol{x}) = \sup_{\boldsymbol{y}}\{\boldsymbol{y}^T\boldsymbol{x} - w^{nl}(\boldsymbol{y})\}, \text{for all } \boldsymbol{x}\in\Delta_{n-1}. \tag{4.12}$$

Since $w^{nl}(\cdot)$ is convex, the first order optimality condition $\nabla w^{nl}(\boldsymbol{y}^*) = \boldsymbol{x}$ is necessary

and sufficient for an $\boldsymbol{y}^*$ to be optimal to (4.12). We now verify that

$$y_i^* = \lambda_k \log x_i + (1-\lambda_k) \log \left(\sum_{j \in B_k} x_j \right),$$

for $i \in B_k$ is an optimal solution, i.e., $\nabla w^{nl}(\boldsymbol{y}^*) = \boldsymbol{x}$. We note that,

$$\begin{aligned} w^{nl}(\boldsymbol{y}^*) &= \log \sum_{k=1}^{K} \left(\sum_{i \in B_k} \exp(y_i^*/\lambda_k) \right)^{\lambda_k} \\ &= \log \sum_{k=1}^{K} \left(\sum_{i \in B_k} x_i \left(\sum_{j \in B_k} x_j \right)^{(1-\lambda_k)/\lambda_k} \right)^{\lambda_k} \\ &= \log \sum_{k=1}^{K} \sum_{i \in B_k} x_i = 0, \end{aligned}$$

where the last equality is because $\boldsymbol{x} \in \Delta_{n-1}$. Thus,

$$\begin{aligned} \frac{\partial w^{nl}(\boldsymbol{y}^*)}{\partial y_i} &= \frac{\exp(y_i^*/\lambda_k) \left(\sum_{j \in B_k} \exp(y_j^*/\lambda_k) \right)^{\lambda_k - 1}}{\sum_{l=1}^{K} \left(\sum_{j \in B_l} \exp(y_j^*/\lambda_l) \right)^{\lambda_l}} \\ &= x_i \left(\sum_{j \in B_k} x_j \right)^{(1-\lambda_k)/\lambda_k} \left(\sum_{j \in B_k} x_j \right)^{(\lambda_k - 1)/\lambda_k} = x_i. \end{aligned}$$

Therefore, $\boldsymbol{y}^*$ is an optimal solution. So we have

$$\begin{aligned} V^{nl}(\boldsymbol{x}) &= (\boldsymbol{y}^*)^T \boldsymbol{x} - w^{nl}(\boldsymbol{y}^*) \\ &= \sum_{k=1}^{K} \left((1-\lambda_k) \sum_{i \in B_k} x_i \log \left(\sum_{j \in B_k} x_j \right) + \lambda_k \sum_{i \in B_k} x_i \log x_i \right). \quad \square \end{aligned}$$

Proof of Proposition 4.5.5: According to Theorem 4.4.4, it suffices to prove that

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

is supermodular if and only if $A_{j,k} - A_{i,k} - A_{i,j} + A_{i,i} \geq 0$ for all distinct $i, j, k \in \mathcal{N}$. For $i \in \mathcal{N}$, $\bar{V}_i$ is an $n-1$ variate quadratic function. Let H^i denote the Hessian matrix of $\bar{V}_i(\boldsymbol{z})$. For $j, k \in \{1, 2, ..., n-1\}$ and $j \neq k$, the off-diagonal element $H^i_{j,k} = A_{\tilde{j},\tilde{k}} - A_{i,\tilde{k}} - A_{i,\tilde{j}} + A_{i,i}$, 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, $\bar{V}_i(\boldsymbol{z})$ 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 $A_{j,k} - A_{i,k} - A_{i,j} + A_{i,i} \geq 0$ for all distinct $i, j, k \in \mathcal{N}$. $\square$

Chapter 5

Online Learning with Non-Convex Losses

In previous chapters, we develop a number of theoretical and practical results regarding discrete choice models. In this chapter , we consider an important application of discrete choice models: the multi-product pricing problem. The multi-product pricing problem is a fundamental decision problem in revenue management and thus has been widely studied in the literature; see for example Talluri and Van Ryzin (2004). In the multi-product pricing problem, a retail store sells a set of substitutable products labeled as $\{1, 2, ..., n\}$. The store is a price setter, i.e., it can set the prices for all n products as $p_1, p_2, ..., p_n$. After observing the prices of different products, consumers come to the store and pick the product they want. Without loss of generality, we assume that each consumer picks one product at a time. Note that consumers can also choose not to buy any products in this store. This option is referred to as the outside option. We denote the vector of purchase probabilities as $\boldsymbol{\pi}(\boldsymbol{p}) = (\pi_1(\boldsymbol{p}), \pi_2(\boldsymbol{p}), ..., \pi_n(\boldsymbol{p}))$. Therefore, the loss (the negative of the revenue) of the store is

$$C(\boldsymbol{p}) = -\sum_{i=1}^{n} p_i \pi_i(\boldsymbol{p}). \tag{5.1}$$

The goal of the store is to minimize the loss.

The customer demand function $\pi_i(\boldsymbol{p})$ is a critical component of the pricing problem.

It is usually assumed that consumers choose products according to a discrete choice model. The deterministic utility of the customer picking product i is usually assumed to linearly decrease in the price of that product, i.e., for any i, $\mu_i = a_i - b_i p_i$, where $b_i > 0$ is called the price sensitivity parameter for product i. It is shown in Hanson and Martin (1996) that $C(\boldsymbol{p})$ under MNL model is not even quasi-convex. Therefore, several other papers propose to inverse the functional relationship between $\boldsymbol{p}$ and $\boldsymbol{\pi}$, and then optimize the loss function over $\boldsymbol{\pi}$. Song and Xue (2007) and Dong et al. (2009) show that the loss function is convex in $\boldsymbol{\pi}$ when the demand is MNL. Li and Huh (2011) extend this method to the pricing problem under the nested logit model and show that the loss function is still convex when scale parameters are in the unit interval and price sensitivity parameters are the same for the products in the same nest.

In reality, however, the customer demand is usually not known as a priori and thus has to learned from pricing experiments. Instead of developing models specifically on dynamic pricing with demand learning (see den Boer 2015 for a recent survey of this stream of literature), we adopt a more general framework rooted in the well celebrated online learning literature. Among the existing online learning models, online convex optimization (OCO) has been studied extensively in the literature. In OCO, at each period $t \in \{1, 2, \ldots, T\}$, an online player chooses a feasible strategy $\boldsymbol{x}_t$ from a decision set $\mathcal{X} \subset \mathbb{R}^n$, and suffers from a loss given by $f_t(\boldsymbol{x}_t)$, where $f_t(\cdot)$ is a convex loss function. One key feature of the OCO is that the player must make a decision for period t without knowing the loss function $f_t(\cdot)$. The performance of an OCO algorithm is usually measured by the *stationary regret*, which compares the accumulated loss suffered by the player with the loss suffered by the best fixed strategy. Specifically, the stationary regret is defined as

$$\text{Regret}_T^S(\{\boldsymbol{x}_t\}_1^T) = \sum_{t=1}^{T} f_t(\boldsymbol{x}_t) - \sum_{t=1}^{T} f_t(\boldsymbol{x}^*), \tag{5.2}$$

where $\boldsymbol{x}^*$ is one best fixed decision in hindsight, i.e. $\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\boldsymbol{x})$. Several sub-linear cumulative regret bounds measured by stationary regret have been established in various papers in the literature. For example, Zinkevich (2003) proposed an online gradient descent algorithm which achieves an regret bound of order $O(\sqrt{T})$ for

convex loss functions. The order of the regret can be further improved to $O(\log T)$ if the loss functions are strongly convex (see Hazan et al. 2007). Moreover, the bounds are shown to be tight for the OCO with convex / strongly convex loss functions respectively in Abernethy et al. (2009). One important extension of the OCO is the so-called *bandit online convex optimization*, where the online player is only supposed to know the function value $f_t(\boldsymbol{x}_t)$ at $\boldsymbol{x}_t$, instead of the entire function $f_t(\cdot)$. In particular, when the player can only observe the function value at a single point, Flaxman et al. (2005) established an $O(T^{3/4})$ regret bound for general convex loss functions by constructing a zeroth-order approximation of the gradient. Assuming that the loss functions are s-mooth, the regret bound can be improved to $O(T^{2/3})$ by incorporating a self-concordant regularizer (see Saha and Tewari 2011). Alternatively, if multiple points can be inquired at the same time, Agarwal et al. (2010) showed that the regrets can be further improved to $O(T^{1/2})$ and $O(\log T)$ for convex / strongly convex loss functions respectively.

As suggested by its name, the loss functions in the OCO are assumed to be convex. As shown in Hanson and Martin (1996), the cost function in (5.1) is not even quasi-convex in the price vector under the simplest MNL choice model. Though the cost function could be convex in the market share under some demand models, policies based on market share vectors are usually difficulty to implement without the knowledge of the demand function. Additional constraints have to be made about the demand function (e.g., Chen et al. 2014 adopts the parametric approach) and the resulting policies are usually complicated. Therefore, the key assumption that the cost function is convex in OCO limits its applicability to the pricing problem. Only a handful of papers study online learning with non-convex loss functions. In the existing works, most of the time heuristic algorithms were proposed (see for example Gasso et al. 2011; Ertekin et al. 2011) without focussing on establishing sublinear regret bounds. There are a few noticeable exceptions though. Hazan and Kale (2012) developed an algorithm that achieves $O(T^{1/2})$ and $O(T^{2/3})$ regret bounds for the full information and bandit settings respectively, by assuming the loss functions to be submodular. Zhang et al. (2015) showed that an $O(T^{2/3})$ regret bound still holds if the loss functions are in the form of composition between a non-increasing scalar function and a linear function.

The stationary regret requires the benchmark strategy to remain unchanged throughout the periods. This assumption may not be relevant in some of the applications.

Recently, a new performance metric known as the *non-stationary regret* is proposed by Besbes et al. (2015). The non-stationary regret compares the cumulative losses of the online player with the losses of the best possible responses:

$$\text{Regret}_T^{NS}(\{\boldsymbol{x}_t\}_1^T) = \sum_{t=1}^{T} f_t(\boldsymbol{x}_t) - \sum_{t=1}^{T} f_t(\boldsymbol{x}_t^*), \tag{5.3}$$

where $\boldsymbol{x}_t^* \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} f_t(\boldsymbol{x})$. Clearly, the non-stationary regret is never less than the stationary regret. Besbes et al. (2015) prove that if there is no restriction on the changes of the loss functions, then the non-stationary regret is linear in T regardless of the strategies. To obtain meaningful bounds, the authors assumed that the temporal change of the sequence of the function $\{f_t\}_1^T$ is bounded. Specifically, the loss functions are assumed to be taken from the set

$$\mathcal{V} := \left\{ \{f_1, f_2, \ldots, f_T\} : \sum_{t=1}^{T-1} \|f_t - f_{t+1}\| \leq V_T \right\}, \tag{5.4}$$

where $\|f_t - f_{t-1}\| = \sup_{\boldsymbol{x}\in\mathcal{X}} |f_t(\boldsymbol{x}) - f_{t-1}(\boldsymbol{x})|$. For nonzero temporal change V_T, Besbes et al. (2015) then proposes algorithms with sub-linear non-stationary regret bounds: $O(V_T^{1/3}T^{2/3})$, $O(V_T^{1/2}T^{1/2})$, and $O(V_T^{1/3}T^{2/3})$ respectively, for the cases where loss functions are: convex with noisy gradients, strongly convex with noisy gradients, and strongly convex with noisy function values. Note that $V_T > 0$ is assumed for the bounds to hold. More recently, Yang et al. (2016) also studied the non-stationary regret bounds for OCO. They proposed an uncertainty set $\mathcal{S}_T^p$ of the sequence of functions in which the worst-case variation of the optimal solution $\boldsymbol{x}_t^*$ of $f_t(\cdot)$ (referred to as the path variation) is bounded:

$$\mathcal{S}_T^p := \Bigg\{ \{f_1, f_2, \ldots, f_T\} : \max_{\boldsymbol{x}_t^* \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} f_t(\boldsymbol{x})} \sum_{t=1}^{T-1} \|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\| \leq V_T \Bigg\}.$$

They then proved some upper and lower bounds for the several different feedback structure and functional classes (within the convex function class). In particular, they showed that some existing algorithms (with some modification) can achieve the V_T ,$\sqrt{TV_T}$ and $\sqrt{TV_T}$ for true gradient (with smooth condition), noisy gradient and two-point bandit

feedback, respectively. Note that $V_T > 0$ is assumed in the bounds.

In this chapter, we consider online non-convex optimization with non-stationary regret as the performance metric. To the best of our knowledge, such a combination had not been studied before. For each period t, even after the decision $\boldsymbol{x}_t$ is made, the online player is not assumed to know the function $f_t(\cdot)$; instead, only some partial information regarding the loss at $\boldsymbol{x}_t$ is revealed. Specifically, only $\nabla f(\boldsymbol{x}_t)$ (in the first-order setting) or $f(\boldsymbol{x}_t)$ (the zeroth-order setting) is available to the player. Similar to Yang et al. (2016), we define the uncertainty set $\mathcal{S}_T$ of the sequence of functions as follows:

$$\mathcal{S}_T := \Bigg\{ \{f_1, f_2, \ldots, f_T\} : \exists \boldsymbol{x}_t^* \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} f_t(\boldsymbol{x}), \\ t = 1, ..., T, \quad \text{s.t.} \sum_{t=1}^{T-1} \|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\| \leq V_T \Bigg\}.$$

Note that $\mathcal{S}_T^p \subseteq \mathcal{S}_T$ for the same V_T. In particular, consider a static sequence $f_t(\cdot) = f(\cdot), t = 1, .., T$ where $f(\cdot)$ has multiple optimal solutions. This sequence would clearly belong to $\mathcal{S}_T$ even when $V_T = 0$. However, V_T would have to be linear in T in order for this sequence to be in $\mathcal{S}_T^p$. We propose the Online Normalized Gradient Descent (ONGD) and the novel Bandit Online Normalized Gradient Descent (BONGD) algorithms for the first-order setting and the zeroth-order setting respectively. For the loss functions satisfying (5.6) and a condition to be introduced later, we show that these two algorithms both achieve $O(\sqrt{T + V_T T})$ regret bound. Compared to the regret bounds in Yang et al. (2016), our regret bound for the first-order setting is worse but is the same for the zeroth-order setting. Note however, that our loss functions are non-convex and we use a weaker version of variational constraints.

Regarding non-convex objective function, a related work in the literature is Hazan et al. (2015), where the authors propose a Normalized Gradient Descent (NGD) method for solving an optimization model with the so-called *strictly locally quasi-convex* (SLQC) function as the objective. They further show that the NGD converges to an ϵ-optimal minimum within $O(1/\epsilon^2)$ iterations. This chapter generalizes the results in Hazan et al. (2015) in the following aspects:

- Hazan et al. (2015) considers an optimization model, while this chapter considers an online learning model.

- Hazan et al. (2015) assumes the objective function to be strictly locally quasi-convex (SLQC). In this chapter we introduce the notion of weak pseudo-convexity (WPC), which will be shown to be a weaker condition than the SLQC. We show that the regret bounds hold if the objective function is weak pseudo-convex.
- Hazan et al. (2015) considers only the first-order setting, while our proposed BONGD algorithm works for the bandit (zeroth-order) setting as well.

The rest of the chapter is organized as follows. Section 5.1 presents some preparations including the assumptions and notations. In Section 5.2, we present the ONGD algorithm and prove its regret bound. In Section 5.3, we present the BONGD algorithm for the zeroth-order setting and show its regret bound under some assumptions. Finally, we conclude this chapter in Section 5.4.

5.1 Problem Setup

In this section, we present the assumptions underlying our online learning model that will be used in the chapter. Let $\mathcal{X} \subset \mathbb{R}^n$ be a convex decision set that is known to the player. For every period $t \in \{1, 2, \ldots, T\}$, the loss function is $f_t(\cdot)$. Throughout the chapter, we assume that $\mathcal{X} \subset \mathbb{R}^n$ is bounded, i.e., there exists $R > 0$ such that $\|\boldsymbol{x}\| \leq R$ for all $\boldsymbol{x} \in \mathcal{X}$. We present the following definitions regarding the loss functions.

Definition 13 (Bounded Gradient) *A function $f(\cdot)$ is said to have bounded gradient if there exists a finite positive value M such that for all $\boldsymbol{x} \in \mathcal{X}$, it holds that $\|\nabla f(\boldsymbol{x})\| \leq M$.*

Note that if $f(\cdot)$ has bounded gradient, then it is also Lipschitz continuous with Lipschitz constant M on the set $\mathcal{X}$.

Definition 14 (Weak Pseudo-Convexity) *A function $f(\cdot)$ is said to be **weakly pseudo-convex** (WPC) if there exists $K > 0$ such that*

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \leq K \frac{\nabla f(\boldsymbol{x})^\top (\boldsymbol{x} - \boldsymbol{x}^*)}{\|\nabla f(\boldsymbol{x})\|},$$

holds for all $\boldsymbol{x} \in \mathcal{X}$, with the convention that $\frac{\nabla f(\boldsymbol{x})}{\|\nabla f(\boldsymbol{x})\|} = 0$ if $\nabla f(\boldsymbol{x}) = 0$, where $\boldsymbol{x}^*$ is one optimal solution, i.e., $\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$.

Here we discuss some implications of the weak pseudo-convexity. If a differentiable function $f(\cdot)$ is Lipschitz continuous and pseudo-convex, then we have (see similar derivation in Nesterov (2004))

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) \leq M \frac{\nabla f(\boldsymbol{x})^\top (\boldsymbol{x} - \boldsymbol{y})}{\|\nabla f(\boldsymbol{x})\|},$$

for all $\boldsymbol{y}, \boldsymbol{x}$ with $f(\boldsymbol{x}) \geq f(\boldsymbol{y})$, where M is Lipschitz constant. Therefore, we can simply let $K = M$, and the function is also weakly pseudo-convex. Moreover, as another example, the star-convex function proposed by Nesterov and Polyak (2006) is weakly pseudo-convexity.

Proposition 5.1.1 *If $f(\cdot)$ is star-convex and smooth with bounded gradient in $\mathcal{X}$, then $f(\cdot)$ is weakly pseudo-convex.*

The proof of Proposition 5.1.1 can be found in Section 5.5. We next introduce a property that is essentially the same as the SLQC property introduced in Hazan et al. (2015).

Definition 15 (Acute Angle) *Gradient of $f(\cdot)$ is said to satisfy the **acute angle** condition if there exists a positive value Z such that*

$$\begin{aligned} \cos(\nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}^*) &= \frac{\nabla f(\boldsymbol{x})^\top (\boldsymbol{x} - \boldsymbol{x}^*)}{\|\nabla f(\boldsymbol{x})\| \cdot \|\boldsymbol{x} - \boldsymbol{x}^*\|} \\ &\geq Z > 0, \end{aligned}$$

holds for all $\boldsymbol{x} \in \mathcal{X}$, with the convention that $\frac{\nabla f(\boldsymbol{x})}{\|\nabla f(\boldsymbol{x})\|} = 0$ if $\nabla f(\boldsymbol{x}) = 0$, where $\boldsymbol{x}^*$ is one optimal solution, i.e., $\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$.

The following proposition shows that the acute angle condition together with the Lipschitz continuity implies the weak pseudo-convexity.

Proposition 5.1.2 *If $f(\cdot)$ has bounded gradient and satisfies the acute angle condition, then $f(\cdot)$ is weakly pseudo-convex.*

The proof of Proposition 5.1.2 can be found in Section 5.5. The class of weakly pseudo-convex functions certainly go beyond the acute angle condition. For example, below is another class of functions satisfying the WPC.

Proposition 5.1.3 *If $f(\cdot)$ has bounded gradient and satisfy the α-homogeneity with respect to its minimum, i.e., there exists $\alpha > 0$ satisfying*

$$f(t(\boldsymbol{x} - \boldsymbol{x}^*) + \boldsymbol{x}^*) - f(\boldsymbol{x}^*) = t^\alpha (f(\boldsymbol{x}) - f(\boldsymbol{x}^*)), \tag{5.5}$$

for all $\boldsymbol{x} \in \mathcal{X}$ and $t \geq 0$ where $\boldsymbol{x}^* = \arg\min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$, then $f(\cdot)$ is weak pseudo-convex.

The proof of Proposition 5.1.3 can be found in Section 5.5. Proposition 5.1.3 suggests that all non-negative homogeneous polynomial satisfies WPC with respect to 0. Take $f(\boldsymbol{x}) = (x_1^2 + x_2^2)^2 + 10(x_1^2 - x_2^2)^2$ as an example. It is easy to verify that $f(\cdot)$ satisfies the condition in Proposition 5.1.3, and thus is weakly pseudo-convex. In Figure 5.1, the curvature of $f(\boldsymbol{x})$ and a sub-level set of this function are plotted. The function is not quasi-convex since the sub-level set is non-convex. However, this function satisfies the acute-angle condition in Definition 15.

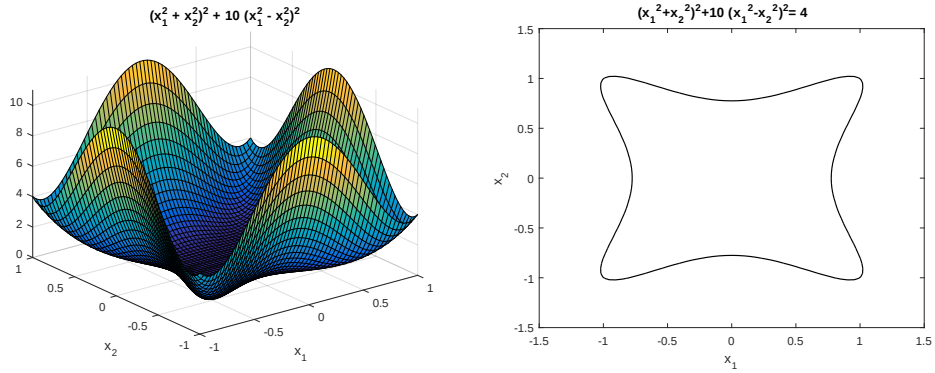

Figure 5.1: Plot of a WPC function that is not quasi-convex

Note that if $f_i(\boldsymbol{x})$ is α_i-homogeneous with respect to the shared minimum $\boldsymbol{x}^*$ for all $1 \leq i \leq I$ with $\alpha_i \geq \alpha > 0$, and the gradient of $f_i(\cdot)$ is uniformly bounded over a set $\mathcal{X}$, then $\sum_{i=1}^I f_i(\boldsymbol{x})$ is WPC. As a result, we can construct functions that are WPC but do not satisfy the acute-angle condition. Consider a two-dimensional function $f(\boldsymbol{x}) = x_1^2 + |x_2|^{3/2}$, and suppose that $\mathcal{X}$ is the unit disc centered at the origin. Clearly, $f(\boldsymbol{x})$ is differentiable and Lipchitz continuous in $\mathcal{X}$. Also, it is the sum of a 2-homogeneous function and a 3/2-homogeneous function with a shared minimum $(0, 0)$. Thus $f(\boldsymbol{x})$ is

WPC. We compute that

$$\begin{aligned}\cos(\nabla f(\boldsymbol{x}), \boldsymbol{x}-\boldsymbol{x}^*) &= \frac{\nabla f(\boldsymbol{x})^\top(\boldsymbol{x}-\boldsymbol{x}^*)}{\|\nabla f(\boldsymbol{x})\|\cdot\|\boldsymbol{x}-\boldsymbol{x}^*\|} \\ &= \frac{2x_1^2+\frac{3}{2}|x_2|^{3/2}}{\sqrt{(4x_1^2+\frac{9}{4}|x_2|)(x_1^2+x_2^2)}}.\end{aligned}$$

Consider a parameterized path $(x_1, x_2) = (t^{1/2}, t^{2/3})$ with $t > 0$. On this path, we have

$$\begin{aligned}\cos(\nabla f(\boldsymbol{x}), \boldsymbol{x}-\boldsymbol{x}^*) &= \frac{2x_1^2+\frac{3}{2}|x_2|^{3/2}}{\sqrt{(4x_1^2+\frac{9}{4}|x_2|)(x_1^2+x_2^2)}} \\ &= \frac{7t}{2\sqrt{(4t+\frac{9}{4}t^{2/3})(t+t^{4/3})}} \\ &= \frac{7t^{1/6}}{2\sqrt{(4t^{1/3}+\frac{9}{4})(1+t^{1/3})}}.\end{aligned}$$

Therefore, along the path, as t approaches to 0, we have $\cos(\nabla f(\boldsymbol{x}), \boldsymbol{x}-\boldsymbol{x}^*) \to 0$. This example shows that a WPC function may fail to satisfy the acute angle condition.

As we mentioned before, in order to establish some sub-linear non-stationary regret bound, we need to confine the loss functions $\{f_t\}_1^T$ in a unified manner. Therefore, we introduce the uncertainty set of the loss functions $\mathcal{S}_T$, as a set of admissible loss functions where their total variation of the minimizers are bounded by V_T.

Definition 16 *The uncertainty set of functions $\mathcal{S}_T$ is defined as*

$$\begin{aligned}\mathcal{S}_T := \Bigg\{ &\{f_1, f_2, \ldots, f_T\} : \exists \boldsymbol{x}_t^* \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} f_t(\boldsymbol{x}), t = 1, ..., T, \\ &\text{s.t.} \sum_{t=1}^{T-1} \|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\| \le V_T \Bigg\},\end{aligned} \tag{5.6}$$

where $V_T \ge 0$.

In the zeroth-order setting to be discussed in Section 5.3, only the function value is available. Therefore, some randomized approaches are needed in the algorithm. To account for this situation, we introduce the expected non-stationary regret for an algorithm that outputs a random sequence $\{\boldsymbol{x}_t\}_1^T$ in the performance metric.

Definition 17 *The expected non-stationary regret for a randomized algorithm $\mathcal{A}$ is defined as*

$$ERegret_T^{NS}(\{\boldsymbol{x}_t\}_1^T) = \mathsf{E}\left[\sum_{t=1}^{T}(f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_t^*))\right], \tag{5.7}$$

where the expectation is taken over the filtration generated by the random sequence $\{\boldsymbol{x}_t\}_1^T$ produced by $\mathcal{A}$.

5.2 The First-Order Setting

In this section, we assume that for each period, the gradient information at the current point is available to the online player after the decision is made. Specifically, at each period t, the sequence of the events is as follows:

1. The online player chooses a strategy $\boldsymbol{x}_t$;
2. The online player receives the feedback $\nabla f_t(\boldsymbol{x}_t)$;
3. Regret $f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_t^*)$ incurs (but is not necessarily known to the online player).

We propose the Online Normalized Gradient Descend algorithm (ONGD) in this setting. The normalized gradient descent method was first proposed in Nesterov (2004) which can be applied to solve the pseudo-convex minimization problem. The ONGD algorithm uses the first-order information $\nabla f_t(\boldsymbol{x}_t)$ to compute the normalized vector $\nabla f_t(\boldsymbol{x}_t)/\|\nabla f_t(\boldsymbol{x}_t)\|$ as the search direction. Similar to the standard gradient method, it moves along that search direction with a specific stepsize $\eta > 0$ and then projects the point back to the decision set $\mathcal{X}$; see Algorithm 2 for the details. Note that in Algorithm 2, $\prod_{\mathcal{X}}(\boldsymbol{y}) := \arg\min_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{y} - \boldsymbol{x}\|$ is the projection operator. The main result is shown in the following theorem which claims an $O(\sqrt{T + V_T T})$ non-stationary regret bound for ONGD.

Theorem 5.2.1 *Let V_T be as defined in (5.6). For any sequence of loss functions $\{f_t\}_1^T \in \mathcal{S}_T$ where f_t is weakly pseudo-convex with common constant K, let the stepsize $\eta = \sqrt{\frac{4R^2+6RV_T}{T}}$. Then, the following regret bound holds for ONGD:*

$$\begin{aligned} \text{Regret}_T^{NS}(\{\boldsymbol{x}_t\}_1^T) &\leq K\sqrt{T(R^2 + 1.5RV_T)} \\ &= O(\sqrt{T + V_T T}). \end{aligned}$$

Algorithm 2: Online Normalized Gradient Descent

Input: feasible set $\mathcal{X}$, # time period T
Initialization: $\boldsymbol{x}_1 \in \mathcal{X}$
for $t = 1$ **to** T **do**
  chooses $\boldsymbol{x}_t$ and receives the feedback $\boldsymbol{g}_t = \nabla f_t(\boldsymbol{x}_t)$
  if $\|\boldsymbol{g}_t\| > 0$ **then**
    $\boldsymbol{x}_{t+1} = \prod_{\mathcal{X}} \left(\boldsymbol{x}_t - \eta \frac{\boldsymbol{g}_t}{\|\boldsymbol{g}_t\|} \right)$
  else
    $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t$
  end if
end for

Proof: Let $\boldsymbol{x}_t^*, t = 1, .., T$ be the sequence of optimal solutions satisfying the condition in Definition 16, and $z_t := \|\boldsymbol{x}_t - \boldsymbol{x}_t^*\|$. Then we have:

$$\begin{aligned}
z_{t+1}^2 &= \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t+1}^*\|^2 \\
&= \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t^*\|^2 + \|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\|^2 + 2(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t^*)^\top (\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*) \\
&\leq \left\| \prod_{\mathcal{X}} \left(\boldsymbol{x}_t - \eta \frac{\boldsymbol{g}_t}{\|\boldsymbol{g}_t\|} \right) - \boldsymbol{x}_t^* \right\|^2 + 6R\|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\| \\
&\leq \left\| \boldsymbol{x}_t - \eta \frac{\boldsymbol{g}_t}{\|\boldsymbol{g}_t\|} - \boldsymbol{x}_t^* \right\|^2 + 6R\|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\| \\
&= z_t^2 + \eta^2 - 2\eta \frac{\boldsymbol{g}_t^\top (\boldsymbol{x}_t - \boldsymbol{x}_t^*)}{\|\boldsymbol{g}_t\|} + 6R\|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\|.
\end{aligned}$$

By rearranging terms and multiplying K on both sides we have

$$K \frac{\boldsymbol{g}_t^\top (\boldsymbol{x}_t - \boldsymbol{x}_t^*)}{\|\boldsymbol{g}_t\|} \leq \frac{K}{2\eta} \left(z_t^2 - z_{t+1}^2 + \eta^2 + 6R\|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\| \right). \tag{5.8}$$

By Definition 14, noting that $\boldsymbol{g}_t = \nabla f_t(\boldsymbol{x}_t)$, we have

$$\begin{aligned}
& f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_t^*) \\
\leq\ & K \frac{\nabla f_t(\boldsymbol{x}_t)^\top (\boldsymbol{x}_t - \boldsymbol{x}_t^*)}{\|\nabla f_t(\boldsymbol{x}_t)\|} \\
\leq\ & \frac{K}{2\eta} \left(z_t^2 - z_{t+1}^2 + \eta^2 + 6R\|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\| \right).
\end{aligned}$$

Summing these inequalities from $t = 1, ..., T$, we have

$$\begin{aligned} & \text{Regret}_T^{NS}(\{\boldsymbol{x}_t\}_1^T) \\ \leq \; & \frac{K}{2\eta}\left(z_1^2 - z_{T+1}^2 + T\eta^2 + 6R\sum_{t=1}^{T}\|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\|\right) \\ \leq \; & \frac{K}{2\eta}\left(4R^2 + T\eta^2 + 6RV_T\right). \end{aligned}$$

As a result, by noting $\eta = \sqrt{\frac{4R^2+6RV_T}{T}}$, we have

$$\begin{aligned} \text{Regret}_T^{NS}(\{\boldsymbol{x}_t\}_1^T) &\leq K\sqrt{T(R^2 + 1.5RV_T)} \\ &= O(\sqrt{T + V_T T}). \end{aligned}$$

□

5.3 The Zeroth-Order Setting

In the previous section, it is assumed that the gradient information is available, which may not be the case in some applications. Such exceptions include the multi-armed bandit problem, dynamic pricing and Bayesian optimization. Therefore, in this section, we consider the setting where the online player only receives the function value $f_t(\boldsymbol{x}_t)$, instead of the gradient $\nabla f_t(\boldsymbol{x}_t)$, as the feedback.

As mentioned above, the zeroth-order (or bandit) setting has been studied in the OCO literature. The main technique in the OCO literature (see Flaxman et al. 2005 for example) is to construct a zeroth-order approximation of the gradient of a smoothed function. That smoothed function is often created by integrating the original loss function with a chosen probability distribution. By querying some random samples of the function value according to a probability distribution, the player is able to create an unbiased zeroth-order approximation of the gradient of the smoothed function. This is, however, not applicable in our online normalized gradient descent algorithm since what we need is the direction of the gradient. Therefore, we shall first develop a new type

of zeroth-order oracle that can approximate the gradient direction without averaging multiple samples of gradients when the norm of the gradient is not too small.

To proceed, we require some additional conditions on the loss function.

Definition 18 (Error Bound) *There exists $D > 0$ and $\gamma > 0$ such that*

$$\|\boldsymbol{x} - \boldsymbol{x}_t^*\| \leq D\|\nabla f_t(\boldsymbol{x})\|^\gamma,$$

for all $\boldsymbol{x} \in \mathcal{X}$, $1 \leq t \leq T$, where $\boldsymbol{x}_t^*$ is the optimal solution to $f_t(\cdot)$, i.e., $\boldsymbol{x}_t^* = \arg\min_{\boldsymbol{x}\in\mathcal{X}} f_t(\boldsymbol{x})$.

Since $\mathcal{X}$ is a compact set, the error bound condition is essentially the requirement for a unique optimal solution and no local minimum.

Definition 19 (Lipschitz Gradient) *There exists a positive number L, such that*

$$\|\nabla f_t(\boldsymbol{x}) - \nabla f_t(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|,$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$, where $1 \leq t \leq T$.

We introduce some notations that will be used in subsequent analysis.

$S(n)$:	the unit sphere in $\mathbb{R}^n$;
$m(A)$:	the measure of set $A \subset \mathbb{R}^n$;
β_n:	the area of the unit sphere $S(n)$;
dS_n:	the differential unit on the unit sphere $S(n)$;
$\mathbf{1}_A(\boldsymbol{x})$:	the indicator function of set A;
$\text{sign}(\cdot)$:	the sign function.

Before we present the main results, several lemmas are in order. The first lemma considers some geometric properties of the unit sphere.

Lemma 5.3.1 *For any non-zero vector $\boldsymbol{d} \in \mathbb{R}^n$ and $\delta < 1$, let $S_\delta^{\boldsymbol{x}}$ be defined as*

$$S_\delta^{\boldsymbol{x}} := \left\{ \boldsymbol{v} \in S(n) | \text{ s.t. } |\boldsymbol{d}^\top \boldsymbol{v}| < \delta^2 \right\}.$$

If $\|\boldsymbol{d}\| \geq \delta$, then there exists a constant $C_n > 0$, such that

$$m(S_\delta^{\boldsymbol{x}}) < C_n \delta.$$

Proof: We have

$$m(S_\delta^{\boldsymbol{x}}) = \int\limits_{v \in S(n) \cap S_\delta^{\boldsymbol{x}}} dS_n.$$

By the symmetry of $S(n)$, we may assume w.l.o.g. that $\boldsymbol{d} = (0, \ldots, 0, \|\boldsymbol{d}\|)^\top$. Let $a = \frac{\delta^2}{\|\boldsymbol{d}\|}$. Since $a < 1$, we have

$$\begin{aligned}
& m(S_\delta^{\boldsymbol{x}}) \\
= & \int\limits_{v \in S(n)} \mathbf{1}_{\left\{-\frac{\delta^2}{\|\boldsymbol{d}\|} \le v_n \le \frac{\delta^2}{\|\boldsymbol{d}\|}\right\}}(v) dS_n \\
= & \; 2 \int\limits_{1-a^2 \le v_1^2 + \cdots + v_{n-1}^2 \le 1} \frac{1}{\sqrt{1 - v_1^2 - \cdots - v_{n-1}^2}} dv_1 \cdots dv_{n-1} \\
= & \; 2 \int\limits_{\sqrt{1-a^2} \le r \le 1} \frac{r^{n-2}}{\sqrt{1-r^2}} dr \cdot dS_{n-1} \\
= & \; 2\beta_{n-1} \int\limits_{\sqrt{1-a^2} \le r \le 1} \frac{r^{n-2}}{\sqrt{1-r^2}} dr \\
\le & \; 2\beta_{n-1} \int\limits_{\sqrt{1-a^2} \le r \le 1} \frac{1}{\sqrt{1-r^2}} dr \\
= & \; 2\beta_{n-1} \left(\frac{\pi}{2} - \arcsin(\sqrt{1-a^2}) \right) \\
= & \; 2\beta_{n-1} (\arcsin a) < 2\beta_{n-1} \frac{\pi}{2} a = \pi \beta_{n-1} \frac{\delta^2}{\|\boldsymbol{d}\|} \le \pi \beta_{n-1} \delta.
\end{aligned}$$

By setting $C_n = \pi \beta_{n-1}$, the desired result follows. □

The next lemma leads to an unbiased first-order estimator of the direction of a vector.

Lemma 5.3.2 *Suppose $\boldsymbol{d} \in \mathbb{R}^n$, and $\boldsymbol{d} \neq 0$. Then,*

$$\int\limits_{\boldsymbol{v} \in S(n)} \operatorname{sign}(\boldsymbol{d}^\top \boldsymbol{v}) \boldsymbol{v} dS_n = P_n \frac{\boldsymbol{d}}{\|\boldsymbol{d}\|},$$

where P_n is a constant.

Proof: By the symmetry of $S(n)$, again we may assume $\boldsymbol{d} = (0, \ldots, 0, \|\boldsymbol{d}\|)^\top$, and

$$\int\limits_{\boldsymbol{v}\in S(n)} \operatorname{sign}(\boldsymbol{d}^\top \boldsymbol{v})\boldsymbol{v} dS_n = 2 \int\limits_{\boldsymbol{v}\in S(n)} \mathbf{1}_{v_n\geq 0}(\boldsymbol{v})\boldsymbol{v} dS_n.$$

Notice that if $\boldsymbol{v} \in S(n)$, then $\boldsymbol{u} = (-v_1, -v_2, \ldots, -v_{n-1}, v_n)^\top$ is also in $S(n)$. As a result, the above integral will be on the direction of $\frac{\boldsymbol{d}}{\|\boldsymbol{d}\|} = (0, 0, \ldots, 0, 1)^\top$, and its length is given by

$$\begin{aligned}
& 2 \int\limits_{v\in S(n)} \mathbf{1}_{v_n\geq 0}(\boldsymbol{v}) v_n dS_n \\
= \; & 2 \int\limits_{0\leq v_1^2+\cdots+v_{n-1}^2\leq 1} \sqrt{1 - v_1^2 - \cdots - v_{n-1}^2} dS_n \\
= \; & 2 \int\limits_{0\leq v_1^2+\cdots+v_{n-1}^2\leq 1} \frac{\sqrt{1 - v_1^2 - \cdots - v_{n-1}^2}}{\sqrt{1 - v_1^2 - \cdots - v_{n-1}^2}} dv_1 \ldots dv_{n-1} \\
= \; & 2 \int\limits_{0\leq r\leq 1} r^{n-2} dr dS_{n-1} \\
= \; & \frac{2\beta_{n-1}}{n-1} := P_n.
\end{aligned}$$

□

Using the previous lemmas, we have the following result which constructs a zeroth-order estimator for the normalized gradient.

Theorem 5.3.3 *Suppose $f(\boldsymbol{x})$ has Lipschitz gradient and $\|\nabla f(\boldsymbol{x})\| \geq \delta$ at $\boldsymbol{x}$. Let $\epsilon = \frac{\delta^2}{L}$. Then we have*

$$\left\| \mathsf{E}_{S(n)} \left[\operatorname{sign}(f(\boldsymbol{x} + \epsilon \boldsymbol{v}) - f(\boldsymbol{x}))\boldsymbol{v} \right] - Q_n \frac{\nabla f(\boldsymbol{x})}{\|\nabla f(\boldsymbol{x})\|} \right\| \leq 2 D_n \delta$$

where $\boldsymbol{v}$ is a random vector uniformly distributed over $S(n)$, and $Q_n = \frac{P_n}{\beta_n}$ and $D_n = \frac{C_n}{\beta_n}$.

Proof: By Definition 19, we have

$$\begin{aligned} & |f(\boldsymbol{x}+\epsilon\boldsymbol{v}) - f(\boldsymbol{x}) - \epsilon\nabla f(\boldsymbol{x})^\top \boldsymbol{v}| \leq \frac{\epsilon L}{2}\|\boldsymbol{v}\|^2 \\ \iff \quad & \nabla f(\boldsymbol{x})^\top \boldsymbol{v} - \frac{\epsilon}{2}L \leq \frac{f(\boldsymbol{x}+\epsilon\boldsymbol{v}) - f(\boldsymbol{x})}{\epsilon} \leq \nabla f(\boldsymbol{x})^\top \boldsymbol{v} + \frac{\epsilon}{2}L. \end{aligned}$$

Since $|\nabla f(\boldsymbol{x})^\top \boldsymbol{v}| \geq \delta^2$ for $\boldsymbol{v} \in S(n) \setminus S_\delta^{\boldsymbol{x}}$, if we let $\epsilon = \frac{\delta^2}{L}$, we have

$$\nabla f(\boldsymbol{x})^\top \boldsymbol{v} - \frac{\delta^2}{2} \leq \frac{f(\boldsymbol{x}+\epsilon\boldsymbol{v}) - f(\boldsymbol{x})}{\epsilon} \leq \nabla f(\boldsymbol{x})^\top \boldsymbol{v} + \frac{\delta^2}{2}.$$

Thus,

$$\begin{aligned} & \operatorname{sign}\left(\nabla f(\boldsymbol{x})^\top \boldsymbol{v}\right) = \operatorname{sign}\left(\nabla f(\boldsymbol{x})^\top \boldsymbol{v} - \frac{\delta^2}{2}\right) \\ \leq \quad & \operatorname{sign}\left(\frac{f(\boldsymbol{x}+\epsilon\boldsymbol{v}) - f(\boldsymbol{x})}{\epsilon}\right) \leq \operatorname{sign}\left(\nabla f(\boldsymbol{x})^\top \boldsymbol{v} + \frac{\delta^2}{2}\right) \\ = \quad & \operatorname{sign}\left(\nabla f(\boldsymbol{x})^\top \boldsymbol{v}\right), \end{aligned}$$

implying $\operatorname{sign}(\nabla f(\boldsymbol{x})^\top \boldsymbol{v}) = \operatorname{sign}\left(\frac{f(\boldsymbol{x}+\epsilon\boldsymbol{v})-f(\boldsymbol{x})}{\epsilon}\right)$. Therefore,

$$\begin{aligned} & \beta_n \mathsf{E}_{S(n)}\left[\operatorname{sign}(f(\boldsymbol{x}+\epsilon\boldsymbol{v}) - f(\boldsymbol{x}))\boldsymbol{v}\right] \\ = \quad & \int\limits_{\boldsymbol{v}\in S(n)\setminus S_\delta^{\boldsymbol{x}}} \left[\operatorname{sign}(f(\boldsymbol{x}+\epsilon\boldsymbol{v}) - f(\boldsymbol{x}))\boldsymbol{v}\right] dS(n) + \int\limits_{\boldsymbol{v}\in S_\delta^{\boldsymbol{x}}} \left[\operatorname{sign}(f(\boldsymbol{x}+\epsilon\boldsymbol{v}) - f(\boldsymbol{x}))\boldsymbol{v}\right] dS(n) \\ = \quad & \int\limits_{\boldsymbol{v}\in S(n)\setminus S_\delta^{\boldsymbol{x}}} \left[\operatorname{sign}(\nabla f(\boldsymbol{x})^\top \boldsymbol{v})\boldsymbol{v}\right] dS(n) + \int\limits_{\boldsymbol{v}\in S_\delta^{\boldsymbol{x}}} \left[\operatorname{sign}(f(\boldsymbol{x}+\epsilon\boldsymbol{v}) - f(\boldsymbol{x}))\boldsymbol{v}\right] dS(n) \\ = \quad & \int\limits_{\boldsymbol{v}\in S(n)} \left[\operatorname{sign}(\nabla f(\boldsymbol{x})^\top \boldsymbol{v})\boldsymbol{v}\right] dS(n) - \int\limits_{\boldsymbol{v}\in S_\delta^{\boldsymbol{x}}} \left[\operatorname{sign}(\nabla f(\boldsymbol{x})^\top \boldsymbol{v})\boldsymbol{v}\right] dS(n) \\ & + \int\limits_{\boldsymbol{v}\in S_\delta^{\boldsymbol{x}}} \left[\operatorname{sign}(f(\boldsymbol{x}+\epsilon\boldsymbol{v}) - f(\boldsymbol{x}))\boldsymbol{v}\right] dS(n) \\ = \quad & P_n \frac{\nabla f(\boldsymbol{x})}{\|\nabla f(\boldsymbol{x})\|} - \int\limits_{\boldsymbol{v}\in S_\delta^{\boldsymbol{x}}} \left[\operatorname{sign}(\nabla f(\boldsymbol{x})^\top \boldsymbol{v})\boldsymbol{v}\right] dS(n) + \int\limits_{\boldsymbol{v}\in S_\delta^{\boldsymbol{x}}} \left[\operatorname{sign}(f(\boldsymbol{x}+\epsilon\boldsymbol{v}) - f(\boldsymbol{x}))\boldsymbol{v}\right] dS(n), \end{aligned}$$

where the last equality is due to Lemma 5.3.2. Putting the estimations together, we

Algorithm 3: Bandit Online Normalized Gradient Descent

Input: feasible set $\mathcal{X}$, # time period T, δ
Initialization: $\boldsymbol{x}_1 \in \mathcal{X}$, $\epsilon = \delta^2/L$
for $t = 1$ **to** T **do**
  Sample $\boldsymbol{v}_t$ uniformly over $S(n) \subset \mathbb{R}^n$;
  play $\boldsymbol{x}_t$ and $\boldsymbol{x}_t + \epsilon \boldsymbol{v}_t$;
  receive feedbacks $f_t(\boldsymbol{x}_t)$ and $f_t(\boldsymbol{x}_t + \epsilon \boldsymbol{v}_t)$;
  set $G_t(\boldsymbol{x}_t, \boldsymbol{v}_t) = \frac{\text{sign}(f_t(\boldsymbol{x}_t+\epsilon \boldsymbol{v}_t) - f_t(\boldsymbol{x}_t))}{Q_n} \boldsymbol{v}_t$;
  update $\boldsymbol{x}_{t+1} = \prod_{\mathcal{X}} (\boldsymbol{x}_t - \eta G_t(\boldsymbol{x}_t, \boldsymbol{v}_t))$.
end for

have

$$\begin{aligned} & \left\| \mathsf{E}_{S(n)} \left[\text{sign}(f(\boldsymbol{x} + \epsilon \boldsymbol{v}) - f(\boldsymbol{x})) \boldsymbol{v} \right] - \frac{P_n}{\beta_n} \frac{\nabla f(\boldsymbol{x})}{\|\nabla f(\boldsymbol{x})\|} \right\| \\ \leq & \frac{1}{\beta_n} \int\limits_{\boldsymbol{v} \in S_\delta^{\boldsymbol{x}}} \left\| \text{sign}(\nabla f(\boldsymbol{x})^\top \boldsymbol{v}) \boldsymbol{v} \right\| dS(n) + \frac{1}{\beta_n} \int\limits_{\boldsymbol{v} \in S_\delta^{\boldsymbol{x}}} \left\| \text{sign}(f(\boldsymbol{x} + \epsilon \boldsymbol{v}) - f(\boldsymbol{x})) \boldsymbol{v} \right\| dS(n) \\ \leq & \frac{2m(S_\delta^{\boldsymbol{x}})}{\beta_n} \leq \frac{2C_n \delta}{\beta_n}. \end{aligned}$$

Note that $Q_n = \frac{P_n}{\beta_n}$ and $D_n = \frac{C_n}{\beta_n}$, the theorem is proved. □

Based on Theorem 5.3.3, for a given $\delta > 0$ we have a zeroth-order estimator for the normalized gradient given as:

$$G_t(\boldsymbol{x}_t, \boldsymbol{v}_t) = \frac{\text{sign}(f_t(\boldsymbol{x}_t + \epsilon \boldsymbol{v}_t) - f_t(\boldsymbol{x}_t))}{Q_n} \boldsymbol{v}_t, \tag{5.9}$$

where $\epsilon = \delta^2/L$ and $\boldsymbol{v}_t$ is an uniformly distributed random vector over $S(n)$. Theorem 5.3.3 implies that the distance between the estimator and the normalized gradient can be controlled up to a factor of δ. Essentially, the Bandit Online Normalized Gradient Descent (BONGD) algorithm replaces the normalized gradient by $G_t(\boldsymbol{x}_t, \boldsymbol{v}_t)$ in the ONGD algorithm.

Note that Algorithm 3 actually outputs a random sequence of vectors $\{\boldsymbol{x}_t\}_1^T$; hence the notion of expected non-stationary regret is applicable here. Let us denote $\{\mathcal{F}_t\}_1^T$ be the filtration generated by $\{\boldsymbol{x}_t\}_1^T$. Then $\boldsymbol{v}_t$ is independent of $\mathcal{F}_t$. Note that in Algorithm 3, at each step, it queries the function at another point $\boldsymbol{x}_t + \epsilon \boldsymbol{v}_t$. Therefore, besides its

output sequence $\{\boldsymbol{x}_t\}_1^T$, we need to include $\{\boldsymbol{x}_t + \epsilon \boldsymbol{v}_t\}_1^T$ in our regret. We thus define

$$\begin{aligned} & \text{ERegret}_T^{NS}(\{\boldsymbol{x}_t\}_1^T, \{\boldsymbol{x}_t + \epsilon \boldsymbol{v}_t\}_1^T) \\ = \; & \mathsf{E}\left[\sum_{t=1}^{T} (f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_t^*))\right] + \mathsf{E}\left[\sum_{t=1}^{T} (f_t(\boldsymbol{x}_t + \epsilon \boldsymbol{v}_t) - f_t(\boldsymbol{x}_t^*))\right]. \end{aligned}$$

The following theorem shows that by choosing η and δ appropriately, we can still achieve an $O(\sqrt{T + V_T T})$ expected non-stationary regret bound.

Theorem 5.3.4 *Let V_T be defined in* (5.6)*. Assume that the loss functions have Lipschitz gradients (Definition 19), satisfying the error bound condition (Definition 18) and are weakly pseudo-convex with bounded gradient. For any sequence of loss functions $\{f_t\}_1^T \in \mathcal{S}_T$, applying BONGD with $\eta = Q_n \sqrt{\frac{4R^2 + 6RV_T}{T}}$ and $\delta = \min\{T^{-\frac{1}{2\gamma}}, T^{-\frac{1}{4}}\}$ where $Q_n = \frac{P_n}{\beta_n}$ and P_n is a constant, the following regret bound holds*

$$\text{ERegret}_T^{NS}(\{\boldsymbol{x}_t\}_1^T, \{\boldsymbol{x}_t + \epsilon \boldsymbol{v}_t\}_1^T) \leq O(\sqrt{T + V_T T}).$$

Proof: Let $z_t := \|\boldsymbol{x}_t - \boldsymbol{x}_t^*\|$. Then,

$$\begin{aligned} z_{t+1}^2 = \; & \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t+1}^*\|^2 \\ = \; & \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t^*\|^2 + \|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\|^2 + 2(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t^*)^\top (\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*) \\ \leq \; & \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t^*\|^2 + 2R\|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\| + 4R\|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\| \\ = \; & \|\textstyle\prod_{\mathcal{X}} (\boldsymbol{x}_t - \eta G_t(\boldsymbol{x}_t, \boldsymbol{v}_t)) - \boldsymbol{x}_t^*\|^2 + 6R\|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\| \\ \leq \; & \|\boldsymbol{x}_t - \eta G_t(\boldsymbol{x}_t, \boldsymbol{v}_t) - \boldsymbol{x}_t^*\|^2 + 6R\|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\| \\ = \; & z_t^2 + \eta^2 \|G_t(\boldsymbol{x}_t, \boldsymbol{v}_t)\|^2 - 2\eta G_t(\boldsymbol{x}_t, \boldsymbol{v}_t)^\top (\boldsymbol{x}_t - \boldsymbol{x}_t^*) + 6R\|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\| \\ \leq \; & z_t^2 + \tfrac{\eta^2}{Q_n^2} - 2\eta G_t(\boldsymbol{x}_t, \boldsymbol{v}_t)^\top (\boldsymbol{x}_t - \boldsymbol{x}_t^*) + 6R\|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\|. \end{aligned}$$

By rearranging the terms, we have:

$$KG_t(\boldsymbol{x}_t, \boldsymbol{v}_t)^\top (\boldsymbol{x}_t - \boldsymbol{x}_t^*) \;\leq\; \frac{K}{2\eta}\left(z_t^2 - z_{t+1}^2 + \frac{\eta^2}{Q_n^2} + 6R\|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\|\right).$$

Now based on $\|\nabla f_t(\boldsymbol{x}_t)\|$, we have two different cases:

- $\|\nabla f_t(\boldsymbol{x}_t)\| \geq \delta$. In this case, by Theorem 5.3.3, we have

$$\|\mathsf{E}\left[G_t(\boldsymbol{x}_t, \boldsymbol{v}_t)|\boldsymbol{x}_t\right] - \frac{\nabla f_t(\boldsymbol{x}_t)}{\|\nabla f_t(\boldsymbol{x}_t)\|}\| \leq \frac{2D_n}{Q_n}\delta.$$

Therefore,

$$\begin{aligned}
& f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_t^*) \\
\leq\ & K\frac{\nabla f_t(\boldsymbol{x}_t)^\top(\boldsymbol{x}_t - \boldsymbol{x}_t^*)}{\|\nabla f_t(\boldsymbol{x}_t)\|} \\
\leq\ & K\mathsf{E}\left[G_t(\boldsymbol{x}_t, \boldsymbol{v}_t)|\boldsymbol{x}_t\right]^\top(\boldsymbol{x}_t - \boldsymbol{x}_t^*) + \frac{2D_nK}{Q_n}\delta\|\boldsymbol{x}_t - \boldsymbol{x}_t^*\| \\
=\ & \frac{K}{2\eta}\left(\mathsf{E}\left[z_t^2|\boldsymbol{x}_t\right] - \mathsf{E}\left[z_{t+1}^2|\boldsymbol{x}_t\right] + \frac{\eta^2}{Q_n^2} + 6R\|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\|\right) + \frac{2D_nK}{Q_n}\delta\|\boldsymbol{x}_t - \boldsymbol{x}_t^*\| \\
\leq\ & \frac{K}{2\eta}\left(\mathsf{E}\left[z_t^2|\boldsymbol{x}_t\right] - \mathsf{E}\left[z_{t+1}^2|\boldsymbol{x}_t\right] + \frac{\eta^2}{Q_n^2} + 6R\|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\|\right) + \frac{4D_nK}{Q_n}R\delta. \qquad (5.10)
\end{aligned}$$

- $\|\nabla f_t(\boldsymbol{x}_t)\| < \delta$. In this case, by the error bound property (Definition 18) we have

$$\|\boldsymbol{x}_t - \boldsymbol{x}_t^*\| \leq D\|\nabla f_t(\boldsymbol{x}_t)\|^\gamma < D\delta^\gamma.$$

Therefore, due to the boundedness of gradient

$$f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_t^*) \leq M\|\boldsymbol{x}_t - \boldsymbol{x}_t^*\| \leq MD\delta^\gamma, \qquad (5.11)$$

and

$$\begin{aligned}
0 \ \leq\ & \frac{K}{2\eta}\left(\mathsf{E}\left[z_t^2|\boldsymbol{x}_t\right] - \mathsf{E}\left[z_{t+1}^2|\boldsymbol{x}_t\right] + \frac{\eta^2}{Q_n^2} + 6R\|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\|\right) - K\mathsf{E}\left[G_t(\boldsymbol{x}_t, \boldsymbol{v}_t)|\boldsymbol{x}_t\right]^\top(\boldsymbol{x}_t - \boldsymbol{x}_t^*) \\
\leq\ & \frac{K}{2\eta}\left(\mathsf{E}\left[z_t^2|\boldsymbol{x}_t\right] - \mathsf{E}\left[z_{t+1}^2|\boldsymbol{x}_t\right] + \frac{\eta^2}{Q_n^2} + 6R\|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\|\right) + K\frac{\beta_n}{Q_n}D\delta^\gamma. \qquad (5.12)
\end{aligned}$$

Adding (5.11) with (5.12), it follows that

$$\begin{aligned}
& f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_t^*) \\
\leq\ & \frac{K}{2\eta}\left(\mathsf{E}\left[z_t^2|\boldsymbol{x}_t\right] - \mathsf{E}\left[z_{t+1}^2|\boldsymbol{x}_t\right] + \frac{\eta^2}{Q_n^2} + 6R\|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\|\right) + \left(K\frac{\beta_n}{Q_n}D + MD\right)\delta^\gamma. \qquad (5.13)
\end{aligned}$$

In view of (5.10) and (5.13), if we let $U = \max\left\{\frac{4C_nK}{P_n}R, (K\frac{\beta_n}{Q_n}D + MD)\right\}$, then in

either case the following inequality holds:

$$f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_t^*) \quad \leq \quad \frac{K}{2\eta}\left(\mathsf{E}\left[z_t^2|\boldsymbol{x}_t\right] - \mathsf{E}\left[z_{t+1}^2|\boldsymbol{x}_t\right] + \frac{\eta^2}{Q_n^2} + 6R\|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\|\right) + U\delta^\gamma.$$

Summing these inequalities over $t = 1, ..., T$, we have

$$\begin{aligned}
& \mathrm{ERegret}_T^{NS}(\{\boldsymbol{x}_t\}_1^T, \{\boldsymbol{x}_t + \epsilon\boldsymbol{v}_t\}_1^T) \\
= \quad & \mathsf{E}\left[\sum_{t=1}^{T}\left(f_t(\boldsymbol{x}_t) + f_t(\boldsymbol{x}_t + \epsilon\boldsymbol{v}_t) - 2f_t(\boldsymbol{x}_t^*)\right)\right] \\
\leq \quad & \mathsf{E}\left[\sum_{t=1}^{T}\left(2f_t(\boldsymbol{x}_t) - 2f_t(\boldsymbol{x}_t^*) + M\epsilon\|\boldsymbol{v}_t\|\right)\right] \\
\leq \quad & \frac{K}{\eta}\left(\mathsf{E}\left[z_1^2\right] - \mathsf{E}\left[z_{T+1}^2\right] + T\frac{\eta^2}{Q_n^2} + 6R\sum_{t=1}^{T}\|\boldsymbol{x}_t^* - \boldsymbol{x}_{t+1}^*\|\right) + 2TU\delta^\gamma + MT\epsilon \\
\leq \quad & \frac{K}{2\eta}\left(4R^2 + T\frac{\eta^2}{Q_n^2} + 6RV_T\right) + 2TU\delta^\gamma + TM\frac{\delta^2}{L}.
\end{aligned}$$

By choosing $\eta = Q_n\sqrt{\frac{4R^2+6RV_T}{T}}$, and $\delta = \min\{T^{-\frac{1}{2\gamma}}, T^{-\frac{1}{4}}\}$, we have

$$\begin{aligned}
& \mathrm{ERegret}_T^{NS}(\{\boldsymbol{x}_t\}_1^T, \{\boldsymbol{x}_t + \epsilon\boldsymbol{v}_t\}_1^T) \\
\leq \quad & \frac{2K}{Q_n}\sqrt{T(4R^2 + 6RV_T)} + (2U + \frac{M}{L})\sqrt{T} \\
\leq \quad & O(\sqrt{T + V_T T}).
\end{aligned}$$

□

Therefore, under the additional error bound condition (Definition 18) and the Lipschitz continuity of the gradient (Definition 19) on the loss functions (e.g. the function depicted in Figure 1 satisfies all the conditions of Theorem 3), the expected regret of BONGD remains $O(\sqrt{T + V_T T})$, which matches both the upper and lower bound in Yang et al. (2016) for the general Lipschitz continuous convex cost functions and two point bandit feedback. Moreover, the zeroth-order estimator for the normalized gradient could be of interest on its own.

5.4 Concluding Remarks

In this chapter, we considered online learning with non-convex loss functions and the non-stationary regret measure, and established $O(\sqrt{T + V_T T})$ regret bounds, where V_T is the total variation of the loss functions, for a gradient-type algorithm and a bandit-type algorithm under some conditions on the non-convex loss function. As a direction for future research, it will be interesting to find out if the same regret bound can still be established without knowing V_T in advance. Moreover, it remains open to extend the results to the setting where the loss functions may be noisy and non-smooth.

5.5 Technical Proofs

Before we prove Proposition 5.1.1, let us recall the definition of star-convexity and show a lemma.

Definition 20 *(Star-convex functions). A function $f : \mathbb{R}^n \to \mathbb{R}$ is star-convex if there is $x^* \in \operatorname*{argmin}_{x \in \mathcal{X}} f(x)$ such that for all $\alpha \in [0, 1]$ and $x \in \mathcal{X}$,*

$$f((1-\alpha)x^* + \alpha x) \leq (1-\alpha)f(x^*) + \alpha f(x). \tag{5.14}$$

The following lemma characterizes the differentiable star-convex functions.

Lemma 5.5.1 *For a differentiable function f, the star convexity condition* (5.14) *is equivalent to the following condition*

$$f(x) - f(x^*) \leq \nabla f(x)^\top (x - x^*), \tag{5.15}$$

where $x^* = \operatorname*{argmin}_{x \in \mathcal{X}} f(x)$.

Proof: Suppose (5.14) holds. Then we have

$$f(x) - f(x^*) \leq \frac{f(x) - f((1-\alpha)x^* + \alpha x)}{1-\alpha}, \tag{5.16}$$

for all $\alpha \in [0, 1]$. Note that

$$\lim_{\alpha \to 1^-} \frac{f(x) - f((1-\alpha)x^* + \alpha x)}{1-\alpha} = \nabla f(x)^\top (x - x^*),$$

which implies (5.15). Conversely, suppose that (5.15) holds. Let us denote

$$d(\alpha) := f((1-\alpha)x^* + \alpha x) - f(x^*).$$

Clearly, (5.14) is equivalent to

$$d(\alpha) \leq \alpha d(1), \text{ for all } 0 \leq \alpha \leq 1. \tag{5.17}$$

It remains to show that if f is differentiable then (5.15) implies (5.17). In fact, (5.15) leads to

$$f((1-\alpha)x^* + \alpha x) - f(x^*) \leq \alpha \nabla f((1-\alpha)x^* + \alpha x)^\top (x - x^*),$$

or,

$$d(\alpha) \leq \alpha d'(\alpha).$$

Hence,

$$\left(\frac{d(\alpha)}{\alpha}\right)' = \frac{\alpha d'(\alpha) - d(\alpha)}{\alpha^2} \geq 0,$$

for all $0 < \alpha \leq 1$, implying that $\frac{d(\alpha)}{\alpha}$ is a nondecreasing function for $\alpha \in (0, 1]$. Therefore,

$$\frac{d(\alpha)}{\alpha} \leq \frac{d(1)}{1},$$

which proves (5.17) for $\alpha \in (0, 1]$. Since $d(0) = f(x^*) = 0$, (5.17) in fact holds for all $\alpha \in [0, 1]$. □

Proof of Proposition 5.1.1 From Lemma 5.5.1, we have

$$\begin{aligned} f(x) - f(x^*) &\leq \nabla f(x)^\top \|x - x^*\| \\ &\leq M \frac{\nabla f(x)^\top (x - x^*)}{\|\nabla f(x)\|} \end{aligned}$$

where the last inequality is due to the bounded gradient condition $\|\nabla f(x)\| \leq M$ for $x \in \mathcal{X}$. □

Proof of Proposition 5.1.2: For all $x \in \mathcal{X}$, we have

$$\begin{aligned} f(x) - f(x^*) &\leq M\|x - x^*\| \\ &\leq \frac{M}{Z}\frac{\nabla f(x)^\top (x - x^*)}{\|\nabla f(x)\|} \end{aligned}$$

where the first inequality follows from the bounded gradient assumption while the second inequality is due to the acute angle condition. □

Proof of Proposition 5.1.3: By taking the derivative of the equation (5.5) with respective to t and letting $t = 1$, we have

$$\nabla f(x)^\top (x - x^*) = \alpha(f(x) - f(x^*)).$$

Therefore, we have

$$\begin{aligned} f(x) - f(x^*) &= \frac{1}{\alpha}\nabla f(x)^\top (x - x^*) \\ &\leq \frac{M}{\alpha}\frac{\nabla f(x)^\top (x - x^*)}{\|\nabla f(x)\|}, \end{aligned}$$

which satisfies the weak pseudo-convexity condition with $K = \frac{M}{\alpha}$. □

Chapter 6

Conclusion

In this dissertation, we studied the application of the convex optimization in the discrete choice modeling. In Chapter 3, we showed that the CMM model, which had already been reformulated as a semi-definite program by Mishra et al. (2012), can be reformulated as maximizing a strongly concave function over a unit simplex. Several desirable properties of the objective function allowed us to prove that the well-known projected gradient algorithm enjoys a local linear convergence rate. In Chapter 4, we proposed the welfare-based approach to study the choice models. The welfare-based approach starts with any convex, monotone and translation invariant function as a potential function, and then defines the choice probability as the gradient of the potential function. We establish the equivalence relation among several known choice models and the proposed welfare-based choice model. The proof of the equivalence relies on the machineries from convex analysis. The substitutability/complementarity concept in the discrete choice model is essentially the submodularity/supermodularity of the welfare function. In Chapter 5, we studied the online learning problem, where loss functions are only assumed to be weakly pseudo-convex. When the gradient of the loss function at the decision point was available, we proposed an online normalized gradient descent algorithm to solve the online learning problem. In another situation, when only the value of the loss function was available, we proposed a bandit online normalized gradient descent algorithm. With the help of continuous optimization theories, we showed that both algorithms achieve a cumulative regret bound of $O(\sqrt{T + V_T T})$, where V_T is the total temporal variations of loss functions.

References

J. Abernethy, A. Agarwal, P. L. Bartlett, and A. Rakhlin. A stochastic view of optimal regret through minimax duality. *arXiv preprint arXiv:0903.5328*, 2009.

A. Agarwal, O. Dekel, and L. Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *COLT*, pages 28–40. Citeseer, 2010.

S. D. Ahipaşaoğlu, R. Meskarian, T. L. Magnanti, and K. Natarajan. Beyond normality: A cross moment-stochastic user equilibrium model. *Transportation Research Part B: Methodological*, 81:333–354, 2015.

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

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

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

M. Ben-Akiva. *Structure of passenger travel demand models.* PhD thesis, Massachusetts Institute of Technology, 1973.

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

A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, volume 2. Siam, 2001.

S. T. Berry. Estimating discrete-choice models of product differentiation. *RAND Journal of Economics*, 25(2):242–262, 1994.

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

D. P. Bertsekas. Nonlinear programming. 1999.

O. Besbes, Y. Gur, and A. Zeevi. Non-stationary stochastic optimization. *Operations Research*, 63(5):1227–1244, 2015.

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

A. Börsch-Supan. On the compatibility of nested logit models with utility maximization. *Journal of Econometrics*, 43(3):373–388, 1990.

A. Börsch-Supan and V. A. Hajivassiliou. Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models. *Journal of Econometrics*, 58(3):347–368, 1993.

J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer Science & Business Media, 2010.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

D. Brownstone, D. S. Bunch, and K. Train. Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles. *Transportation Research B*, 34(5): 315–338, 2000.

J. Bunch, C. Nilsen, and D. Sorensen. Rank-one modification of the symmetric eigenproblem. *Numerische Mathematik*, 31:31–48, 1978.

Q. G. Chen, S. Jasin, and I. Duenyas. Adaptive parametric and nonparametric multiproduct pricing via self-adjusting controls. 2014.

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

C. F. Daganzo and M. Kusnic. Technical notetwo properties of the nested logit model. *Transportation Science*, 27(4):395–400, 1993.

A. Daly and S. Zachary. Improved multiple choice models. In *Determinants of Travel Choice*, pages 335–357. Teakfields, 1978.

J. M. Davis, G. Gallego, and H. Topaloglu. Assortment optimization under variants of the nested logit model. *Operations Research*, 62(2):250–273, 2014.

A. V. den Boer. Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in operations research and management science*, 20(1):1–18, 2015.

L. Dong, P. Kouvelis, and Z. Tian. Dynamic pricing and inventory control of substitute products. *Manufacturing & Service Operations Management*, 11(2):317–339, 2009.

D. C. Dowson and B. V. Landau. The fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12:450–455, 1982.

M. D. Ekstrand, J. T. Riedl, and J. A. Konstan. Collaborative filtering recommender systems. *Foundations and Trends in Human-Computer Interaction*, 4(2):81–173, 2010.

S. Ertekin, L. Bottou, and C. L. Giles. Nonconvex online support vector machines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):368–381, Feb 2011. ISSN 0162-8828. doi: 10.1109/TPAMI.2010.109.

T. Evgeniou, M. Pontil, and O. Toubia. A convex optimization approach to modeling consumer heterogeneity in conjoint estimation. *Marketing Science*, 26(6):805–818, 2007.

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

A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394. Society for Industrial and Applied Mathematics, 2005.

D. Fudenberg, R. Iijima, and T. Strzalecki. Stochastic choice and revealed perturbed utility. *Econometrica*, 83(6):2371–2409, 2015.

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

G. Gasso, A. Pappaioannou, M. Spivak, and L. Bottou. Batch and online learning algorithms for nonconvex neyman-pearson classification. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):28, 2011.

J. Geweke. Bayesian inference in econometric models using monte carlo integration. *Econometrica*, 57(6):1317–1339, 1989. ISSN 00129682. URL `http://www.jstor.org/stable/1913710`.

J. Geweke. *Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints*. Defense Technical Information Center, 1992. URL `http://books.google.com.sg/books?id=PrTSNwAACAAJ`.

G. H. Golub and C. F. V. Loan. *Matrix Computations*. John Hopkins University Press, 1996.

V. A. Hajivassiliou and D. L. McFadden. The method of simulated scores for the estimation of ldv models. *Econometrica*, 66(4):863–896, 1998. ISSN 00129682. URL `http://www.jstor.org/stable/2999576`.

V. A. Hajivassiliou, D. L. McFadden, and P. Ruud. Simulation of multivariate normal rectangle probabilities and their derivatives: Theoretical and computational results. *Journal of Econometrics*, 72:85–134, 1996.

W. Hanson and K. Martin. Optimizing multinomial logit profit functions. *Management Science*, 42(7):992–1003, 1996.

E. Hazan and S. Kale. Online submodular minimization. *The Journal of Machine Learning Research*, 13(1):2903–2922, 2012.

E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

E. Hazan, K. Levy, and S. Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. In *Advances in Neural Information Processing Systems*, pages 1594–1602, 2015.

E. Heinz. Beiträge zur Störungstheorie der Spektralzerlegung. *Math. Ann.*, 123:415–438, 1951. ISSN 0025-5831. URL `MR0044747.pdf`.

J. A. Herriges and C. L. Kling. Testing the consistency of nested logit models with utility maximization. *Economics Letters*, 50(1):33–39, 1996.

N. J. Higham. *Functions of matrices: theory and computation*, volume 104. Siam, 2008.

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

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

A. Iusem. On the convergence properties of the projected gradient method for convex optimization. *Computational & Applied Mathematics*, 22(1):37–52, 2003.

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

M. P. Keane. *Four Essays in Empirical Macro and Labor Economics*. Brown University, 1990. URL `http://books.google.com.sg/books?id=C0EjHQAACAAJ`.

M. P. Keane. A computationally practical simulation estimator for panel data. *Econometrica*, 62(1):95–116, January 1994. URL `http://ideas.repec.org/a/ecm/emetrp/v62y1994i1p95-116.html`.

C. Kenney and A. J. Laub. Condition estimates for matrix functions. *SIAM Journal on Matrix Analysis and Applications*, 10(2):191 – 209, 1989.

C. L. Kling and J. A. Herriges. An empirical investigation of the consistency of nested logit models with utility maximization. *American Journal of Agricultural Economics*, 77(4):875–884, 1995.

B. Lee. Calling patterns and usage of residential toll service under self selecting tariffs. *Journal of Regulatory Economics*, 16(1):45–82, 1999.

H. Li and W. T. Huh. Pricing multiple products with the multinomial logit and nested logit models: Concavity and implications. *Manufacturing & Service Operations Management*, 13(4):549–563, 2011.

K. Löwner. Über monotone matrixfunktionen. *Mathematische Zeitschrift*, 38(1):177–216, 1934. ISSN 0025-5874. doi: 10.1007/BF01170633. URL `http://dx.doi.org/10.1007/BF01170633`.

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

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

D. McFadden. Quantitative methods for analyzing travel behaviour of individuals: Some recent developments. Technical report, Cowles Foundation for Research in Economics, Yale University, 1977.

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

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

D. McFadden and K. Train. Mixed MNL models for discrete responses. *Journal of Applied Econometrics*, 15(5):447–470, 2000.

D. McFadden et al. Conditional logit analysis of qualitative choice behavior. 1973.

S. McShane and M. Von Glinow. *Organizational Behavior*. McGraw-Hill Higher Education, 2015.

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

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

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

K. Natarajan and C.-P. Teo. On reduced semidefinite programs for second order moment bounds with applications. *Mathematical Programming*, 161(1-2):487–518, 2017.

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

Y. Nesterov. *Introductory lectures on convex optimization*, volume 87. Springer Science & Business Media, 2004.

Y. Nesterov and B. T. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

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

I. Olkin and F. Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48(0):257 – 263, 1982. ISSN 0024-3795. doi: http://dx.doi.org/10.1016/0024-3795(82)90112-4. URL `http://www.sciencedirect.com/science/article/pii/0024379582901124`.

G. Parsons and M. Kealy. Randomly drawn opportunity sets in a random utility model of lake recreation. *Land Economics*, 68(1):93–206, 1992.

R. T. Rockafellar. *Convex Analysis*, volume 28. 1970.

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

P. Rosenzweig. *The Halo Effect and the Eight Other Business Delusions That Deceive Managers*. Simon and Schuster, 2014.

A. Saha and A. Tewari. Improved regret guarantees for online smooth convex optimization with bandit feedback. In *International Conference on Artificial Intelligence and Statistics*, pages 636–642, 2011.

A. Shapiro. Extremal problems on the set of nonnegative definite matrices. *Linear Algebra and its Applications*, 67:7–18, 1985.

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

J.-S. Song and Z. Xue. Demand management and inventory control for substitutable products. *Working paper*, 2007.

K. Talluri and G. Van Ryzin. Revenue management under a general discrete choice model of consumer behavior. *Management Science*, 50(1):15–33, 2004.

E. L. Thorndike. A constant error in psychological ratings. *Journal of Applied Psychology*, 4(1):25–29, 1920.

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

P. Tiwari and H. Hasegawa. Demand for housing in Tokyo: A discrete choice analysis. *Regional Studies*, 38(1):27–42, 2004.

O. Toubia, D. I. Simester, J. R. Hauser, and E. Dahan. Fast polyhedral adaptive conjoint estimation. *Marketing Science*, 22(3):273–303, 2003.

K. Train, D. McFadden, and M. Ben-Akiva. The demand for local telephone service: A fully discrete model of residential calling patterns and service choices. *RAND Journal of Economics*, 18(1):109–123, 1987.

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

F. Verboven. The nested logit model and representative consumer theory. *Economics Letters*, 50(1):57–63, 1996.

J. H. Wilkinson. *The algebraic eigenvalue problem*, volume 87. Clarendon Press Oxford, 1965.

H. C. Williams. On the formation of travel demand models and economic evaluation measures of user benefit. *Environment and planning A*, 9(3):285–344, 1977.

L. Yang, D. Sun, and K.-C. Toh. Sdpnal + : a majorized semismooth newton-cg augmented lagrangian method for semidefinite programming with nonnegative constraints. *Mathematical Programming Computation*, 7(3):331–366, 2015.

T. Yang, L. Zhang, R. Jin, and J. Yi. Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. In *International Conference on Machine Learning*, pages 449–457, 2016.

J. Yates and D. F. Mackay. Discrete choice modelling of urban housing markets: A critical review and an application. *Urban Studies*, 43(3):559–581, 2006.

L. Zhang, T. Yang, R. Jin, and Z.-H. Zhou. Online bandit learning for a special class of non-convex losses. In *AAAI*, pages 3158–3164, 2015.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. 2003.