Robustness of Consistent Loss Functions for Multinomial Outcome Models

by

Suchan Vivatsethachai

B.S., Massachusetts Institute of Technology (2020)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author .....
Department of Electrical Engineering and Computer Science
May 14, 2021

Certified by .....
Daniel Pollmann
Vice President, QuantCo, Inc.
Thesis Supervisor

Certified by .....
Constantinos Daskalakis
Professor of Computer Science
Thesis Supervisor

Accepted by .....
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Robustness of Consistent Loss Functions for Multinomial Outcome Models

by

Suchan Vivatsethachai

Submitted to the Department of Electrical Engineering and Computer Science
on May 14, 2021, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Maximum likelihood estimation, which uses the logarithmic loss function, is the default method used to estimate latent parameters consistently in multinomial outcome models. However, it is sensitive to even a tiny fraction of corruption in the training data. Alternatively, other loss functions in the family of strictly consistent loss functions can be used to consistently estimate model parameters. In this thesis, we study the robustness properties of different loss functions in the family, mainly the logarithmic loss function, the quadratic loss function, and the spherical loss function. We introduce two notions of robustness properties of loss functions. A loss function is *partially robust* if its corresponding influence function, a proxy for the bias from corruption, has bounded 2-norm. On the other hand, a loss function is *strongly robust* if the 2-norm of the bias itself is bounded. When some mild assumptions are met, the quadratic loss function can be shown to be both partially robust and strongly robust, while the logarithmic loss function is not. We also demonstrate that the behaviors of each loss function agree with their theoretical properties when used to estimate parameter in two synthetic models: a price-purchase model and a multinomial logit with intercepts model for two products. This thesis thus not only advocates more use of the quadratic loss function in parameter estimation of multinomial outcome models but also serves as a framework to conduct future research of the cross section between the robustness of loss functions and the consistency of parameter estimation.

Thesis Supervisor: Daniel Pollmann
Title: Vice President, QuantCo, Inc.

Thesis Supervisor: Constantinos Daskalakis
Title: Professor of Computer Science

Acknowledgments

This Master's thesis would not have been possible without a great deal of support I have received from QuantCo, Professor Costis, all of my instructors and friends during my 5 years at MIT, and, most importantly, my parents.

First and foremost, I would like to express my gratitude to Professor Constantinos Daskalakis for advising me throughout the past 5 years, for agreeing to be my MEng thesis supervisor, and for providing countless valuable advice and feedback. I would also like to thank Daniel Pollmann for agreeing to join 6-A program as my supervisor, for assigning me this thesis topic, and for helping me polish results time after time.

I would like to next thank QuantCo and everyone there for making this journey possible. Gea Hyun Shin and Kosti Takala have been my mentors on numerous projects at QuantCo and have actively involved in the weekly update of this thesis. Knowledge and experience I have received from them is no less important than what I have gained from classes at MIT. I want to thank Andreas for introducing me to QuantCo and for being an amazing mentor for everything since day one. I also want to thank Liz, Ben, and Evan for sharing their valuable experience and making every team lunch and dinner very enjoyable. My participation in 6-A program would not have happened if you all hadn't made my summer internship absolutely enjoyable.

At MIT, I owe a lot to the professors, the instructors, and the TAs in all the classes I have taken, who have taught me many more things than I could have imagined in the past 5 years. I was especially fortunate to have Fredrik Johansson and Professor David Sontag as my supervisors for the entire SuperUROP program. The experience I gained helped build the foundation of how I think about statistical modeling. For all these reasons, I am grateful to everyone who has contributed to my learning in any way. Regarding the 6-A program, I also want to thank Kathleen M Sullivan for introducing me to this wonderful program and helping me through all the administrative documents.

In addition to MIT professors, instructors, and colleagues at QuantCo, I am also indebted to multiple friends in the Thai community and other friends I crossed path

with at MIT. I would like to first express my gratitude for Anupong Tangpeerachaikul for playing one of the biggest roles in this chapter of my life by teaching, guiding, and being with me for more times than I can count and constantly supporting me through all of my ups and downs. I sincerely want to thank Kritkorn Karntikoon, Korrawat Pruegsanusak, Cattalyya Nuengsigkapian and Nipun Pitimanaaree for sharing numerous classes, problem sets, final projects, meals, game nights, grocery trips, and travels with me. Half of my classes at MIT would not have been this enjoyable and my life would have been much more miserable if you had not been there. Special thanks to Kritkorn for introducing me to Loomis, MIT, and even QuantCo. It is an understatement to say that my life trajectory would have been very different without you. I want to thank Warittha and Sorawit for despite being in their Senior year constantly made me feel welcome at MIT and showed me how to enjoy life at MIT to the fullest extent. I am also very grateful for having Chanoot, Nonravee, Panchanok, Paween, and Chaiwat constantly visit and bring joy into my MIT life. Next, I want to thank all of my other friends at MIT including Winnie, Yinzhan, Celia, Gloria, Christina, Sophie, Hyo, and many more than I can list here for sharing many days and nights working on problem sets and class projects and for sharing meals with me on many occasions. Last but not least, I couldn't possibly thank Apisada Chulakadabba enough for being one of the most important people in my life who constantly support me for all these years. Thank you for sharing many adventures with me in the past years, for mentally supporting me everyday, and for staying with me throughout the pandemic and beyond.

Finally, I want to express my gratitude to my family members. I would like to first thank my two older sisters for influencing me, both in good and bad ways, since my childhood. Thank you also for helping mom and dad when I am not there. Most importantly, I'm eternally grateful to my dad and my mom for their constant support through all phases of my life. Thank you for painstakingly working throughout all these years to raise us your three children. I wouldn't have been here without your support and sacrifice.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivations

Maximum likelihood estimation (MLE), since Fisher's introduction in [Fisher, 1922, Fisher, 1925], has been the method of choice that is widely used to estimate the true latent parameters in multinomial outcome models. It has many desirable theoretical guarantees on consistency, efficiency, and minimal asymptotic variance, as summarized in [Norden, 1972]. However, despite these properties, MLE can be sensitive to even a tiny fraction of corruption or outliers in the training data, resulting in severely biased parameter estimates that are no longer consistent with the true parameters.

We first demonstrate the sensitivity of MLE to the corruption through a simple toy example. Consider a synthetic dataset of price-purchase history with training data $\{(x_i, y_i)\}_{i=1}^n$ where x_i is a price and y_i is a binary that indicates a purchase. Each price x_i is uniformly drawn from 10 to 200. Each label y_i is drawn from Bernoulli distribution with probability $p = \sigma(c^* - \alpha^* x)$ when σ is a sigmoid function, and $(\alpha^*, c^*) = (0.1, 5.0)$ are the true latent parameters we want to estimate. One can use maximum likelihood estimation to estimate parameter from the training data by simply finding α, c such that

$$\operatorname*{argmin}_{c,\alpha} \frac{1}{n} \sum_{i=1}^{n} -y_i \log(\sigma(c - \alpha x_i)) - (1 - y_i) \log(1 - \sigma(c - \alpha x_i)).$$

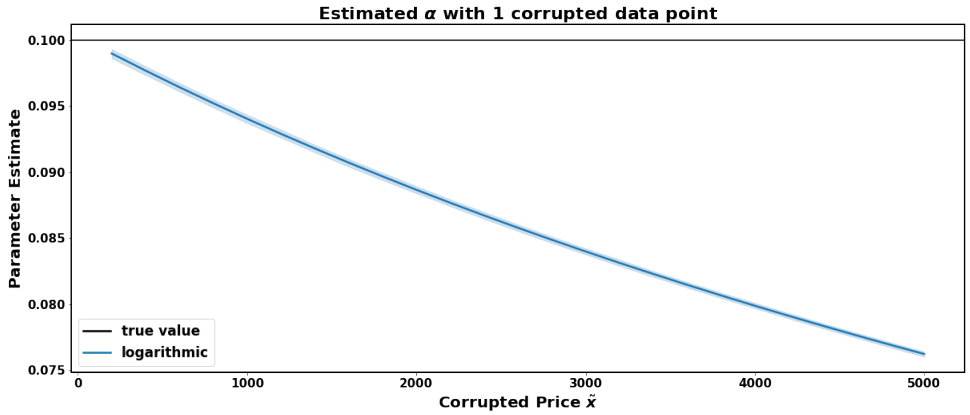

Figure 1-1: The parameter estimates of price sensitivity α from the maximum likelihood estimation. For each iteration, we generate 10000 data points from the data-generating process of price-purchase model and alter 1 data point to have price $\tilde{x}$ and label 1 when $\tilde{x}$ ranges from 200 to 5000. The process is repeated for 200 iterations and the mean parameter estimates are plotted with 95% bootstrapping confidence interval.

As the number of data points n increases, the MLE will produce parameter estimates that are converging to the true parameter $\alpha^* = 0.1$ and $c^* = 5.0$. However, when we corrupt just one data point with a faulty data with high price $\tilde{x}$, but is labeled a purchase ($y = 1$), the resulting parameter estimates can be unboundedly biased. In this example, we simply generate $n = 10000$ data points and change one data point to $x = \tilde{x}, y = 1$ as described. As shown in Figure 1-1, the greater the corrupted price $\tilde{x}$, the greater the bias in the parameter estimate α. We will revisit this example in detail together with other interesting synthetic datasets in Chapter 7.

This behavior can be problematic because data corruption comes in many forms. Training data can contains label noise or label poisoning [Frénay and Verleysen, 2013], contains data poisoning [Barreno et al., 2010], is misspecified [White, 1982], or suffers from adversarial attack [Qiu et al., 2019]. Any of these corruption can potentially lead to biased parameter estimates that can greatly affect subsequent optimization or inference tasks, especially from the industry perspective. A biased parameter estimate can misleadingly alter business decision because effects of some business factors

are overestimated or underestimated. Prices of products might be set too high if they are determined from an estimate of price sensitivity that is smaller than the truth.

When the consistency of the parameter estimate is the main concern, MLE is not the only option we can use. Any loss function in a family of strictly consistent loss functions can also be utilized to consistently estimate parameters through loss minimization method. Some loss functions, e.g., the quadratic loss function, as will be shown later in this thesis, are more robust than the logarithmic loss function, which is the equivalence of the MLE method. For that reason, we will study in detail the robustness properties of different loss functions in this family, both from the theoretical and empirical perspective. We hope that the results in this work can be used to solve data corruption problems by serving as a guideline for choosing robust loss functions in the family of strictly consistent loss functions in different scenarios.

1.2 Contributions

To the best of our knowledge, this thesis is the first to compare the robustness properties of loss functions in the family of strictly consistent loss functions for categorical labels. As a result, the framework we provide in this work can be used as a foundation to conduct research on the cross section between the robustness of loss functions and the consistency of the parameter estimates in multinomial outcome models. We connect these two fields into one unified framework and rigorously define the definitions of multinomial outcome models, the family of strictly consistent loss functions, (ϵ, G)-corruption, and necessary tools used in the analysis such as influence functions.

We then establish multiple novel theoretical results relating to the robustness properties of strictly consistent loss functions. We first define the notions of *partial robustness* and *strong robustness* properties. The first concerns the behavior of influence functions which are easier to analyze, while the second considers the bias from the corruption directly. We show that in both notions, the logarithmic loss function is not robust whereas the quadratic loss function is, thus encouraging more use of the quadratic loss function when corruption is expected. We also state necessary

assumptions and provide concrete theoretical bounds on both the influence functions and the bias of the parameter estimates from the quadratic loss function.

Lastly, we demonstrate the robustness behaviors of loss functions in synthetic datasets and confirm that the empirical behavior matches with the theoretical results. In the settings where the assumptions for the quadratic loss function hold, we show that the quadratic loss function is more robust than the logarithmic loss function. On the other hand, when the assumptions are violated, we demonstrate that the quadratic loss function can also fail.

1.3 Thesis Outline

This thesis is mainly divided into 4 parts: the background (Chapter 1 - 3), the theoretical results (Chapter 4 - 6), the empirical results (Chapter 7), and the conclusion (Chapter 8).

In chapter 2, we define the multinomial outcome models and the family of strictly consistent loss functions. We then show that a loss function in this family can be used to consistently estimate parameters of multinomial outcome models. Next, we list in Chapter 3 related works that serve as building blocks for the concepts used to build this thesis. In Chapter 4, we first define the main corruption framework we use for analysis, the (ϵ, G)-corruption, and then the crucial tool we use to analyze the robustness property, the influence functions. Afterward, we introduce the concepts of *partial robustness* and *strong robustness* and their properties in Chapter 5 and 6 respectively together with their corresponding theoretical results. Next, we show in Chapter 7 that the empirical robustness behaviors of loss functions in synthetic data agree with the theory in the previous two chapters. Lastly, we conclude our findings and discuss potential research directions in Chapter 8.

Chapter 2

Background

2.1 Multinomial Outcome Models

In this thesis, we will only consider a specific class of models: multinomial outcome models. These can be any model in which the label of the data is categorical, or multinomial, in nature. Specifically, throughout this work, we consider a model in which each data point consists of a pair of a covariate $X \in \mathcal{X}$ and a categorical label Y that is among one of the C possible categories. Each multinomial outcome model consists of 3 components:

1. A probability distribution F_X on $\mathcal{X}$ from which covariate X is drawn.

2. A probability function $\boldsymbol{f}(\cdot; \boldsymbol{\theta}) : \mathcal{X} \to \mathcal{P}_C$ which maps a covariate X to a finite discrete probability distribution of size C.

3. A model parameter $\boldsymbol{\theta}^* \in \Theta$.

Each data point is generated by first sampling a covariate X from the distribution F_X on $\mathcal{X}$, then, given the covariate X, a categorical label Y is sampled from the multinomial distribution with probability vector $\boldsymbol{f}(X; \boldsymbol{\theta}^*)$. In other words, (X, Y) is

sampled using the following data-generating process

$$\begin{aligned} X &\sim F_X, \\ Y|X &\sim Multi(1, \boldsymbol{f}(X; \boldsymbol{\theta}^*)). \end{aligned} \tag{2.1}$$

Through the observations of (X, Y), our goal is to reliably estimate the true latent parameters $\boldsymbol{\theta}^*$. For conciseness, we use Z to represent a pair of random variables (X, Y). We use notation $F = (F_X, \boldsymbol{f}, \boldsymbol{\theta}^*)$ to represent the model and the data-generating process of X and Y described earlier. Lastly, when it is clear that the model F consists of $(F_X, \boldsymbol{f}, \boldsymbol{\theta}^*)$, we use notation $Z \sim F$ or $(X, Y) \sim F$ to represent that data Z or (X, Y) are sampled from the process described in Eq. (2.1) with the covariate distribution F_X, the conditional label distribution $\boldsymbol{f}$, and the true parameter $\boldsymbol{\theta}^*$.

Most of the models with categorical labels fit right into our framework. Therefore, the notations and the results in this thesis can be applied to most of these models. In Example 1 and 2, we will demonstrate how the two most well-known multinomial outcome models, a logistic regression model and a multinomial logit model, fit into this framework.

Example 1. (Logistic Regression) A logistic regression model is one of the most basic multinomial outcome models. It uses a logistic function to maps a linear function of a covariate X to a probability of label Y. The number of labels in a logistic regression model is $C = 2$. The covariate $X \in \mathbb{R}^d$ is a d-dimensional real vector and the label $Y \in \{0, 1\}$ is a binary number. Given a covariate X, the probability of label $Y = 1$ is

$$\mathbb{P}(Y = 1 \mid X) = \frac{\exp\left(\beta_0 + \boldsymbol{\beta}^T X\right)}{\exp\left(\beta_0 + \boldsymbol{\beta}^T X\right) + 1}$$

when $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T$ is a parameter vector. Here the covariate probability distribution F_X can be any distribution on $\mathbb{R}^d$ space. The parameter of the model is

$\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_d)^T$. Lastly, the probability function of the label is

$$\boldsymbol{f}(X; \boldsymbol{\theta}) = \begin{bmatrix} \frac{\exp(\beta_0 + \boldsymbol{\beta} X)}{\exp(\beta_0 + \boldsymbol{\beta}^T X) + 1} \\ \frac{1}{\exp(\beta_0 + \boldsymbol{\beta}^T X) + 1} \end{bmatrix}$$

Example 2. (Multinomial Logit) A multinomial logistic regression model is a generalization of a logistic regression model to multiclass problem with C choices, or classes. Each observations contains a covariate $X = (X^{(0)}, X^{(1)}, \ldots, X^{(C-1)})$, and each component $X^{(i)} \in \mathbb{R}^d$ is a feature for class i for every $0 \leq i < C$. Given a covariate X, the probability of label $Y = y$ is

$$\mathbb{P}(Y = y \mid X) = \frac{\exp(\boldsymbol{\beta}^T X^{(y)})}{\sum_{j=0}^{C-1} \exp(\boldsymbol{\beta}^T X^{(j)})}.$$

Again, here, the covariate probability distribution F_X can be any distribution on $\mathbb{R}^{d \times C}$, the model parameter is $\boldsymbol{\theta} = \boldsymbol{\beta}$, and the probability function of the label is

$$\boldsymbol{f}(X; \boldsymbol{\theta}) = \begin{bmatrix} \frac{\exp(\boldsymbol{\beta}^T X^{(0)})}{\sum_{j=0}^{C-1} \exp(\boldsymbol{\beta}^T X^{(j)})} \\ \frac{\exp(\boldsymbol{\beta}^T X^{(1)})}{\sum_{j=0}^{C-1} \exp(\boldsymbol{\beta}^T X^{(j)})} \\ \vdots \\ \frac{\exp(\boldsymbol{\beta}^T X^{(C-1)})}{\sum_{j=0}^{C-1} \exp(\boldsymbol{\beta}^T X^{(j)})} \end{bmatrix}$$

2.2 Strictly Consistent Loss Functions

Our goal is to consistently estimate the model parameter $\boldsymbol{\theta}^*$ from data $Z \sim F$. One way to estimate the latent parameter $\boldsymbol{\theta}^*$ from data $Z = (X, Y)$ is to minimize the expected loss between the outcome Y and the probability vector $\boldsymbol{f}(X; \boldsymbol{\theta})$, which depends on the covariate X and the parameter $\boldsymbol{\theta}$ we want to estimate. We first introduce a type of loss function than is needed for this purpose in Definition 1. Furthermore, because we want to consistently estimate the true parameter, we need to use a specific class of these loss functions called the family of strictly consistent loss functions, which is formally defined in Definition 2. Lastly, we will show in Theorem 1 that any loss function in the family of strictly consistent loss functions can be used

in loss minimization method to consistently estimate the true parameter. Note that we mainly adopt the notations from [Gneiting and Raftery, 2007].

Definition 1. A loss function ℓ for probabilistic forecast of a categorical variable is a collection of C functions, each of which maps a probability vector of size C to a real number:

$$\forall\, 0 \leq y < C, \quad \ell(y, \cdot) : \mathcal{P}_C \to \mathbb{R}$$

Definition 2. A loss function ℓ is consistent if

$$\forall\, \boldsymbol{p}, \boldsymbol{q} \in \mathcal{P}_C, \quad \mathbb{E}_{Y \sim \boldsymbol{p}} \left[\ell(Y, \boldsymbol{p})\right] \leq \mathbb{E}_{Y \sim \boldsymbol{p}} \left[\ell(Y, \boldsymbol{q})\right],$$

and is strictly consistent if the equality happens only when $\boldsymbol{p} = \boldsymbol{q}$. Let $\mathcal{L}$ represent the family of such strictly consistent loss functions.

It turns out that the family of (strictly) consistent loss functions can be described in a closed-form formula. As stated in [Savage, 1971], a loss function for probability forecasts is (strictly) consistent if and only if it can be written as follows:

$$\forall\, 0 \leq y < C, \forall\, \boldsymbol{p} \in \mathcal{P}_C, \quad \ell(y, \boldsymbol{p}) = \langle G'(\boldsymbol{p}), \boldsymbol{p} \rangle - G(\boldsymbol{p}) - G'_y(\boldsymbol{p}), \tag{2.2}$$

where $G : \mathcal{P}_C \to \mathbb{R}$ is any (strictly) convex function, and $G'(\cdot)$ is the subgradient of G.

The logarithmic loss function and the quadratic loss function, the two well-known loss functions for parameter estimation in multinomial outcome models, are the members of the family of strictly consistent loss functions. A less well-known loss function, the spherical loss function, can also be derived from the closed-form formula of the family of strictly consistent loss functions. In this work, we mainly investigate these three loss functions and their robustness behaviors.

1. (logarithmic) With $G(\boldsymbol{p}) = \sum_{i=0}^{C-1} p_j \log(p_j)$, the logarithmic loss function is

$$\ell_{log}(y, \boldsymbol{p}) = -\log(p_y). \tag{2.3}$$

   Using the logarithmic loss function in loss minimization is equivalent to the maximum likelihood estimation method. The logarithmic loss function is thus the most widely-used loss function out of the three loss functions.

2. (quadratic) With $G(\boldsymbol{p}) = \sum_{i=0}^{C-1} p_i^2 - 1$, the quadratic loss function is

$$\begin{aligned}\ell_{quad}(y, \boldsymbol{p}) &= (1 - p_y)^2 + \sum_{\substack{i=0 \\ i \neq y}}^{C-1} p_i^2 \\ &= p_0^2 + \cdots + p_{y-1}^2 + (1 - p_y)^2 + p_{y+1}^2 + \ldots + p_{C-1}^2 \end{aligned} \tag{2.4}$$

3. (spherical) For any $\beta > 1$, a convex funvtion $G(\boldsymbol{p}) = \left(\sum_{i=0}^{C-1} p_i^\beta\right)^{\frac{1}{\beta-1}}$ produces the spherical loss function with parameter β as follows:

$$\ell_{\text{sphere}}^{(\beta)}(y, \boldsymbol{p}) = \frac{p_y^{\beta-1}}{(\sum_{i=0}^{C-1} p_i^\beta)^{\frac{\beta-1}{\beta}}}. \tag{2.5}$$

The following Theorem 1 guarantees that for any strictly consistent loss function $\ell \in \mathcal{L}$, we can consistently estimate the true latent parameter by minimizing the expectation of the loss $L(Z, \boldsymbol{\theta}) = \ell(Y, \boldsymbol{f}(X; \boldsymbol{\theta}))$. The reason is that in theory, the expectation of the loss is uniquely minimized at the true parameter itself.

Theorem 1. *Let the model distribution be $F = (F_X, \boldsymbol{f}, \boldsymbol{\theta}^*)$. The true latent parameter $\boldsymbol{\theta}^*$ minimizes the expected loss $L(Z, \boldsymbol{\theta}) = \ell(Y, \boldsymbol{f}(X; \boldsymbol{\theta}))$ with respect to the model distribution F. That is*

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_F \left[L(Z, \boldsymbol{\theta}) \right] = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_F \left[\ell(Y, \boldsymbol{f}(X; \boldsymbol{\theta})) \right].$$

We provide the proof of Theorem 1 in Appendix A.1.

2.3 Vector and Matrix Notations

We use vector and matrix notations in many parts of this thesis. For clarity, we will define all the notations we use here. We first define in Definition 3 what 2-norms of a vector and a matrix are. Note that the latter is not used to state the main results but instead in their proofs. Lastly, we define in Definition 4 a Frobenius norm of a matrix, which is crucial in the proofs of our main results.

Definition 3. (2-norm) For a vector $v \in \mathbb{R}^n$, the 2-norm of v is

$$\|v\|_2 = \sqrt{\sum_{i=1}^{n} v_i^2},$$

and for a matrix $A \in \mathbb{R}^{m \times n}$, the 2-norm of A is

$$\|A\|_2 = \sup_{v \neq 0} \frac{\|Av\|_2}{\|v\|_2}.$$

Definition 4. (Frobenius norm) Consider a matrix $A \in \mathbb{R}^{m \times n}$. Let a_{ij} be the element in the i-th row and the j-th column of matrix A. The Frobenius norm of A is

$$\|A\|_{Frob} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}.$$

Chapter 3

Related Work

As shown in Chapter 2, any loss function in the family of strictly consistent loss functions can consistently estimate latent parameters through loss minimization method. This family of loss functions, which is crucial to this thesis, is introduced together with its closed-form formula back in [Savage, 1971], and is revisited again with a modern mathematical notation in [Gneiting and Raftery, 2007]. There exists past literature that studies the behaviors of loss functions in the family of strictly consistent loss function. For example, [de Leeuw, 2019] studies the robustness of loss functions in this family when they are used as evaluation metrics for probability forecast. Similar to this thesis, they also consider three main loss functions: the logarithmic loss function, the quadratic loss function, and the spherical loss function. The main difference is that our work study the robustness property of parameter estimation instead.

The logarithmic loss function, the equivalence of MLE method, is the most prevalent loss function since its introduction by Fisher [Fisher, 1922, Fisher, 1925]. Many existing works advocate its use and study its good properties as summarized in [Norden, 1972]. The logarithmic loss function is also proven to be a universal loss function [Painsky and Wornell, 2018] in the sense that when we minimize the logarithmic loss, we are also guaranteed to drive the losses corresponding to other loss functions to be small as well. However, as we will show in this paper, the logarithmic loss function can be very sensitive to corruption or outliers. Our work is not the first to study the robustness property of the logarithmic loss function or MLE. [White, 1982] studies

the effect of model misspecification on MLE and [Pregibon et al., 1981] demonstrates the sensitivity of MLE method in a logistic regression model. This thesis is, however, the first to compare the robustness properties across different loss functions in the family of strictly consistent loss functions.

One of the main tools we use to analyze the robustness property is an influence function, a classical tool in robust statistics [Cook and Weisberg, 1980]. Influence functions can be written in a closed-form formula [Cook and Weisberg, 1982], making the analysis of robustness property easier, as demonstrated in Chapter 4 and 5. Although it originates in the field of robust statistics, many recent machine learning literature starts to utilize influence functions to study the robustness behaviors of machine learning models. In [Koh and Liang, 2017], the influence functions are used to study the behaviors of black-box machine learning models by identifying a data point that affects the prediction the most. Similarly, [Broderick et al., 2020] uses influence functions to identify the most influential data points and study how robust the models are when these few data points are dropped out.

Similar to our thesis, many existing works also try to mitigate the effect from corruption or outliers through the properties of loss functions. Previous works either modify loss functions to minimize the effect of corruption or demonstrate robustness behaviors of specific loss functions in some settings. [Ghosh et al., 2017] studies corruption in the multinomial outcome models in the form of uniform label noise and concludes that the absolute loss function provides the most robust prediction. In robust statistics literature, [Huber, 1992] uses a hybrid loss function, called Huber loss, which combines the absolute loss function and the quadratic loss function together to reduce the effect from extreme outliers. On the other hand, [Barron, 2019] parameterizes a family of robust loss functions and let model training process discovers the parameter that is most robust for the training data by itself. However, none of these works studies the robustness of loss functions when the consistency of the parameter estimates is required. To the best of our knowledge, we are the first to study the robustness behaviors of loss functions specific to the family of strictly consistent loss functions for multinomial outcomes models.

Another way to deal with corruption in the training data is to use outlier-filtering based algorithms to filter out potential outliers before or while performing parameter estimation. Examples of such algorithms are described in [Fischler and Bolles, 1981, Diakonikolas et al., 2019, Breunig et al., 2000]. These algorithms, however, can be detrimental if the training data does not contain corruption or significant noise because they unnecessarily filter out data and can substantially impact the parameter estimates if the initial training data is limited in number. Achieving robustness through the choice of loss functions thus has an advantage in that aspect.

Chapter 4

Corruption & Robustness

Theorem 1 implies that we can consistently estimate model parameters from the observations using any loss function that is in the form in Eq. (2.2) if the data is not corrupted. These loss functions include the logarithmic loss function, the quadratic loss function, and the spherical loss function. However, in the presence of corruption, different loss functions have different robustness properties for parameter estimation. In Chapter 4 - 6, we study the theoretical behaviors of parameter estimation using different loss functions when the model distribution contains some corruption. In this chapter, we provide definitions and frameworks to study the robustness of loss functions. We mainly focus on the definitions of corruption frameworks and influence functions. We will also state properties of influence functions that are crucial for the theoretical results in the following chapters. In Chapter 5, we will explore the concept of *partial robustness* which concerns the magnitude of influence functions of each loss function. We then state theorems that provide concrete upperbounds of the influence functions of the quadratic loss function. In Chapter 6, we will expand the results on the influence functions by considering instead the properties of the bias of parameter estimates in a one-dimensional logistic regression model when the logarithmic loss function and the quadratic loss function are used. These properties are parts of the concept of *strong robustness* that is explored in the chapter.

To study the robustness of loss functions, we first define what corruption is. In this work, we consider two types of corruption: a distributional corruption and an

empirical corruption. For each type of corruption, we introduce the corresponding definition of influence functions as the frameworks to analyze the robustness behavior of each loss function.

4.1 Distributional Corruption

In this corruption framework, we assume that the corruption happens on the data distribution itself. Specifically, we consider the training data in which $1-\epsilon$ fraction of the data is generated from a model distribution F, whereas the remaining ϵ fraction of the data is generated from any other distribution G, which we refer to as a corrupted distribution. We formally define this distributional corruption framework as (ϵ, G)-corruption in Definition 5.

Definition 5. Given a corruption fraction $\epsilon > 0$, a model distribution F, and any corrupted distribution G of (X, Y), a data distribution is (ϵ, G)-corrupted if data $Z = (X, Y)$ is generated from the following process

$$\begin{aligned} p &\sim Bernoulli(\epsilon) \\ Z &\sim \begin{cases} F & \text{if } p = 0 \\ G & \text{if } p = 1 \end{cases} \end{aligned} \tag{4.1}$$

For simplicity, we use $F_{\epsilon,G}$ to represent the (ϵ, G)-corrupted data distribution, and we use notation $Z \sim F_{\epsilon,G}$ to represent that data Z is generated according to the data-generating process in Eq. (4.1).

Following Definition 5, we have that for any function h,

$$\mathbb{E}_{F_{\epsilon,G}}\left[h\left(Z\right)\right] = \left(1-\epsilon\right)\mathbb{E}_{F}\left[h\left(Z\right)\right] + \epsilon\mathbb{E}_{G}\left[h\left(Z\right)\right]. \tag{4.2}$$

Because the size of the bias from (ϵ, G)-corruption is hard to quantify, we instead measure the effect of corruption through the lens of influence functions. In this distributional corruption case, the influence function is the rate of change in the parameter

estimates from loss minimization method right when corruption G is inserted into the model distribution. We formally define influence functions in Definition 6.

Definition 6. (Influence Function) Let $\hat{\boldsymbol{\theta}}_{\epsilon,G}$ be the parameter that minimizes the expected loss with respect to the (ϵ, G)-corrupted distribution $F_{\epsilon,G}$:

$$\hat{\boldsymbol{\theta}}_{\epsilon,G} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \mathbb{E}_{F_{\epsilon,G}} \left[L(Z, \boldsymbol{\theta})\right] = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \mathbb{E}_{F_{\epsilon,G}} \left[\ell(Y, \boldsymbol{f}(X; \boldsymbol{\theta}))\right].$$

The influence of the corruption G is

$$\mathcal{I}(G) = \left.\frac{d\,\hat{\boldsymbol{\theta}}_{\epsilon,G}}{d\,\epsilon}\right|_{\epsilon=0}.$$

If the parameter is one-dimensional θ, then $\mathcal{I}(G)$ is also one-dimensional. If the parameter is multi-dimensional $\boldsymbol{\theta}$, then $\mathcal{I}(G)$ is multi-dimensional. Similar to the celebrated result from [Cook and Weisberg, 1982], the influence function in Definition 6 turns out to also have a closed-form formula. We first derive $\frac{d\,\hat{\boldsymbol{\theta}}_{\epsilon,G}}{d\,\epsilon}$ in the following lemma, which immediately leads to the closed-form formula of the influence function.

Lemma 1. The derivative $\frac{d\,\hat{\boldsymbol{\theta}}_{\epsilon,G}}{d\,\epsilon}$ is equal to

$$-\left((1-\epsilon)^2\mathbb{E}_F\left[\nabla^2_{\boldsymbol{\theta}} L(Z, \hat{\boldsymbol{\theta}}_{\epsilon,G})\right] + \epsilon(1-\epsilon)\mathbb{E}_G\left[\nabla^2_{\boldsymbol{\theta}} L(Z, \hat{\boldsymbol{\theta}}_{\epsilon,G})\right]\right)^{-1} \mathbb{E}_G\left[\nabla_{\boldsymbol{\theta}} L(Z, \hat{\boldsymbol{\theta}}_{\epsilon,G})\right]$$

Proof of Lemma 1 is in Appendix A.2. Using lemma 1, we can directly calculate the influence function $\mathcal{I}(G)$ by noticing that at $\epsilon = 0$, $F_{0,G} = F$, and thus from Theorem 1, $\hat{\boldsymbol{\theta}}_{0,G} = \boldsymbol{\theta}^*$. This results in the following theorem.

Theorem 2. *The influence function of the corruption G is the negative of the inverse of the expected hessian of the loss at $\boldsymbol{\theta}^*$ with respect to distribution F multiplied with the expected gradient of the loss at $\boldsymbol{\theta}^*$ with respect to distribution G:*

$$\mathcal{I}(G) = -\mathbb{E}_F\left[\nabla^2_{\boldsymbol{\theta}} L(Z, \boldsymbol{\theta}^*)\right]^{-1} \mathbb{E}_G\left[\nabla_{\boldsymbol{\theta}} L(Z, \boldsymbol{\theta}^*)\right], \tag{4.3}$$

and for one-dimensional parameter θ,

$$\mathcal{I}(G) = -\frac{\mathbb{E}_G\left[\nabla_\theta L(Z, \theta^*)\right]}{\mathbb{E}_F\left[\nabla_\theta^2 L(Z, \theta^*)\right]}. \tag{4.4}$$

4.2 Empirical Corruption

Consider a multinomial outcome model $F = (F_X, \boldsymbol{f}, \theta^*)$. Let $\{z_i\}_{i=1}^n$ be n training data that are independent and identically distributed from the model distribution F. Each $z_i = (x_i, y_i)$ consists of a covariate $x_i \in \mathcal{X}$ and a label $y_i \in \mathcal{Y}$. To empirically estimate the model parameter θ^*, we perform an empirical loss minimization similar to the loss minimization in Theorem 1. Given a strictly consistent loss function $\ell \in \mathcal{L}$, a loss between a model parameter $\boldsymbol{\theta} \in \Theta$ and a data point z is

$$L(z, \boldsymbol{\theta}) = \ell(y, \boldsymbol{f}(x; \boldsymbol{\theta})).$$

In the empirical setting, we again estimate model parameter by performing empirical loss minimization. That is, we simply determine a model parameter $\hat{\boldsymbol{\theta}}$ that minimizes the average loss across all n training data points:

$$\hat{\boldsymbol{\theta}} := \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{i=1}^{n} L(z_i, \boldsymbol{\theta}).$$

We consider an empirical corruption as a new data point that is added to the training data. The new data point can be adversarially chosen to alter the estimated parameter $\hat{\boldsymbol{\theta}}$ the most. Let z_{new} be the new data point that is added to the training data. Ideally, we want the parameter estimates to be robust in the sense that the parameter $\hat{\boldsymbol{\theta}}$ estimated from the training data $\{z_1, z_2, \ldots, z_n\}$ does not differ too much from the parameter $\hat{\boldsymbol{\theta}}_{new}$ estimated from the training data $\{z_1, z_2, \ldots, z_n, z_{new}\}$. However, to the best of our knowledge, the quantity $\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{new}$ is generally hard to quantify and does not have a closed-form formula. We instead use empirical influence functions, a classical tool used in robust statistics [Cook and Weisberg, 1980], as a proxy to quantify $\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{new}$.

We will now explain what empirical influence functions of the new data point z_{new} is. Using notations adapted mostly from [Koh and Liang, 2017], we formally define the influence function as follows.

Definition 7. (Empirical Influence Function) Let $\hat{\boldsymbol{\theta}}_{\epsilon,z}$ be the parameter that minimizes the empirical loss with the loss on a data point z upweighted by ϵ:

$$\hat{\boldsymbol{\theta}}_{\epsilon,z} := \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} L(z_i,\boldsymbol{\theta}) + \epsilon L(z,\boldsymbol{\theta}).$$

The empirical influence function $\mathcal{I}(z)$ is defined as the derivative of $\hat{\boldsymbol{\theta}}_{\epsilon,z}$ with respect to ϵ at $\epsilon = 0$:

$$\mathcal{I}(z) := \left.\frac{d\,\hat{\boldsymbol{\theta}}_{\epsilon,z}}{d\,\epsilon}\right|_{\epsilon=0}.$$

In other words, the influence function is the rate of change in the optimal parameter when the loss on one data point z is upweighted from 0. Note that with this definition, $\hat{\boldsymbol{\theta}}_{\frac{1}{n},z_{\text{new}}}$ is the parameter that minimizes the empirical loss after adding the new data point z_{new}. The influence function $\mathcal{I}(z_{\text{new}})$ can also be seen as a linear approximation of the effect from the new training data z_{new}. That is the change in the parameter estimate after adding training data z_{new} can be linearly approximated as

$$\hat{\boldsymbol{\theta}}_{\frac{1}{n},z_{\text{new}}} - \hat{\boldsymbol{\theta}} \approx \frac{1}{n}\mathcal{I}(z_{\text{new}}).$$

Unlike the bias $\hat{\boldsymbol{\theta}}_{\frac{1}{n},z_{\text{new}}} - \hat{\boldsymbol{\theta}}$, the influence function $\mathcal{I}(z_{\text{new}})$ can be written in a closed-form formula, making it easier to be analyzed. Using a classical result of empirical influence functions from [Cook and Weisberg, 1982], we have the following theorem.

Theorem 3. *The empirical influence function of a new data point z is the negative of the inverse of the average hessian of the loss at $\hat{\boldsymbol{\theta}}$ across the n original data points*

$\{z_i\}_{i=1}^n$ *multiplied with the gradient of the loss of the new data point* z *at* $\hat{\boldsymbol{\theta}}$*:*

$$\mathcal{I}(z) = -H_{\hat{\boldsymbol{\theta}}}^{-1} \nabla_{\boldsymbol{\theta}} L\left(z, \hat{\boldsymbol{\theta}}\right),$$

where

$$H_{\hat{\boldsymbol{\theta}}} := \frac{1}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}}^2 L\left(z_i, \hat{\boldsymbol{\theta}}\right).$$

We provide the proof of Theorem 3 in Appendix A.3 for completeness. Note that this formula for the empirical influence functions is very similar to that of the distributional influence function in Theorem 2. The proofs of these two formulas are also based on the same approach.

The influence functions, both distributional and empirical versions, of different loss functions behave differently, and thus have different robustness property. In this thesis, we will mostly study the distributional (ϵ, G)-corruption and the corresponding influence functions. Therefore, most results are based on the distributional corruption. We first introduce a concept of *partial robustness* in Chapter 5. A loss function ℓ is *partially robust* if its corresponding (distributional) influence function is bounded. It turns out that different loss functions have different partial robustness property. Although we do not demonstrate in this thesis, it is worth to note that many of the results about partial robustness also hold for the empirical corruption and the corresponding empirical influence function. We next introduce a concept of *strong robustness* in Chapter 6. A loss function ℓ is *strongly robust* if the 2-norm of the bias vector of the parameter estimate in the presence of (ϵ, G)-corruption is bounded. We will show some strong robustness results for a one-dimensional logistic regression model. Note that the results about the strong robustness property only apply to the distributional corruption framework.

Chapter 5

Partial Robustness Property

Now that we have introduced all the necessary tools, we can formally define the robustness property of loss functions. In this chapter, we will introduce the concept of *partial robustness* property and show that only some loss functions in the family of strictly consistent loss functions are partially robust. Specifically, among the three loss functions we introduce in Chapter 2, the quadratic loss function and the spherical loss function with parameter $\beta \geq 2$ are partially robust. On the other hand, this property is not shared with the logarithmic loss function, which is the default loss function used in parameter estimation for multinomial outcome models. Afterward, we further investigate the quadratic loss function, which is the main loss function of our interest. We will show that under some regularity assumptions, the quadratic loss function is not only partially robust, but its influence function from any corruption G can also be bounded by a concrete term that depends only on the true parameter. We divide the results for this property into two cases. For the one-dimensional parameter case, we provide a bound on the absolute value of the influence function of the quadratic loss function. For the multi-dimensional parameter case, we instead bound the 2-norm of the influence function vector.

5.1 Partial Robustness

In multinomial outcome models, a predicted outcome from a covariate x is a probability vector $\boldsymbol{f}(x;\theta)$ over the space of categorical labels. Let C be the number of categories. Before stating the main definitions and results, we first define some notations. Given a covariate x, a predicted probability vector from the model $F = (F_X, \boldsymbol{f}, \boldsymbol{\theta}^*)$ is

$$\boldsymbol{f}(x;\theta) = \begin{bmatrix} f_0(x;\theta) \\ f_1(x;\theta) \\ \vdots \\ f_{C-1}(x;\theta) \end{bmatrix}.$$

Let $J(x, \boldsymbol{\theta})$ represent a Jacobian matrix of $\boldsymbol{f}(x;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$:

$$J(x,\boldsymbol{\theta}) = \begin{bmatrix} \frac{d\, f_0(x;\boldsymbol{\theta})}{d\,\theta_0} & \cdots & \frac{d\, f_0(x;\boldsymbol{\theta})}{d\,\theta_{d-1}} \\ \vdots & & \\ \frac{d\, f_{C-1}(x;\boldsymbol{\theta})}{d\,\theta_0} & \cdots & \frac{d\, f_{C-1}(x;\boldsymbol{\theta})}{d\,\theta_{d-1}} \end{bmatrix}. \tag{5.1}$$

Before we can define the notion of *partial robustness*, we need to first introduce two regularity assumptions about the model distribution F. We will only consider model distribution F that satisfies these two assumptions when we study *partial robustness* property of loss functions. The two regularity assumptions are as follows.

Assumption 1. There is a common upperbound $K > 0$ such that

$$\forall\, 0 \le i < C, \forall\, 0 \le j < d, \forall\, x \in \mathcal{X}, \quad \left| \frac{d\, f_i(x;\boldsymbol{\theta})}{d\,\theta_j}\, \theta_j \right| \le K. \tag{5.2}$$

Note that Assumption 1 is easily satisfied if θ_j is bounded for every $0 \le j < d$ and if the functions $h(\theta) = f_i(x;\theta)$ are K-Lipschitz continuous under Euclidean metric for every $0 \le i < C$ and for every $x \in \mathcal{X}$.

Assumption 2. For any loss function $\ell \in \mathcal{L}$, the Frobenius norm of the expectation of the hessian of loss $L(Z;\boldsymbol{\theta}^*) = \ell(Y, \boldsymbol{f}(X;\boldsymbol{\theta}^*))$ is finite:

$$\forall\, \ell \in \mathcal{L}, \quad \left\| \mathbb{E}_F \left[\nabla^2_{\boldsymbol{\theta}} L(Z;\boldsymbol{\theta}^*) \right] \right\|_{Frob} < \infty.$$

Now, we can define partial robustness property as follows.

Definition 8. (Partial Robustness) A loss function $\ell \in \mathcal{L}$ is *partially robust* if and only if for any model F that satisfies Assumption 1 and 2, the influence function from any corruption G has bounded 2-norm:

$$\exists\, U \in \mathbb{R}^+, \forall\, \text{corruption } G, \quad \left\| \mathcal{I}_\ell(G) \right\|_2 < U.$$

We next establish a theoretical result that can be used to determine if a loss function is partially robust. First, consider that the gradient of $L(z, \boldsymbol{\theta})$ can be expanded as follows:

$$\nabla_{\boldsymbol{\theta}} L(z, \boldsymbol{\theta}) = \sum_{i=0}^{C-1} \frac{\partial \ell(y, \boldsymbol{f})}{\partial f_i} \nabla_{\boldsymbol{\theta}} f_i(x;\boldsymbol{\theta}) = J(x;\boldsymbol{\theta})^T \nabla_{\boldsymbol{f}} \ell(y, \boldsymbol{f}). \tag{5.3}$$

Note that the term $J(x;\boldsymbol{\theta})$ depends only on the model distribution and does not depend on the choice of loss function. Because we only consider model F that follows Assumption 1 and 2, the boundedness of the 2-norm of the influence function thus only depends on the choice of loss functions and the corresponding term $\nabla_{\boldsymbol{f}} \ell(y, \boldsymbol{f})$.

Theorem 4. *If a loss function $\ell \in \mathcal{L}$ satisfies*

$$\exists\, U_1 > 0, \forall\, 0 \le y < C, \forall\, \boldsymbol{p} \in \mathcal{P}_C, \quad \left\| \nabla_{\boldsymbol{p}} \ell(y, \boldsymbol{p}) \right\|_2 \le U_1, \tag{5.4}$$

then ℓ is partially robust.

Before we begin to prove Theorem 4, we first state the following lemma.

Lemma 2. For any matrix $A \in \mathbb{R}^{m \times n}$ and any vector $v \in \mathbb{R}^n$

$$\|Av\|_2 \leq \|A\|_{Frob} \|v\|_2 .$$

We provide the proof for Lemma 2 in Appendix A.4. Now, we provide the proof of Theorem 4 below.

Proof. (Theorem 4) Consider a loss function $\ell \in \mathcal{L}$ that satisfies the property in Eq. (5.4). That is, there exists $U_1 \in \mathbb{R}^+$ such that

$$\forall\, 0 \leq y < C, \forall\, \boldsymbol{p} \in \mathcal{P}_C, \quad \|\nabla_{\boldsymbol{p}} \ell(y, \boldsymbol{p})\|_2 \leq U_1.$$

To prove that ℓ is partially robust, we consider a model F which satisfies Assumption 1 and 2. Let

$$\left\| \mathbb{E}_F \left[\nabla^2_{\boldsymbol{\theta}} L(Z; \boldsymbol{\theta}^*) \right]^{-1} \right\|_{Frob} = U_2,$$

when $U_2 \in \mathbb{R}^+$. We have that for any corruption G,

$$\begin{aligned}
\|\mathcal{I}_\ell(G)\|_2 &= \left\| \mathbb{E}_F \left[\nabla^2_{\boldsymbol{\theta}} L(Z, \boldsymbol{\theta}^*) \right]^{-1} \mathbb{E}_G \left[\nabla_{\boldsymbol{\theta}} L(Z, \boldsymbol{\theta}^*) \right] \right\|_2 && (5.5) \\
&\leq \left\| \mathbb{E}_F \left[\nabla^2_{\boldsymbol{\theta}} L(Z; \boldsymbol{\theta}^*) \right]^{-1} \right\|_{Frob} \left\| \mathbb{E}_G \left[\nabla_{\boldsymbol{\theta}} L(Z, \boldsymbol{\theta}^*) \right] \right\|_2 && (5.6) \\
&\leq \left\| \mathbb{E}_F \left[\nabla^2_{\boldsymbol{\theta}} L(Z; \boldsymbol{\theta}^*) \right]^{-1} \right\|_{Frob} \mathbb{E}_G \left[\left\| \nabla_{\boldsymbol{\theta}} L(Z, \boldsymbol{\theta}^*) \right\|_2 \right] && (5.7) \\
&= U_2 \mathbb{E}_G \left[\left\| J(X; \boldsymbol{\theta}^*)^T \nabla_{\boldsymbol{f}} \ell(Y, \boldsymbol{f}(X; \boldsymbol{\theta}^*)) \right\|_2 \right] && (5.8) \\
&\leq U_2 \mathbb{E}_G \left[\left\| J(X; \boldsymbol{\theta}^*)^T \right\|_{Frob} \left\| \nabla_{\boldsymbol{f}} \ell(Y, \boldsymbol{f}(X; \boldsymbol{\theta}^*)) \right\|_2 \right] && (5.9) \\
&\leq U_2 \mathbb{E}_G \left[\sqrt{dC} K U_1 \right] = \sqrt{dC} K U_1 U_2. && (5.10)
\end{aligned}$$

The equality (5.5) comes from Theorem 2 and the fact that a sign disappears under 2-norm. The inequalities (5.6) and (5.9) are the applications of Lemma 2. The inequality (5.7) utilizes Jensen's inequality and the fact that 2-norm is a convex function. Lastly, the inequality (5.10) results from Assumption 1 and the condition of loss function assumed in Eq. (5.4). Therefore, ℓ is partially robust as desired. □

5.2 The Three Loss Functions

We now analyze the influence functions under the three different loss functions introduced in Chapter 2: the logarithmic loss function, the quadratic loss function, and the spherical loss function.

5.2.1 Logarithmic Loss Function

Consider the logarithmic loss function $\ell_{log}(y, \boldsymbol{p}) = -\log(p_y)$. We have that

$$\frac{\partial \ell_{log}(y, \boldsymbol{p})}{\partial p_i} = \frac{\partial}{\partial p_i} - \log(p_y) = \begin{cases} -\frac{1}{p_y} & \text{if } i = y, \\ 0 & \text{otherwise.} \end{cases}$$

This makes $\|\nabla_{\boldsymbol{p}} \ell_{log}(y, \boldsymbol{p})\|_2$ unbounded for some y and $\boldsymbol{p}$. It becomes unbounded when, for example, $y \neq 0$ but $p_0 = 1$ and $p_j = 0$ for $0 < j < C$. Therefore, the logarithmic loss function doesn't satisfy the condition (5.4) in Theorem 4 and thus is not partially robust. This is consistent with what we have already seen in Chapter 1 that even one faulty data point can have unbounded impact on the parameter estimates when the logarithmic loss function is used.

5.2.2 Quadratic Loss Function

Consider the quadratic loss function $\ell_{quad}(y, \boldsymbol{p}) = \sum_{j=0}^{C-1} p_j^2 - 2p_j + 1$. The gradient is as follows:

$$\frac{\partial \ell_{quad}(y, \boldsymbol{p})}{\partial p_i} = \frac{\partial}{\partial p_i} \sum_{j=0}^{C-1} p_j^2 - 2\frac{\partial}{\partial p_i} p_y = \begin{cases} 2p_i - 2 & \text{if } i = y, \\ 2p_i & \text{otherwise.} \end{cases}$$

Because $0 \leq p_i \leq 1$ for every $0 \leq i < C$, we have that $-2 \leq \frac{\partial \ell_{quad}(y,\boldsymbol{p})}{\partial p_i} \leq 2$ for every $0 \leq i < C$. Therefore,

$$\|\nabla_{\boldsymbol{p}} \ell_{quad}(y, \boldsymbol{p})\|_2 = \sqrt{\sum_{i=0}^{C-1} \left(\frac{\partial \ell_{quad}(y, \boldsymbol{p})}{\partial p_i} \right)^2} \leq \sqrt{\sum_{i=0}^{C-1} 4} = 2\sqrt{C},$$

and thus is bounded for any $0 \leq y < C$ and $\boldsymbol{p} \in \mathcal{P}_C$. From Theorem 4, the quadratic loss function is therefore partially robust.

5.2.3 Spherical Loss Function

Lastly, we consider the spherical loss function with parameter β when $\beta > 1$:

$$\ell^{(\beta)}_{sphere}(y, \boldsymbol{p}) = \frac{p_y^{\beta-1}}{\left(\sum_{i=0}^{C-1} p_i^{\beta}\right)^{\frac{\beta-1}{\beta}}}.$$

Unlike the logarithmic loss function and the quadratic loss function, the gradient of $\ell^{(\beta)}_{sphere}(y, \boldsymbol{p})$ is more complex, making its robustness behavior differ across different β:

$$\frac{\partial \ell^{(\beta)}_{sphere}(y, \boldsymbol{p})}{\partial p_i} = \begin{cases} \frac{(\beta-1)p_y^{\beta-2}}{\left(\sum_{j=0}^{C-1} p_j^{\beta}\right)^{\frac{\beta-1}{\beta}}} \left[\frac{p_y^{\beta}}{\sum_{j=0}^{C-1} p_j^{\beta}} - 1 \right] & \text{if } i = y, \\ \frac{(\beta-1)p_i^{\beta-1} p_y^{\beta-1}}{\left(\sum_{j=0}^{C-1} p_j^{\beta}\right)^{\frac{2\beta-1}{\beta}}} & \text{otherwise.} \end{cases}$$

It turns out that not all spherical loss functions can be shown to have partial robustness property. Theorem 5 will state a subset of the spherical loss functions that are partially robust.

Theorem 5. *The spherical loss function with parameter β is partially robust when $\beta \geq 2$.*

The proof of Theorem 5 is provided in Appendix A.5. We will also see in the results from synthetic data in Section 7.1.1 that the empirical behaviors of different spherical loss functions match the result in Theorem 5. In that specific example,

the spherical loss functions with $\beta \in \{2, 3, 4\}$ are more robust than the spherical loss functions with $\beta \in \{1.5, 1.4, 1.3, 1.2, 1.1, 1.01\}$. The spherical loss function also becomes less robust when β converges to 1.

5.3 The Bounds on the Influence Function

We have shown that for the quadratic loss function and the spherical loss function with $\beta \geq 2$, the 2-norm of the influence functions are bounded regardless of the corruption G. However, apart from the fact that it is bounded, we do not know how large the magnitude of the influence function is compared to the true parameter. It turns out that under some mild assumptions about the model distribution F, we can derive concrete upper bounds of the influence function for the quadratic loss function.

We will first show the results in the one-dimensional parameter case, followed by the results in the multi-dimensional parameter case. We consider these two cases separately because the results in the former case are more simple and the assumptions used are more intuitive. In both cases, we again make an assumption that the model probability function $\boldsymbol{f}$ has bounded derivative as stated in Assumption 1.

5.3.1 Quadratic Loss Function: One-Dimensional Case

We will show that under some mild assumptions, when the quadratic loss function ℓ_{quad} is used for parameter estimation, the influence function from any corruption G is bounded by the absolute value of the true parameter times a constant. We first state the necessary assumption we will use.

Assumption 3. There exists a lower bound $M > 0$ such that

$$\forall\, \theta \in \Theta, \quad \mathbb{E}_{F_X}\left[\sum_{i=0}^{C-1}\left(\frac{d\, f_i(X;\theta)}{d\,\theta}\,\theta\right)^2\right] \geq M.$$

Note that Assumption 3 only depends on covariate X and its true distribution F_X. It is a fair assumption to be made because if violated, changing θ will on average barely

change any terms in the outcome probability $\boldsymbol{f}(X;\theta)$. In other words, two different parameters will generate two sets of observations that distribute very similarly. In that case, the value of θ itself will not be important for the model distribution, and there will not be any need to estimate parameter θ in the first place.

Assume that Assumption 1 and 3 hold, we will show in Theorem 6 that under the quadratic loss function, the influence function from any distribution G is bounded.

Theorem 6. *Assume that the model distribution $F = (F_X, \boldsymbol{f}, \theta^*)$ follows Assumption 1 and 3. The absolute value of the influence function of any corruption G on the parameter estimation with the quadratic loss function is upper bounded by $\frac{CK}{M}|\theta^*|$:*

$$\forall \textit{ corruption } G, \quad |\mathcal{I}_{quad}(G)| \leq \frac{CK}{M}|\theta^*|.$$

To prove Theorem 6, we first establish a lemma that greatly simplifies the influence function formula of any strictly consistent loss function.

Lemma 3. For any strictly consistent loss function $\ell \in \mathcal{L}$,

$$\mathbb{E}_F\left[\sum_{i=0}^{C-1} \frac{d^2 f_i(X;\theta^*)}{d\theta^2} \frac{\partial \ell(Y, \boldsymbol{f}(X;\theta^*))}{\partial f_i}\right] = 0$$

The proof is in Appendix A.6. Using Lemma 3, we can reduce the denominator of the influence function, in Eq. (4.4), of the quadratic loss function to a simpler term. We show the proof here to highlight the simplicity that results from the nature of the quadratic loss function.

Corollary 1. For the quadratic loss function ℓ_{quad},

$$\mathbb{E}_F\left[\nabla_\theta^2 \ell_{quad}(Y, \boldsymbol{f}(Y,\theta^*))\right] = 2\mathbb{E}_{F_X}\left[\sum_{i=0}^{C-1}\left(\frac{d\, f_i(X;\theta^*)}{d\theta}\right)^2\right].$$

Proof. From Lemma 3, for any strictly consistent loss function $\ell \in \mathcal{L}$,

$$\begin{aligned}
&\mathbb{E}_F\left[\nabla_\theta^2 \ell(Y, \boldsymbol{f}(Y;\theta^*))\right] \\
&= \mathbb{E}_F\left[\sum_{i=0}^{C-1} \frac{d^2 f_i(X;\theta^*)}{d\theta^2} \frac{\partial \ell(Y,\boldsymbol{f})}{\partial f_i}\right] + \mathbb{E}_F\left[\sum_{i=0}^{C-1}\sum_{j=0}^{C-1} \frac{d\, f_i(X;\theta^*)}{d\theta} \frac{d\, f_j(X;\theta^*)}{d\theta} \frac{\partial^2 \ell(Y,\boldsymbol{f})}{\partial f_i \partial f_j}\right] \\
&= \mathbb{E}_F\left[\sum_{i=0}^{C-1}\sum_{j=0}^{C-1} \frac{d\, f_i(X;\theta^*)}{d\theta} \frac{d\, f_j(X;\theta^*)}{d\theta} \frac{\partial^2 \ell(Y,\boldsymbol{f})}{\partial f_i \partial f_j}\right].
\end{aligned}$$

Specifically for the quadratic loss function ℓ_{quad}, we have

$$\frac{\partial^2 \ell_{quad}(y,\boldsymbol{f})}{\partial f_i \partial f_j} = \begin{cases} 2 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{5.11}$$

Therefore,

$$\mathbb{E}_F\left[\nabla_\theta^2 \ell_{quad}(Y, \boldsymbol{f}(Y,\theta^*))\right] = 2\mathbb{E}_{F_X}\left[\sum_{i=0}^{C-1}\left(\frac{d\, f_i(X;\theta^*)}{d\theta}\right)^2\right]$$

as desired. □

Another great property of the quadratic loss function is that it has bounded derivative. Because

$$\frac{\partial \ell_{quad}(y,\boldsymbol{f})}{\partial f_i} = \begin{cases} 2p_i - 2 & \text{if } i = y, \\ 2p_i & \text{otherwise,} \end{cases} \tag{5.12}$$

we have that

$$\forall\, 0 \le i < C \quad \left|\frac{\partial \ell_{quad}(y,\boldsymbol{f})}{\partial f_i}\right| \le 2. \tag{5.13}$$

With Corollary 1 and all the assumptions introduced, we prove our main robustness result in the one-dimensional parameter case.

Proof. (Theorem 6) Consider any corruption G. We have that

$$|\mathcal{I}_{quad}(G)| = \frac{|\mathbb{E}_G\left[\nabla_\theta L_{quad}(Z,\theta^*)\right]|}{|\mathbb{E}_F\left[\nabla^2_\theta L_{quad}(Z,\theta^*)\right]|} \tag{5.14}$$

$$= \frac{\left|\mathbb{E}_G\left[\sum_{i=0}^{C-1} \frac{d\, f_i(X;\theta)}{d\theta} \frac{\partial \ell_{quad}(Y,\boldsymbol{f})}{\partial f_i}\right]\right|}{\left|2\mathbb{E}_{F_X}\left[\sum_{i=0}^{C-1}\left(\frac{d\, f_i(X;\theta^*)}{d\theta}\right)^2\right]\right|} \tag{5.15}$$

$$\leq \frac{\mathbb{E}_G\left[\sum_{i=0}^{C-1}\left|\frac{d\, f_i(X;\theta^*)}{d\theta}\right|\left|\frac{\partial \ell_{quad}(Y,\boldsymbol{f})}{\partial f_i}\right|\right]}{2\mathbb{E}_{F_X}\left[\sum_{i=0}^{C-1}\left(\frac{d\, f_i(X;\theta^*)}{d\theta}\right)^2\right]} \tag{5.16}$$

$$\leq \frac{\mathbb{E}_G\left[\sum_{i=0}^{C-1} 2\frac{K}{|\theta^*|}\right]}{2\frac{M}{\theta^{*2}}} = \frac{CK}{M}\left|\theta^*\right|. \tag{5.17}$$

The equality (5.14) comes directly from Theorem 2. The equality (5.15) results from the chain-rule expansion of $\nabla_\theta L(Z,\theta)$ and from the identity in Corollary 1. The inequality (5.16) is the application of triangle inequality. Finally, the inequality (5.17) results directly from Assumption 1, Assumption 3, and the property of the quadratic loss function in Eq. (5.13). □

5.3.2 Quadratic Loss Function: Multi-Dimensional Case

In the multi-dimensional parameter case, the influence function becomes a vector, rather than a scalar as in the one-dimensional case. For that reason, our result establishes a bound on the 2-norm of the influence function vector, instead of its absolute value.

Similar to the one-dimensional case, we need an assumption on the probability function $\boldsymbol{f}$ in order to establish a robustness result. However, unlike the one-dimensional case, the assumption we make here is less intuitive and more technical. Let $\lambda_{min}(A)$ represents the smallest eigenvalue of a squared matrix $A \in \mathbb{R}^{n\times n}$. Similar to the Jacobian matrix $J(x,\boldsymbol{\theta})$ defined in Eq. (5.1), we define $\tilde{J}(x,\boldsymbol{\theta})$ as

$$\tilde{J}(x, \boldsymbol{\theta}) = \begin{bmatrix} \frac{d\, f_0(x;\boldsymbol{\theta})}{d\theta_0}\theta_0 & \dots & \frac{d\, f_0(x;\boldsymbol{\theta})}{d\theta_{d-1}}\theta_{d-1} \\ \vdots & & \\ \frac{d\, f_{C-1}(x;\boldsymbol{\theta})}{d\theta_0}\theta_0 & \dots & \frac{d\, f_{C-1}(x;\boldsymbol{\theta})}{d\theta_{d-1}}\theta_{d-1} \end{bmatrix}.$$

Also, for any parameter vector $\boldsymbol{\theta}$, let $D(\boldsymbol{\theta})$ be a $d \times d$ diagonal matrix with diagonal values $\theta_0, \dots, \theta_{d-1}$ in that order. We therefore have that

$$\tilde{J}(x, \boldsymbol{\theta}) = J(x, \boldsymbol{\theta}) D(\boldsymbol{\theta}). \tag{5.18}$$

The necessary assumption is as follows.

Assumption 4. There exists a lower bound $M > 0$ such that

$$\forall\, \theta \in \Theta, \quad \lambda_{min}\left(\mathbb{E}_{F_X}\left[\tilde{J}(x, \boldsymbol{\theta})^T \tilde{J}(x, \boldsymbol{\theta})\right]\right) \geq M.$$

Theorem 7. *Assume that the model distribution $F = (F_X, \boldsymbol{f}, \theta^*)$ follows Assumption 1 and 4. The 2-norm of the influence function vector of any corruption G on the parameter estimate using the quadratic loss function is bounded by $\frac{CK\sqrt{d}}{M} \|\boldsymbol{\theta}^*\|_2$:*

$$\forall\, \textit{corruption } G, \quad \|\mathcal{I}_{quad}(G)\|_2 \leq \frac{CK\sqrt{d}}{M} \|\boldsymbol{\theta}^*\|_2 .$$

Note that this doesn't necessary mean that the magnitude of the influence function is scaled up by the dimension d of the parameter $\boldsymbol{\theta}$. The 2-norm is, by definition, already scaled up by $\sqrt{d}$ because its formula contains d terms inside the square root. Therefore, on average, the influence function of each individual parameter θ_i will be upper bounded by $\frac{CK}{M} \|\boldsymbol{\theta}^*\|_2$, a term that does not depend on the dimension d.

To prove Theorem 7, we first establish a lemma that is very similar to Lemma 3 in the one-dimensional case. This lemma also shares the exact same proof with Lemma 3.

Lemma 4. For any strictly consistent loss function $\ell \in \mathcal{L}$,

$$\mathbb{E}_F \left[\sum_{i=0}^{C-1} \nabla^2_{\boldsymbol{\theta}} f_i(X; \boldsymbol{\theta}^*) \frac{\partial \ell(Y, \boldsymbol{f})}{\partial f_i} \right] = \mathbf{0}$$

The proof of Lemma 3 is also in Appendix A.6. Using the identity

$$\nabla^2_{\boldsymbol{\theta}} L(Z, \boldsymbol{\theta}) = \left(\sum_{i=0}^{C-1} \nabla^2_{\boldsymbol{\theta}} f_i(x; \boldsymbol{\theta}) \frac{\partial \ell(y, \boldsymbol{f})}{\partial f_i} \right) + J(x, \theta)^T \nabla^2_{\boldsymbol{f}} \, \ell(y, \boldsymbol{f}(x; \boldsymbol{\theta})) J(x, \theta)$$

and the fact that $\nabla^2_{\boldsymbol{f}} \, \ell_{quad}(y, \boldsymbol{f}(x; \boldsymbol{\theta})) = 2I$ (an identity that results from Eq. (5.11)), we have the following corollary for the quadratic loss function.

Corollary 2. For the quadratic loss function ℓ_{quad},

$$\mathbb{E}_F \left[\nabla^2_{\boldsymbol{\theta}} L_{quad}(Z, \boldsymbol{\theta}^*) \right] = 2\mathbb{E}_{F_X} \left[J(X, \boldsymbol{\theta}^*)^T J(X, \boldsymbol{\theta}^*) \right].$$

Lastly, we state two lemmas that give upper bounds to the 2-norms of multiplications of matrices and vectors.

Lemma 5. For any symmetric matrix $A \in \mathbb{R}^{n \times n}$ and any vector $v \in \mathbb{R}^n$,

$$\left\| A^{-1} v \right\|_2 \leq \frac{\|v\|_2}{\lambda_{min}(A)}.$$

Lemma 6. Let $u, v \in \mathbb{R}^n$ be any two vectors. Let $D(u)$ be a $n \times n$ diagonal matrix with the i-th diagonal element equals to the i-th element of the vector u. We have that

$$\|D(u)v\|_2 \leq \|u\|_2 \|v\|_2 .$$

The proofs of Lemma 5 and Lemma 6 are in Appendix A.7 and A.8 respectively. Using these lemmas, we proceed to prove our main result in Theorem 7.

Proof. (Theorem 7) We have that

$$
\begin{aligned}
&\left\|\mathcal{I}_{quad}(G)\right\|_2 \\
&= \left\|\mathbb{E}_F\left[\nabla^2_{\boldsymbol{\theta}} L_{quad}(Z,\boldsymbol{\theta}^*)\right]^{-1}\mathbb{E}_G\left[\nabla_{\boldsymbol{\theta}} L_{quad}(Z,\boldsymbol{\theta}^*)\right]\right\|_2 && (5.19)\\
&= \frac{1}{2}\left\|\mathbb{E}_{F_X}\left[J(X,\boldsymbol{\theta}^*)^T J(X,\boldsymbol{\theta}^*)\right]^{-1}\mathbb{E}_G\left[\nabla_{\boldsymbol{\theta}} L_{quad}(Z,\boldsymbol{\theta}^*)\right]\right\|_2 && (5.20)\\
&= \frac{1}{2}\left\|D(\boldsymbol{\theta}^*)\mathbb{E}_{F_X}\left[\tilde{J}(X,\boldsymbol{\theta}^*)^T\tilde{J}(X,\boldsymbol{\theta}^*)\right]^{-1} D(\boldsymbol{\theta}^*)\mathbb{E}_G\left[\nabla_{\boldsymbol{\theta}} L_{quad}(Z,\boldsymbol{\theta}^*)\right]\right\|_2 && (5.21)\\
&\leq \frac{1}{2}\left\|\boldsymbol{\theta}^*\right\|_2\left\|\mathbb{E}_{F_X}\left[\tilde{J}(X,\boldsymbol{\theta}^*)^T\tilde{J}(X,\boldsymbol{\theta}^*)\right]^{-1} D(\boldsymbol{\theta}^*)\mathbb{E}_G\left[\nabla_{\boldsymbol{\theta}} L_{quad}(Z,\boldsymbol{\theta}^*)\right]\right\|_2 && (5.22)\\
&\leq \frac{1}{2}\frac{\left\|D(\boldsymbol{\theta}^*)\mathbb{E}_G\left[\nabla_{\boldsymbol{\theta}} L_{quad}(Z,\boldsymbol{\theta}^*)\right]\right\|_2}{\lambda_{min}\left(\mathbb{E}_{F_X}\left[\tilde{J}(X,\boldsymbol{\theta}^*)^T\tilde{J}(X,\boldsymbol{\theta}^*)\right]\right)}\left\|\boldsymbol{\theta}^*\right\|_2 && (5.23)\\
&\leq \frac{\left\|D(\boldsymbol{\theta}^*)\mathbb{E}_G\left[\nabla_{\boldsymbol{\theta}} L_{quad}(Z,\boldsymbol{\theta}^*)\right]\right\|_2}{2M}\left\|\boldsymbol{\theta}^*\right\|_2 && (5.24)
\end{aligned}
$$

The equality (5.19) is the formula of the influence function from Theorem 2. The equality (5.20) results from Corollary 2. The equality (5.21) comes from the identity in Eq. (5.18). The inequalities (5.22) and (5.23) result from Lemma 6 and 5 respectively. Lastly, the inequality (5.24) comes from Assumption 4.

The remaining task is to establish an upper bound on the numerator of Eq. (5.24), the 2-norm of $D(\boldsymbol{\theta}^*)\mathbb{E}_G\left[\nabla_{\boldsymbol{\theta}} L_{quad}(Z,\boldsymbol{\theta}^*)\right]$. Consider any corruption G, we have

$$
\begin{aligned}
\left\|D(\boldsymbol{\theta}^*)\mathbb{E}_G\left[\nabla_{\boldsymbol{\theta}} L_{quad}(Z,\boldsymbol{\theta}^*)\right]\right\|_2 &\leq \mathbb{E}_G\left[\left\|D(\boldsymbol{\theta}^*)\nabla_{\boldsymbol{\theta}} L_{quad}(Z,\boldsymbol{\theta}^*)\right\|_2\right] && (5.25)\\
&= \mathbb{E}_G\left[\left\|D(\boldsymbol{\theta}^*)J(X;\boldsymbol{\theta}^*)^T\nabla_{\boldsymbol{f}}\ell(Y,\boldsymbol{f}(X;\boldsymbol{\theta}^*)\right\|_2\right] && (5.26)\\
&= \mathbb{E}_G\left[\left\|\tilde{J}(X;\boldsymbol{\theta}^*)^T\nabla_{\boldsymbol{f}}\ell(Y,\boldsymbol{f}(X;\boldsymbol{\theta}^*)\right\|_2\right] && (5.27)\\
&\leq \mathbb{E}_G\left[\left\|\tilde{J}(X;\boldsymbol{\theta}^*)^T\right\|_{Frob}\left\|\nabla_{\boldsymbol{f}}\ell(Y,\boldsymbol{f}(X;\boldsymbol{\theta}^*)\right\|_2\right] && (5.28)\\
&\leq \mathbb{E}_G\left[\sqrt{dCK^2}\sqrt{4C}\right] = 2CK\sqrt{d}. && (5.29)
\end{aligned}
$$

The inequality (5.25) is the application of Jensen's inequality and the fact that 2-norm is a convex function. The equality (5.26) and (5.27) come from the identity in Eq. (5.3) and (5.18) respectively. The inequality (5.28) results from Lemma 2. Lastly, the inequality (5.29) comes from Assumption 1 and the property specific to

the quadratic loss function in Eq. (5.13).

Combining Eq. (5.29) with Eq. (5.24), we get the desired result. □

Chapter 6

Strong Robustness Property

We have shown in Chapter 5 that under some mild assumptions, parameter estimation using the quadratic loss function has bounded influence function from any corruption G. This however does not necessarily imply that the realized bias of the parameter estimates with (ϵ, G)-corrupted data is also bounded. In this chapter, we study the magnitude of the realized bias of the parameter estimates itself instead of the influence function. Ideally, we want a loss function to produce a small bias in the presence of any (ϵ, G)-corruption. Specifically, we want the realized bias to be bounded and scale proportionally to the corruption fraction ϵ and the norm of the parameter vector $\boldsymbol{\theta}^*$ (in other words, on the order of magnitude of $\mathcal{O}(\epsilon \|\boldsymbol{\theta}^*\|)$). We consider a loss function that exhibits such robust behavior to be *strongly robust*. We formally define the concept of *strong robustness* below.

Definition 9. (Strong Robustness) A loss function $\ell \in \mathcal{L}$ is *strongly robust* with respect to a multinomial outcome model $F = (F_X, \boldsymbol{f}, \theta^*)$ up to a corruption fraction ϵ' if and only if

$$\exists\, U > 0, \forall\, 0 \leq \epsilon < \epsilon', \forall\, \text{Corruption } G, \quad \left\| \boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_{\epsilon,G} \right\|_2 \leq \epsilon U \|\boldsymbol{\theta}^*\|$$

when $\hat{\boldsymbol{\theta}}_{\epsilon,G}$ is the parameter that minimizes the expected loss with respect to the

(ϵ, G)-corrupted distribution $F_{\epsilon,G}$:

$$\hat{\boldsymbol{\theta}}_{\epsilon,G} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \mathbb{E}_{F_{\epsilon,G}} \left[\ell(Y, \boldsymbol{f}(X;\boldsymbol{\theta}))\right].$$

In this chapter, we will show that the quadratic loss function is strongly robust with respect to a simple multinomial outcome model up to a certain corruption fraction. Specifically, we study the behavior of parameter estimation under (ϵ, G)-corruption of a one-dimensional logistic regression model. We first formally define the model:

Definition 10. (One-dimensional Logistic Regression Model) A one-dimensional logistic regression model is any model with

1. Any covariate distribution F_X on $\mathcal{X} \subseteq \mathbb{R}^+$.

2. A conditional probability $\boldsymbol{f}(X;\theta)$ with

$$f_0(x;\theta) = \frac{1}{1+e^{\theta x}},$$
$$f_1(x;\theta) = \frac{e^{\theta x}}{1+e^{\theta x}}.$$

3. A model parameter $\theta^* \in \Theta = \mathbb{R}^+$.

Note that when the parameter is one-dimensional, the realized bias of the parameter estimate is simply the absolute difference between the parameter estimate and the true parameter. In other words, the realized bias is equivalent to $\left|\theta^* - \hat{\theta}_{\epsilon,G}\right|$, and we would like it to be bounded and be on the order of magnitude of $\mathcal{O}(\epsilon\,|\theta^*|)$. We will show below that for the one-dimensional logistic regression model that satisfies Assumption 3, the bias is on the order of $\mathcal{O}(\epsilon\theta^*)$ when the quadratic loss function is used for parameter estimation. The quadratic loss function is thus strongly robust with respect to any one-dimensional logistic regression model that satisfies Assumption 3. On the other hand, we will also show that when the logarithmic loss function is used, the realized bias can be as large as $|\theta^*|$ regardless of the corruption fraction ϵ. The logarithmic loss function is thus not strongly robust with respect to any of

the one-dimensional logistic regression model. This agrees with the result in Chapter 5 which states that the logarithmic loss function is not partially robust.

Consider a one-dimensional logistic regression model F that follows Definition 10. Recall that for any corruption G and corruption fraction $0 < \epsilon < 1$, the (ϵ, G)-corrupted distribution follows the process described in Eq. (4.1). Let $\hat{\theta}^{log}_{\epsilon,G}$ and $\hat{\theta}^{quad}_{\epsilon,G}$ be the parameters that minimize the expected loss with respect to $F_{\epsilon,G}$ using the logarithmic loss function and the quadratic loss function respectively:

$$\hat{\theta}^{log}_{\epsilon,G} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}\, \mathbb{E}_{F_{\epsilon,G}} \left[\ell_{log}(Y, \boldsymbol{f}(X;\boldsymbol{\theta}))\right],$$
$$\hat{\theta}^{quad}_{\epsilon,G} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}\, \mathbb{E}_{F_{\epsilon,G}} \left[\ell_{quad}(Y, \boldsymbol{f}(X;\boldsymbol{\theta}))\right].$$

First, in Theorem 8, we will show that a corruption G can be adversarially chosen to severely increase the realized bias of the parameter estimate from the logarithmic loss function.

Theorem 8. *Let F be any one-dimensional logistic regression model in Definition 10. For any corruption fraction $\epsilon > 0$, there exists a corruption G that makes the realized bias of the parameter estimate $\hat{\theta}^{log}_{\epsilon,G}$ using the logarithmic loss function greater than or equal to the absolute value of the true parameter θ^*:*

$$\forall\, \epsilon > 0, \exists\, \textit{Corruption } G, \quad \left|\hat{\theta}^{log}_{\epsilon,G} - \theta^*\right| \geq |\theta^*|.$$

We provide the proof of Theorem 8 in Appendix A.9. Theorem 8 directly implies that the logarithmic loss function is not strongly robust with respect to any one-dimensional logistic regression model. This is because for any $U > 0$, Theorem 8 implies that

$$\forall\, \frac{1}{U} > \epsilon > 0, \exists\, \text{Corruption } G, \quad \frac{\left|\hat{\theta}^{log}_{\epsilon,G} - \theta^*\right|}{\epsilon\,|\theta^*|} > U.$$

Therefore, there is no corruption fraction level ϵ' that makes the logistic loss function strongly robust with respect to a one-dimensional logistic regression model. In fact,

we can generalize the result in Theorem 8 to a wider class of models. We will state in Theorem 9 the sufficient condition for a model F that makes the logarithmic loss function not strongly robust with respect to F up to any corruption fraction ϵ'.

Theorem 9. *Consider any model F that satisfies*

$$\forall\, U \in \mathbb{R}^+, \exists\, y \in \mathcal{Y}, \exists\, x \in \mathcal{X}, \forall\, \theta \in \Theta, \quad \left| \frac{d\, f_y(x;\theta)}{d\theta} \frac{1}{f_y(x;\theta)} \right| > U. \tag{6.1}$$

There exists a corruption G that makes the realized bias of the parameter estimate $\hat{\theta}^{log}_{\epsilon,G}$ using the logarithmic loss function greater than or equal to the absolute value of the true parameter θ^*:

$$\forall\, \epsilon > 0, \exists\, \textit{Corruption } G, \quad \left| \hat{\theta}^{log}_{\epsilon,G} - \theta^* \right| \geq |\theta^*|\,.$$

We again provide the proof of Theorem 9 in Appendix A.10. Theorem 8 can also be proven with Theorem 9. It suffices to show that any one-dimensional logistic regression model in Definition 10 satisfies the condition (6.1) of model F in Theorem 9. This is true because for a one-dimensional logistic regression model F and for any $U > 0$,

$$\left| \frac{\partial f_0(x;\theta)}{\partial \theta} \frac{1}{f_0(x;\theta)} \right| = \left| -\frac{x e^{\theta x}}{1 + e^{\theta x}} \right| = x \frac{e^{\theta x}}{1 + e^{\theta x}} \geq \frac{x}{2} > U$$

for any $x > 2U$.

On the other hand, the quadratic loss function does not suffer from the same problem. We show in Theorem 10 that under some mild conditions, the total bias of the estimated parameter $\hat{\theta}^{quad}_{\epsilon,G}$ is bounded by $\mathcal{O}(\epsilon\theta^*)$ for any small corruption fraction ϵ.

Theorem 10. *Let F be any one-dimensional logistic regression model in Definition 10. Assume that F follows Assumption 3 with the lower bound M. Then, for any $\epsilon < \frac{M}{M+8}$, the difference between the true parameter θ^* and the estimated parameter*

using the quadratic loss function $\hat{\theta}^{quad}_{\epsilon,G}$ *is bounded by* $\frac{4\epsilon}{M(1-\epsilon)}\left|\theta^*\right|$*:*

$$\forall\, \epsilon < \frac{M}{M+8}, \forall\, \textit{Corruption } G, \quad \left|\hat{\theta}^{quad}_{\epsilon,G} - \theta^*\right| \leq \frac{4\epsilon}{M(1-\epsilon)}\left|\theta^*\right|.$$

Theorem 10 immediately implies that the quadratic loss function is strongly robust with respect to a one-dimensional logistic regression model up to a corruption fraction of $\frac{M}{M+8}$. This results from the fact that for any $\epsilon < \frac{M}{M+8}$ and any corruption G,

$$\begin{aligned}
\left|\hat{\theta}^{quad}_{\epsilon,G} - \theta^*\right| &\leq \frac{4\epsilon}{M(1-\epsilon)}\left|\theta^*\right| \\
&= \left(\frac{4}{M(1-\epsilon)}\right)\epsilon\left|\theta^*\right| \\
&\leq \left(\frac{4}{M\left(1-\frac{M}{M+8}\right)}\right)\epsilon\left|\theta^*\right| \\
&= \left(\frac{1}{2}+\frac{4}{M}\right)\epsilon\left|\theta^*\right|.
\end{aligned}$$

Thus, the desired upper bound constant U in Definition 9 is $\frac{1}{2}+\frac{4}{M}$.

Chapter 7

Synthetic Data

In this chapter, we will explore different synthetic datasets and study how each of the three strictly consistent loss functions introduced in Chapter 2 behaves in the presence of corruption. We find that for the data-generating processes we consider, it is possible to add a small fraction of corruption to significantly mislead the logarithmic loss function. The resulting parameter estimates can be very biased from the truth, and thus when these parameter estimates are used to optimize other subsequent tasks such as profit maximization, the results are far from optimal. On the other hand, the quadratic loss function and the spherical loss function with $\beta = 2$ do not suffer from the same problem for most of the data-generating processes. We will also show that when the data-generating process violates Assumption 3 for the quadratic loss function introduced in Chapter 5, it is possible to find corruption that biases the quadratic loss function more than the logarithmic loss function. These behaviors are in accordance with the theoretical results in Chapter 5 and 6 which are applicable to the quadratic loss function only when the assumptions are satisfied.

We consider two main data-generating processes: a price-purchase model and a multinomial logit with intercepts model for two products.

7.1 Price-Purchase Model

In the price-purchase model, each data point contains just a price X and a binary Y that indicates whether there is a purchase ($Y = 1$) or not ($Y = 0$). Intuitively, when the price goes up, we expect that the probability of purchase decreases, and vise versa. Specifically in this model, the conditional probability $\mathbb{P}(Y = 1 \mid X) = \sigma(c - \alpha X)$ when σ is a sigmoid function ($\sigma(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$) and α and c are a price sensitivity and an intercept respectively.

We consider two types of corruption scenarios for the price-purchase model. The first scenario is a simple corruption where ϵ of training data points are faulty data points that do not follow the data-generating process. We simply set this ϵ of training data points to have price $X = \tilde{x}$ and label $Y = 1$. The second scenario is when the model is misspecified. All data points are modeled to have the same intercept c when in reality there are ϵ fraction of training data points that are associated with "rich" customer, and thus have a larger intercept $c + \gamma$ (higher purchase probability). Both scenarios fit perfectly into our multunomial-outcome model framework with (ϵ, G)-corruption. We will also see that even for a tiny ϵ, the parameter estimates from the logarithmic loss function are very biased compared to those from the other two loss functions, leading to suboptimal results in the subsequent optimization task. We now explain the two corruption scenarios in details.

7.1.1 Faulty Data

We first formally describe the data-generating process of the price-purchase model. The data point (X, Y) are generated as follows:

$$X \sim Uniform(10, 200),$$
$$Y|X \sim Bernoulli\left(\sigma(c^* - \alpha^* X)\right).$$

Equivalently, the price-purchase model can be described with the notation of the multinomial outcome model. The model can be written as $F = (F_X, \boldsymbol{f}, \theta^*)$ when the

covaritate distribution $F_X = Uniform(10, 200)$, the model parameter $\theta^* = (\alpha^*, c^*)$, and the label probability vector function

$$\boldsymbol{f}(X; \boldsymbol{\theta}) = \begin{bmatrix} 1 - \sigma(c - \alpha X) \\ \sigma(c - \alpha X) \end{bmatrix}.$$

For each iteration, we generate $n = 10000$ independent and identically distributed data points from this data-generating process with the true parameter $\alpha^* = 0.1$ and $c^* = 5.0$. The corruption in this first scenario comes from faulty data points. We change $\epsilon = 0.01$ fraction of data points to all have price $X = \tilde{x}$ and label $Y = 1$ when $\tilde{x}$ is a fixed price that ranges from $10, 20, \ldots, 200$ in each separated iteration. This corruption can be seen as (ϵ, G)-corruption when G is a Dirac distribution that concentrates at the point $X = \tilde{x}$ and $Y = 1$. Then we estimate the model parameter (α, c) by performing loss minimization using the three strictly consistent loss functions. We repeat 200 iterations of parameter estimation process for each different $\tilde{x}$ and each loss function. The mean of the parameter estimates α and c from different loss functions are shown in Figure 7-1 with 95% confidence interval calculated using bootstrapping.

Intuitively, the higher the faulty price $\tilde{x}$, the more corrupted that data point is because the probability of purchasing decreases as the price increases. We thus expect that the parameter estimates will be affected more as the faulty price $\tilde{x}$ increases. As show in Figure 7-1, this is exactly the case for the parameter estimates from the logarithmic loss function. The parameter estimates diverge from the true parameters when the values of the corrupted price $\tilde{x}$ become larger even though the corruption fraction remains at $\epsilon = 0.01$. On the other hand, the quadratic loss function and the spherical loss function with $\beta = 2$ do not have such behavior, demonstrating that they are more robust to faulty training data than the logarithmic loss function. The parameter estimates using the quadratic loss function and the spherical loss function with $\beta = 2$ move only slightly from the true values regardless of the value of $\tilde{x}$. These results agree with the partial robustness property of the three loss functions. The quadratic loss function and the spherical loss function with $\beta = 2$ are shown to be

partially robust in Chapter 5. The 2-norm of the influence function of the quadratic loss function is also bounded according to Theorem 7.

We further demonstrate that the robustness property of the spherical loss function agrees with the theoretical results we have shown. We repeat the experiment of parameter estimation from faulty data, but instead use only the spherical loss functions with β values in the list $[1.01, 1.05, 1.1, 1.2, 1.3, 1.4, 1.5, 2, 3]$. The behaviors of the parameter estimates are shown in 7-2. The spherical loss functions become more robust as β increases. They are less robust when parameter β converges to 1 such as when $\beta = 1.01$ and $\beta = 1.05$. These results again agree with Theorem 5 which guarantees that the spherical loss functions become partially robust when $\beta \geq 2$.

In addition to demonstrating that the parameter estimates from the logarithmic loss function are more biased that those from the other two loss functions, we also show that the biased parameter estimates lead to prices that create suboptimal profits when used for profit maximization tasks. Given a parameter estimate $\hat{\theta} = (\hat{\alpha}, \hat{c})$, we can find a price that maximizes profit for the price-purchase model by solving

$$\underset{p}{\text{argmax}} \quad \sigma(\hat{c} - \hat{\alpha}p)(p - t), \tag{7.1}$$

when t is the cost of the product. We assume that the cost of the product is $t = 50$. Again, we perform 200 iterations of parameter estimation for each loss function and each corrupted price $\tilde{x}$. The resulting parameter estimates are used to calculate optimal prices that maximize the term in Eq. (7.1). The realized expected profit at price p is simply $\sigma(c^* - \alpha^* p)(p - 50)$. The results for each of the three loss functions are visualized in Figure 7-3. The profits associated with the logarithmic loss function keep decreasing as the corrupted price $\tilde{x}$ increases. The profits associated with the quadratic loss function and the spherical loss function with $\beta = 2$ are almost always around the optimal profit and only drop a little bit when the corrupted price $\tilde{x}$ is between 50 and 100. These results confirm that the logarithmic loss function is not robust to the corruption even with a tiny corrupted fraction ϵ, resulting in biased parameter estimates that lead to suboptimal results in subsequent optimization tasks.

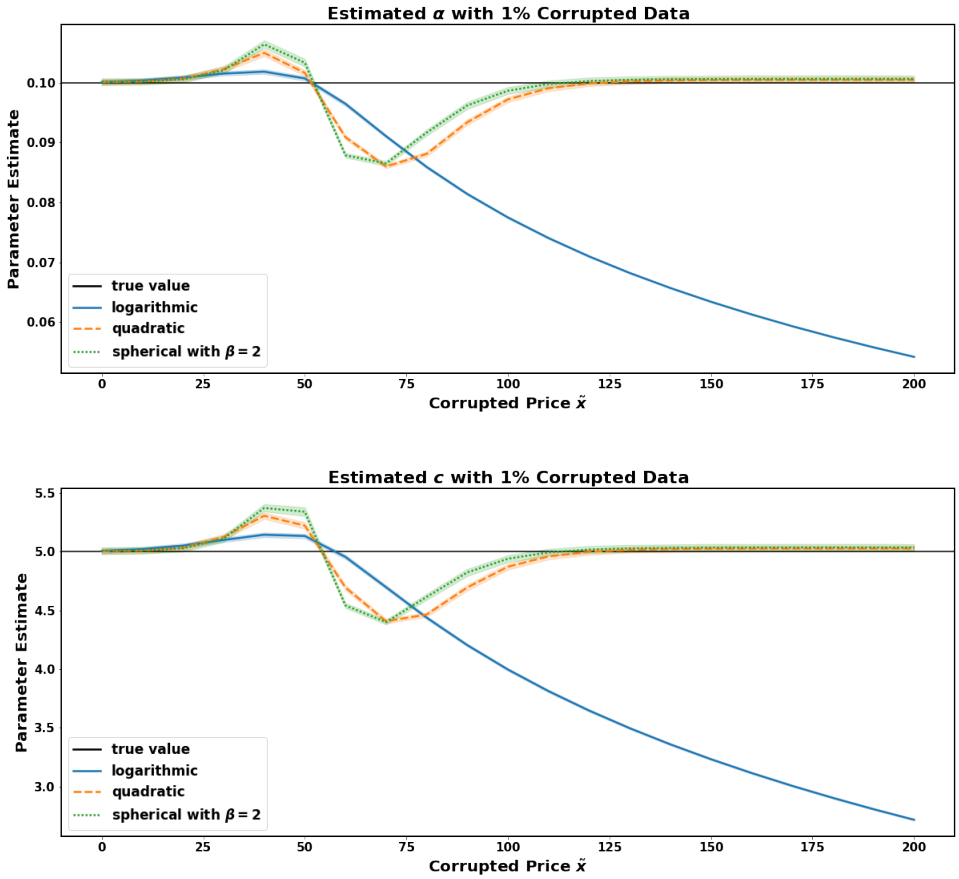

Figure 7-1: The parameter estimates of price sensitivity α and intercept c in the price-purchase model with 1% corrupted training data. For each iteration, we generate 10000 data points from the data-generating process of the price-purchase model and corrupt $\epsilon = 0.01$ fraction of data points with faulty data. The faulty data points are all set to have price $\tilde{x}$ and label 1 when $\tilde{x}$ ranges from 10 to 200. The process is repeated for 200 iterations and the mean parameter estimates are plotted with 95% bootstrapping confidence interval.

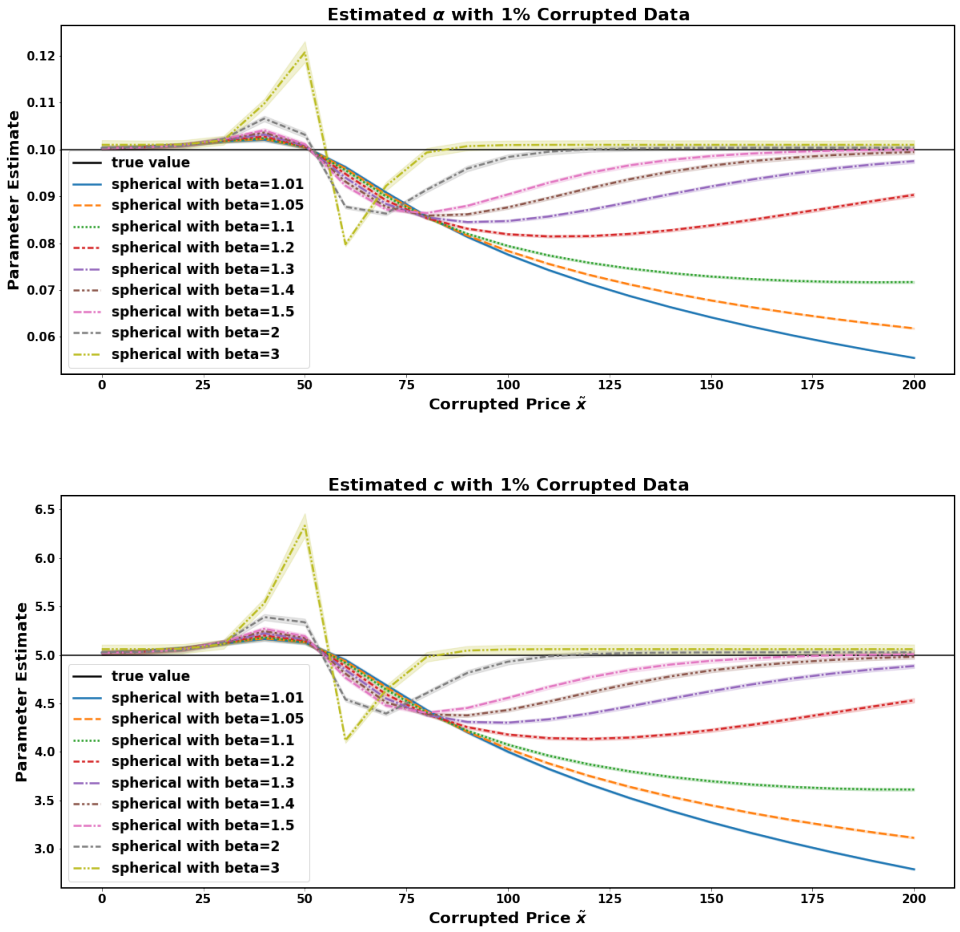

Figure 7-2: The parameter estimates of price sensitivity α and intercept c in the price-purchase model with 1% corrupted training data. The parameter β of the spherical loss function are set to different values in $[1.01, 1.05, 1.1, 1.2, 1.3, 1.4, 1.5, 2, 3]$ to demonstrate how the degree of robustness depends on β. For each iteration, we generate 10000 data points from the data-generating process of the price-purchase model and corrupt $\epsilon = 0.01$ fraction of data points with faulty data. The faulty data points are all set to have price $\tilde{x}$ and label 1 when $\tilde{x}$ ranges from 10 to 200. The process is repeated for 200 iterations and the mean parameter estimates are plotted with 95% bootstrapping confidence interval.

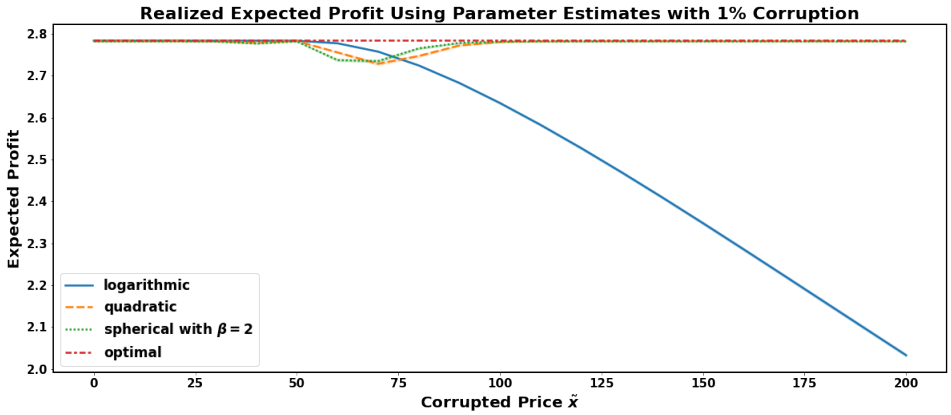

Figure 7-3: The realized expected profit of the price-purchase model using parameter estimates from the three loss functions in the presence of 1% corrupted training data. For each iteration, we generate 10000 data points from the data-generating process of the price-purchase model and corrupt $\epsilon = 0.01$ fraction of data points with faulty data. The faulty data points are all set to have price $\tilde{x}$ and label 1 when $\tilde{x}$ ranges from 10 to 200. The parameter estimates using these data points are used to estimate the optimal prices for the product assuming that the cost is 50. The process is repeated for 200 iterations and the mean of the realized profit are plotted with 95% bootstrapping confidence interval.

7.1.2 Misspecification

Instead of the corruption through the faulty data, we consider a scenario when the model is misspecified. Assume that in reality, a small ϵ fraction of the training data points involve rich customers who have higher purchase probability than average customers. The difference in the purchase probability is captured by the difference in the intercepts. The purchase probability remains as the sigmoid of the product of the price sensitivity and the price subtracted from the intercept for both rich and average customers. A rich customer, however, has a higher intercept $(c+\gamma)$ than that of an average customer (c). We can explain the correctly specified model through the following data-generating process. For each data point, there is an unobserved attribute R which indicates whether the customer is rich $(R = 1)$ or not $(R = 0)$.

The variables X, R, Y are generated as follows:

$$\begin{aligned} X &\sim Uniform(10, 200), \\ R &\sim Bernoulli(\epsilon), \\ Y|X, R &\sim Bernoulli\left(\sigma(c - \alpha X + \gamma R)\right), \end{aligned}$$

when we refer to γ as a rich coefficient and ϵ as a misspecification fraction.

We again set the true intercept to be $c^* = 5.0$ and the true price sensitivity to be $\alpha^* = 0.1$. We vary the values of rich coefficient γ and misspecification fraction ϵ to study their effects to the parameter estimates and the realized expected profits that use optimal prices calculated from the parameter estimates. Similar to the results in the corruption from faulty data in the first scenario, the logarithmic loss function produces parameter estimates that are very biased, both for the price sensitivity α and the intercept c, whereas the quadratic loss function and the spherical loss function with $\beta = 2$ produce parameter estimates that are biased only slightly. The mean of the parameter estimates are shown in Figure 7-4 and 7-5 for the misspecification fraction ϵ of 0.01 and 0.05 respectively. We can see that the biases in the quadratic loss function and the spherical loss function with $\beta = 2$ remain relatively the same regardless of the value of γ. The bias in the logarithmic loss function however increases as γ increases and is consistently higher than the biases from the other two loss functions.

We next consider the impact of misspecification on the profit maximization. This scenario is fundamentally different from the first scenario of faulty data. Because the ϵ fraction of data points come from the misspecification of the model, not from the external corruption, we have to also take these ϵ fraction into account when we optimize for the profit. Thus, we have to instead find the optimal price by optimizing

$$\underset{p}{\text{argmax}} \quad (1 - \epsilon)\sigma(c - \alpha p)(p - t) + \epsilon\sigma(c - \alpha p + \gamma)(p - t),$$

when t is the cost of the product. Again, we assume that the cost is $t = 50$. Note that different γ and ϵ produce different optimal prices and optimal profits. When γ

or ϵ becomes larger, the optimal price and the optimal profit also increase because there is more buying power from the rich customers. The results are shown in Figure 7-6 and 7-7. The quadratic loss function and the spherical loss function with $\beta = 2$ consistently achieve near-optimal profit for all values of ϵ and γ. The logarithmic loss function, on the other hand, achieves realized profits that diverge from the optimal as γ increases for every misspecification fraction ϵ, as seen in Figure 7-6, where we fix ϵ and vary γ. This behavior is confirmed in Figure 7-7 where we instead fix γ and vary ϵ. Every line plot is shown with 95% confidence interval calculated from bootstrapping of the values from 200 independent iterations.

7.2 Multinomial Logit with Intercepts

We now investigate a model that is more realistic than the price-purchase model. We consider a multinomial logit model with intercepts that has three categorical labels instead of the two binary labels in the price-purchase model. Consider a discrete choice model of two similar products; the first product has higher quality, is more costly, and is more popular, while the second product has lower quality, costs less, and rarely gets bought. We will simulate historical transaction data where for each transaction, a customer sees prices of these two products, p_1 and p_2, and chooses to buy the first product ($Y = 1$), the second product ($Y = 2$), or an outside option ($Y = 0$).

We use multinomial logit model as the underlying choice mechanism for this discrete choice model. For a data point with covariate $\boldsymbol{x} = (p_1, p_2)$, the probability of choosing the outside option, the first product, and the second product respectively are

$$
\begin{aligned}
f_0(\boldsymbol{x}; \boldsymbol{\theta}) &= \frac{1}{1 + \exp(c_1 - \alpha \log(p_1)) + \exp(c_2 - \alpha \log(p_1))}, \\
f_1(\boldsymbol{x}; \boldsymbol{\theta}) &= \frac{\exp(c_1 - \alpha \log(p_1))}{1 + \exp(c_1 - \alpha \log(p_1)) + \exp(c_2 - \alpha \log(p_2))}, \\
f_2(\boldsymbol{x}; \boldsymbol{\theta}) &= \frac{\exp(c_2 - \alpha \log(p_2))}{1 + \exp(c_1 - \alpha \log(p_1)) + \exp(c_2 - \alpha \log(p_2))},
\end{aligned}
$$

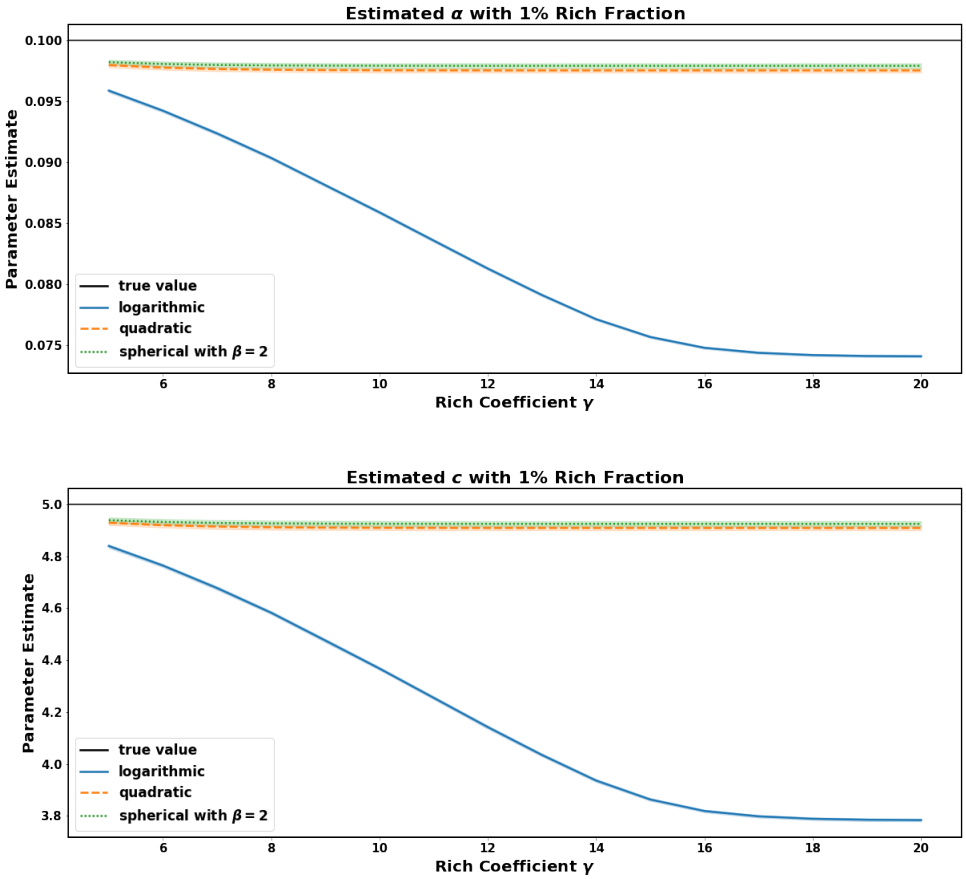

Figure 7-4: The parameter estimates of price sensitivity α and intercept c in the price-purchase model that is misspecified by not taking rich coefficient into account. For each iteration, we generate 10000 data points from the data-generating process of the price-purchase model with $\epsilon = 0.01$ fraction of rich customers. The rich coefficient γ is varied from 5.0 to 20.0. The process is repeated for 200 iterations and the mean parameter estimates are plotted with 95% bootstrapping confidence interval.

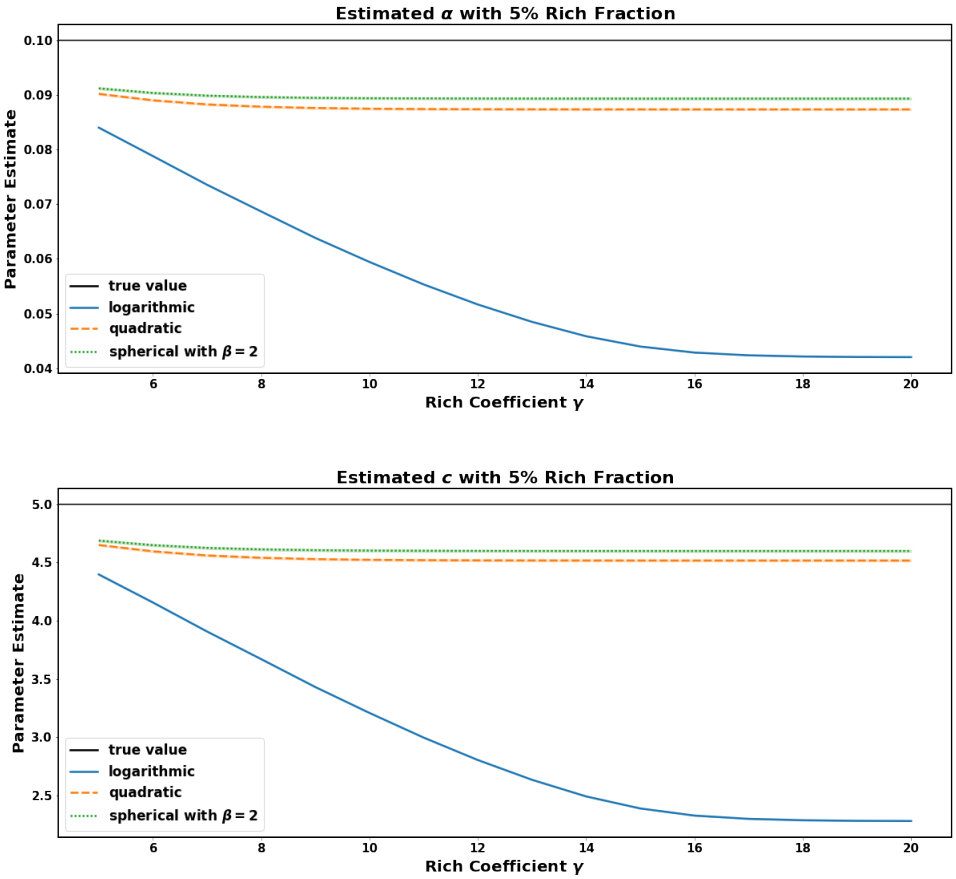

Figure 7-5: The parameter estimates of price sensitivity α and intercept c in the price-purchase model that is misspecified by not taking rich coefficient into account. For each iteration, we generate 10000 data points from the data-generating process of the price-purchase model with $\epsilon = 0.05$ fraction of rich customers. The rich coefficient γ is varied from 5.0 to 20.0. The process is repeated for 200 iterations and the mean parameter estimates are plotted with 95% bootstrapping confidence interval.

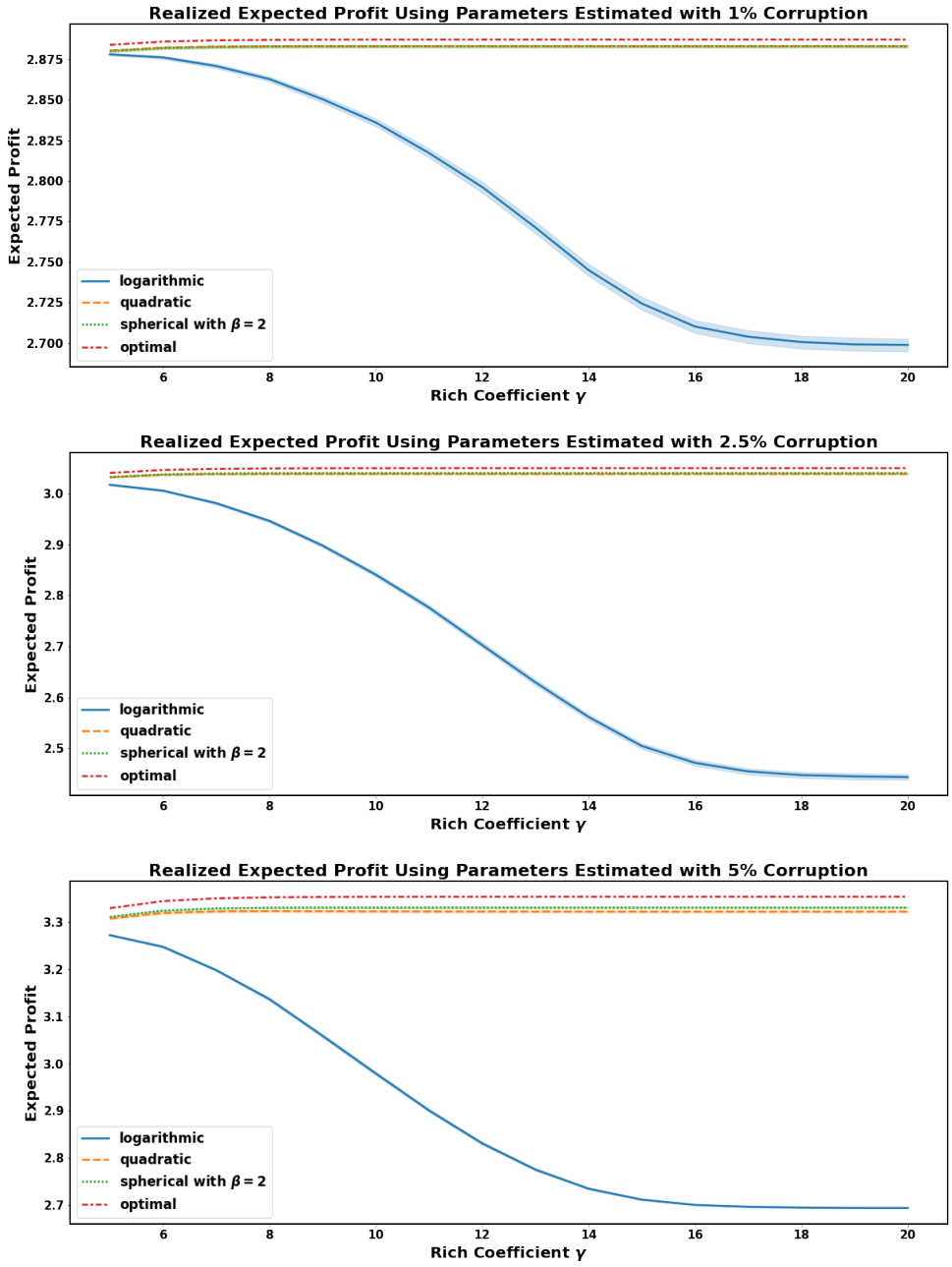

Figure 7-6: The realized expected profit of the price-purchase model using parameter estimates from the three loss functions when the model is misspecified by not taking rich coefficient into account. For each iteration, we generate 10000 data points from the data-generating process of the price-purchase model with ϵ fraction of rich customers. The rich coefficient γ is varied from 5.0 to 20.0. The parameter estimates using these data points are used to estimate the optimal prices for the product assuming that the product costs 50. The process is repeated for 200 iterations and the mean of the realized profit are plotted with 95% bootstrapping confidence interval. The three plots are for $\epsilon = 0.01, 0.25$ and 0.5 respectively from the top to the bottom.

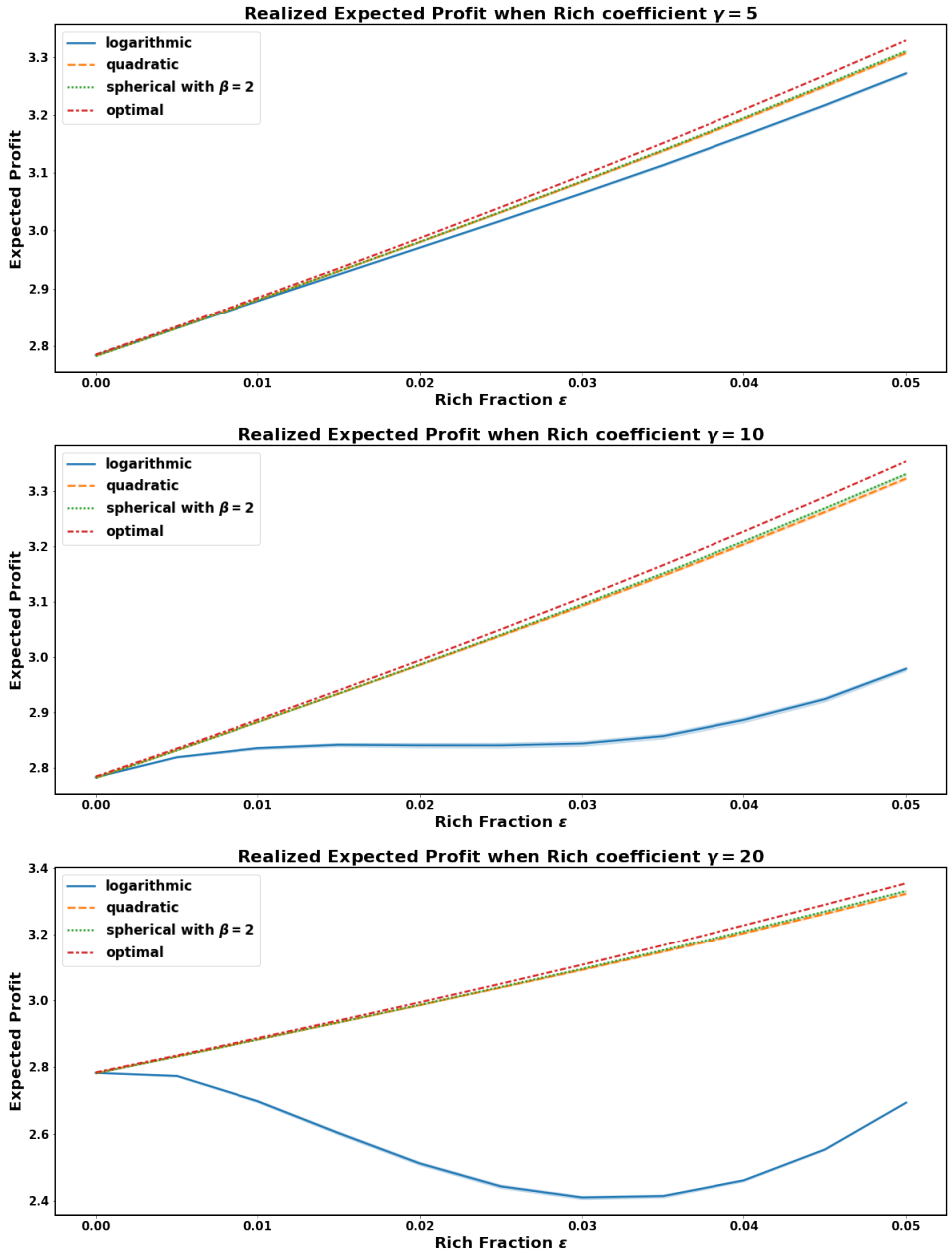

Figure 7-7: The realized expected profit of the price-purchase model using parameter estimates from the three loss functions when the model is misspecified by not taking rich coefficient into account. For each iteration, we generate 10000 data points from the data-generating process of the price-purchase model with fixed rich coefficient γ and varying ϵ fraction of rich customers. The misspecification fraction ϵ is varied from 0.005 to 0.05. The parameter estimates using these data points are used to estimate the optimal prices for the product assuming that the product costs 50. The process is repeated for 200 iterations and the mean of the realized profit are plotted with 95% bootstrapping confidence interval. The three plots are for $\gamma = 5.0, 10.0$ and 20.0 respectively from the top to the bottom.

when $\boldsymbol{\theta} = (\alpha, c_1, c_2)$ is the model parameters that need to be estimated. The parameter α is the price sensitivity that reflects a decrease in a chance of purchasing when a price increases. The price sensitivity are shared across both the first and the second products. The parameter c_1 and c_2 are the intercepts of the first and the second product respectively. They reflect the differences in the popularity of the two products and the outside option.

Similar to the use case of the parameter estimates in the price-purchase model, we can use the parameter estimates to determine the product prices that maximize the expected profit in the multinomial logit model. Given a parameter estimate $\boldsymbol{\theta}$, the optimal prices of products that maximize the expected profit are

$$\underset{\boldsymbol{x}=(p_1,p_2)}{\operatorname{argmax}} \quad f_1(\boldsymbol{x};\boldsymbol{\theta})(p_1 - t_1) + f_2(\boldsymbol{x};\boldsymbol{\theta})(p_2 - t_2), \tag{7.2}$$

when t_1 and t_2 are the costs of the first and the second product respectively. Note again that given estimated optimal prices $\hat{p}_1, \hat{p}_2$, the realized expected profit is

$$f_1((\hat{p}_1, \hat{p}_2);\boldsymbol{\theta}^*)(\hat{p}_1 - t_1) + f_2((\hat{p}_1, \hat{p}_2);\boldsymbol{\theta}^*)(\hat{p}_2 - t_2).$$

We consider two different sets of settings for this multinomial logit model. In the first setting, we set the true parameter and the cost of each product at reasonable values. The resulting marginal purchase probabilities are reasonable at 29.6% for the first product and at 0.4% for the second product. The rest goes to the outside option. In this first setting, we also consider two types of corruption. On one hand, we corrupt both prices and labels of ϵ fraction of data points. On the other hand, we corrupt only the label, which is more realistic to happen in the real world.

The second setting is a pedagogical example that shows when the quadratic loss function fails. We set the true parameter and the product costs at unreasonably extreme values in order to violate the Assumption 3 for the quadratic loss function. The second setting demonstrates that when the assumption is violated, the quadratic loss function is not necessary more robust than the quadratic loss function.

Note that in all settings, we generate 100000 training data points to be used in parameter estimation. We also repeat all processes for 200 independent iterations to calculate the 95% confidence interval by bootstrapping.

7.2.1 The Standard Setting

We let the cost be $t_1 = 20$ and $t_2 = 15$, and let the true parameter be

$$\boldsymbol{\theta}^* = (\alpha^*, c_1^*, c_2^*) = (3.0, 10.0, 5.0).$$

The optimal prices given this true parameter, calculated by optimizing Equation (7.2) with Nelder-Mead optimization method, are $p_1^* = 37.0$ and $p_2^* = 30.0$. For each observation, we randomly sample prices p_1 and p_2 with equal probability from the price lists centered at p_1^* and p_2^*:

$$\begin{aligned} p_1 &\sim [0.85p_1^*, 0.9p_1^*, \ldots, 1.1p_1^*, 1.15p_1^*], \\ p_2 &\sim [0.85p_2^*, 0.9p_2^*, \ldots, 1.1p_2^*, 1.15p_2^*]. \end{aligned} \tag{7.3}$$

Another way to interpret these price distributions is to consider when the prices p_1 and p_2 of the first and the second product independently appreciate or depreciate in the range of $-15\%, -10\%, \ldots, +10\%, +15\%$ from their corresponding optimal prices. With this choice of the true parameters, the costs and the price distribution, the expected probability vector $\boldsymbol{f}(X; \boldsymbol{\theta})$ is $(0.700, 0.296, 0.004)$. That is, on average, purchases happen only in 30% of all observations. For those observations with purchase, 98.67% of the purchases go to the first product while only 1.33% of all purchases go to the unpopular second product. We now consider two types of corruption.

Price and Label Corruption. We corrupt the training data by randomly selecting ϵ fraction of all data, changing label from 0 (the outside option) to 2 (the unpopular product), and changing prices to $p_1 = 0.85p_1^*$ and $p_2 = 1.15p_2^*$. That is, for each corrupted data point, the price of the first product is set at the lowest among the existing prices for the first product, while the price of the second product is set at

the highest among the existing prices for the second product. These data points are also set to have label 2, meaning that the second unpopular product is still selected despite being less popular and having a higher price than usual.

As shown in Figure 7-9, we can see that the effect of the corruption is devastating for the logarithmic loss function. As the corruption fraction ϵ goes up to 0.05, the logarithmic loss function produces parameter estimates that are much more biased than those from the quadratic loss function and the spherical loss function with $\beta = 2$. This holds true for all parameter α, c_1 and c_2. As seen in Figure 7-8, the pattern also holds for the 2-norm of the bias of the parameter vector, $\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2$. The 2-norm of the bias from the logarithmic loss function increases at a much higher rate than the other two loss functions as the corruption fraction ϵ increases. This agrees with the theoretical result in Theorem 7 which states that the 2-norm of the influence function is bounded for the quadratic loss function.

Finally, we compare the realized expected profit when the parameter estimates are used to calculate the optimal prices by optimizing Eq. (7.2). The results are shown in 7-10. Again, the logarithmic loss function is severely affected by the corruption that the realized expected profits drop rapidly compared to those from the other two loss functions, from which the profits barely drop. The expected profits from the logarithmic loss function even decrease to 0 after the corruption fraction ϵ becomes greater than 0.025. This again reiterates that the logarithmic loss function must be used carefully in the presence of corruption.

Label Only Corruption. The corruption where both prices and label are corrupted might rarely happen in the real world. However, we will show that even if we corrupt only the label of a tiny fraction of the training data, the logarithmic loss function still ends up performing worse than the other loss functions. Specifically we randomly select ϵn observation with label 0 (the outside option) and change it to 2 (the unpopular product). This label corruption is much more realistic and can happen in the real data in many scenarios. For example, when the first popular product is out of stock, the second unpopular product might be selected instead even though it has a lower purchase probability in the model. Another example is when a tiny

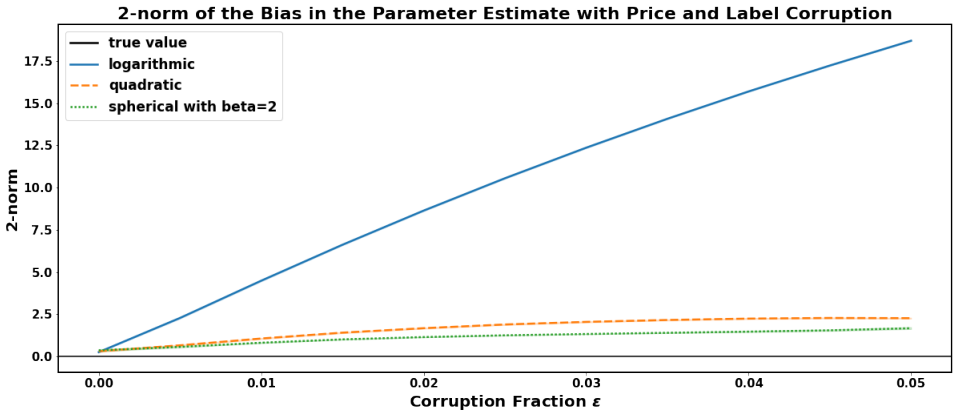

Figure 7-8: The 2-norm bias of the parameter estimates from the three loss functions in the multinomial logit model when the prices and labels of ϵ fraction of the training data are corrupted. For each iteration, 100000 data points are generated from the multinomial logit model with two products with true parameter $\alpha = 3.0, c_1 = 10.0, c_2 = 5.0$. Then, ϵ fraction of the data points have labels changed from 0 (the outside option) to 2 (the unpopular product) and prices set at $p_1 = 0.85p_1^*$ and $p_2 = 1.15p_2^*$. The corruption fraction ϵ is varied from $0.00, 0.005, \ldots, 0.05$. The process is repeated for 200 iterations and the mean parameter estimates are plotted with 95% bootstrapping confidence interval.

fraction of the data is mislabeled from other labels to label 2 (the unpopular product) due to human error.

Because the corruption in this case is relatively weaker, the logarithmic loss function does not perform as poorly as when we corrupt both the price and the label. The quadratic loss function and the spherical loss function still nevertheless perform better than the logarithmic loss function. We can see in Figure 7-11 that the bias from the logarithmic loss function is higher for the parameter α and c_1. The bias for the parameter c_2, however, is smaller from the logarithmic loss function than from the other two loss functions. As a result, the 2-norms of the bias vectors from all the three loss functions, as shown in Figure 7-12, are about at the same level, with the logarithmic loss function performing relatively better. Nevertheless, when we instead consider the realized expected profits from the parameter estimates in Figure 7-13, the logarithmic loss function performs the worst. This is due to the fact that the price sensitivity α is shared across all products and is thus more important for price

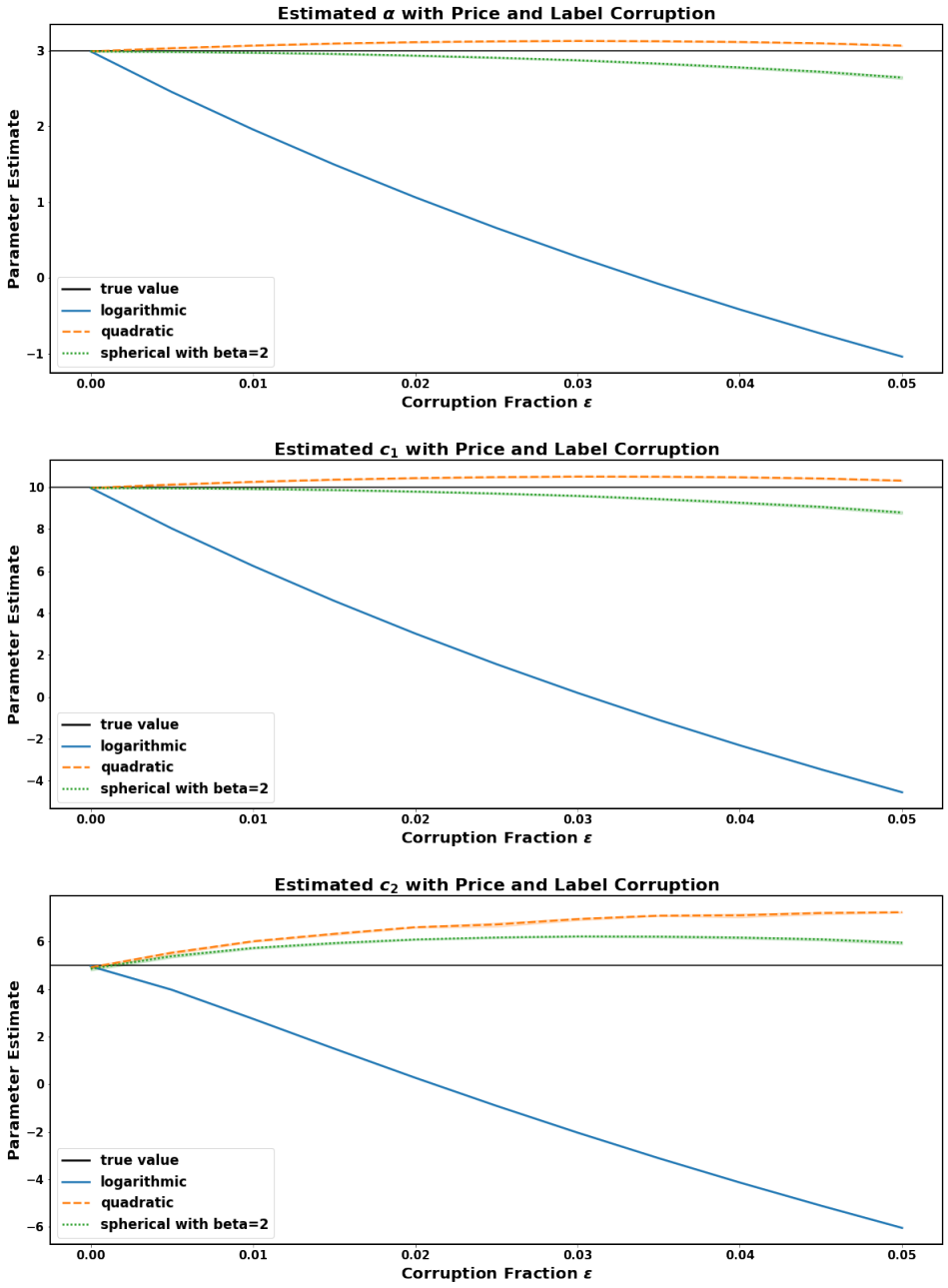

Figure 7-9: The parameter estimates of price sensitivity α and two intercepts c_1 and c_2 from the three loss functions in the multinomial logit model when the prices and labels of ϵ fraction of the training data are corrupted. For each iteration, 100000 data points are generated from the multinomial logit model with two products with true parameter $\alpha = 3.0, c_1 = 10.0, c_2 = 5.0$. Then, ϵ fraction of the data points have labels changed from 0 (the outside option) to 2 (the unpopular product) and prices set at $p_1 = 0.85p_1^*$ and $p_2 = 1.15p_2^*$. The corruption fraction ϵ is varied from $0.00, 0.005, \ldots, 0.05$. The process is repeated for 200 iterations and the mean parameter estimates are plotted with 95% bootstrapping confidence interval.

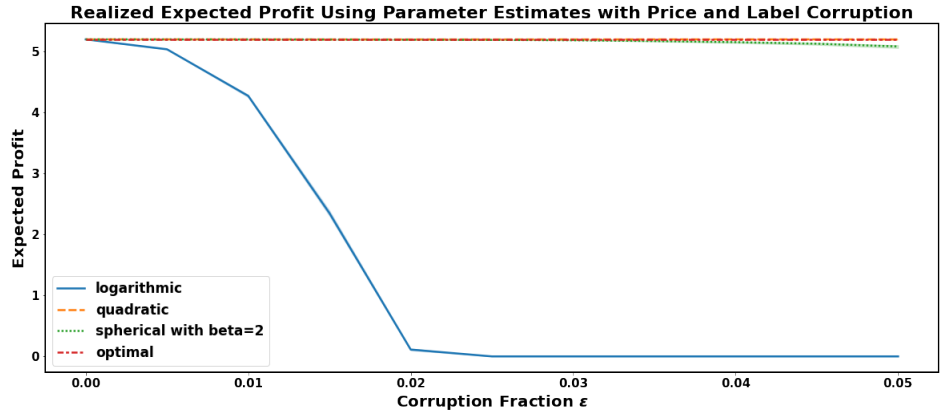

Figure 7-10: The realized expected profit of the multinomial logit model using parameter estimates from the three loss functions when the prices and labels of ϵ fraction of the training data are corrupted. For each iteration, 100000 data points are generated from the multinomial logit model with two products with true parameter $\alpha = 3.0, c_1 = 10.0, c_2 = 5.0$. Then, ϵ fraction of the data points have labels changed from 0 (the outside option) to 2 (the unpopular product) and prices set at $p_1 = 0.85p_1^*$ and $p_2 = 1.15p_2^*$. The parameter estimates using these data points are used to estimate the optimal prices for the two products assuming that the products cost $t_1 = 20$ and $t_2 = 15$. The corruption fraction ϵ is varied from $0.00, 0.005, \ldots, 0.05$. The process is repeated for 200 iterations and the mean parameter estimates are plotted with 95% bootstrapping confidence interval.

optimization task. The logarithmic loss function that has a much higher bias in the parameter estimate of α therefore performs worse than the other two loss functions.

7.2.2 The Pedagogical Setting: When the Quadratic Fails

In this setting, we will show that the quadratic loss function is not always more robust against the corruption. This happens when the model distribution doesn't follow Assumption 3, which we use to prove the theoretical robustness properties of the quadratic loss function. In order to make the model distribution violate the assumption, we set the true model parameter for the multinomial logit model to extreme values at

$$\theta^* = (\alpha^*, c_1^*, c_2^*) = (5.0, 16.0, 9.0),$$

and we set the costs of each product at $t_1 = 55$ and $t_2 = 30$. With this choice of the true parameter and the costs, the optimal prices calculated by optimizing Eq. (7.2) are $p_1^* = 68.85$ and $p_2^* = 37.60$. We again set the price distribution to center around p_1^* and p_2^* exactly like in Eq. (7.3) in the standard setting. The resulting expected probability vector $\boldsymbol{f}(X;\theta)$ is $(0.9932, 0.0067, 0.0001)$, meaning that the outside option is almost always chosen. Note that in this setting, we set the corruption fraction ϵ to range from $0.0000, 0.0005, \ldots, 0.0050$, which are 10 times smaller than those used in the previous setting.

The results are the opposite of those we have seen previously. Even with only 0.05% corruption, the parameter estimates from the quadratic loss function are already very biased at a much higher level than the logarithmic loss function, as seen in Figure 7-15. The quadratic loss function consistently has a higher bias level than the logarithmic loss function for both the parameter α and c_1. For the parameter c_2, the biases of different loss functions have different signs, and the magnitudes are harder to compare. Nevertheless, when we instead consider the 2-norm of the bias vector shown in Figure 7-14, the 2-norm of the bias is consistently larger for the quadratic loss function than for the logarithmic loss function. The same can be said for the

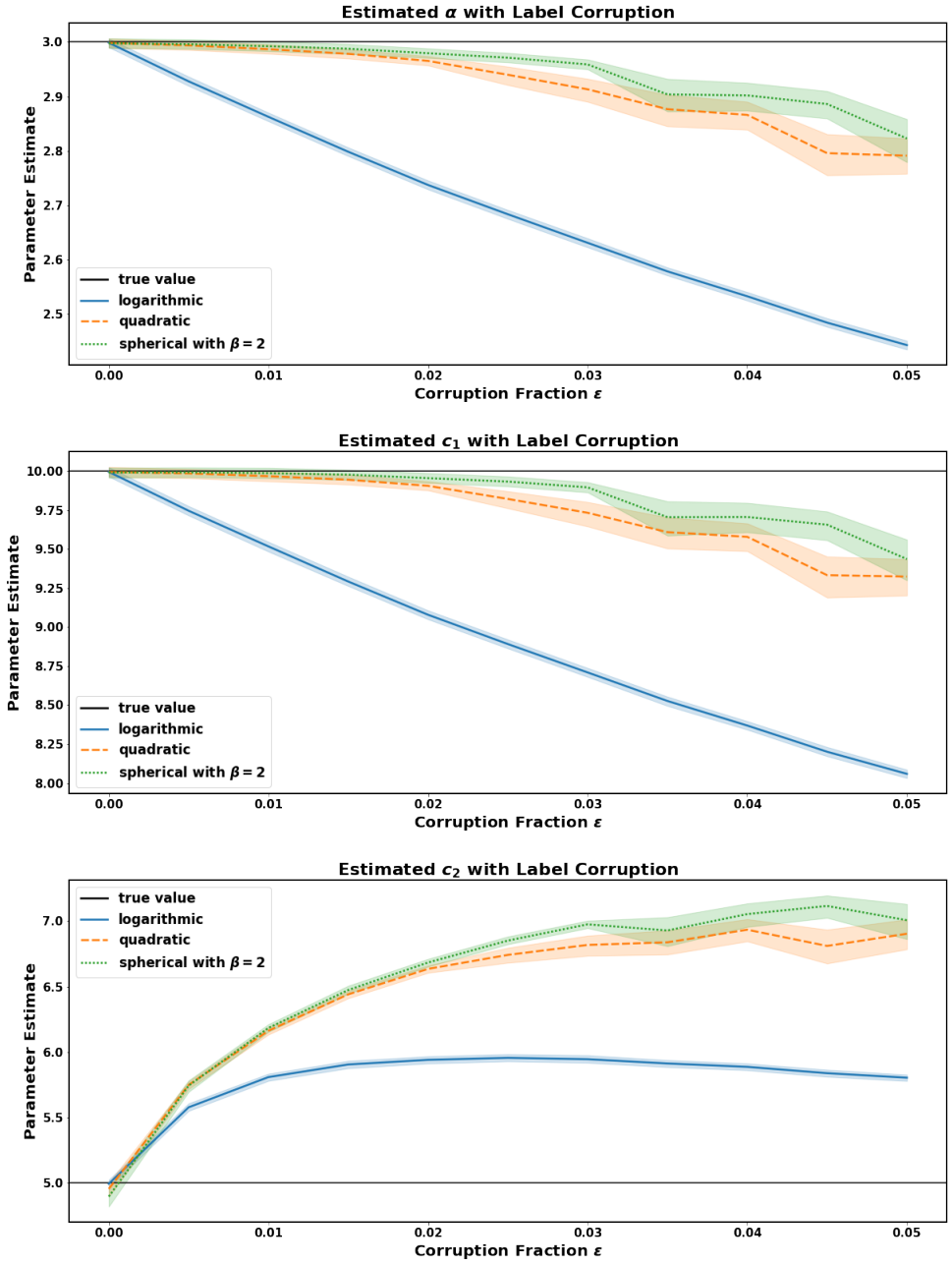

Figure 7-11: The parameter estimates of price sensitivity α and two intercepts c_1 and c_2 from the three loss functions in the multinomial logit model when the labels of ϵ fraction of training data are corrupted. For each iteration, 100000 data points are generated from the multinomial logit model with two products with true parameter $\alpha = 3.0, c_1 = 10.0, c_2 = 5.0$. Then, ϵ fraction of the data points have labels changed from 0 (the outside option) to 2 (the unpopular product). The corruption fraction ϵ is varied from $0.00, 0.005, \ldots, 0.05$. The process is repeated for 200 iterations and the mean parameter estimates are plotted with 95% bootstrapping confidence interval.

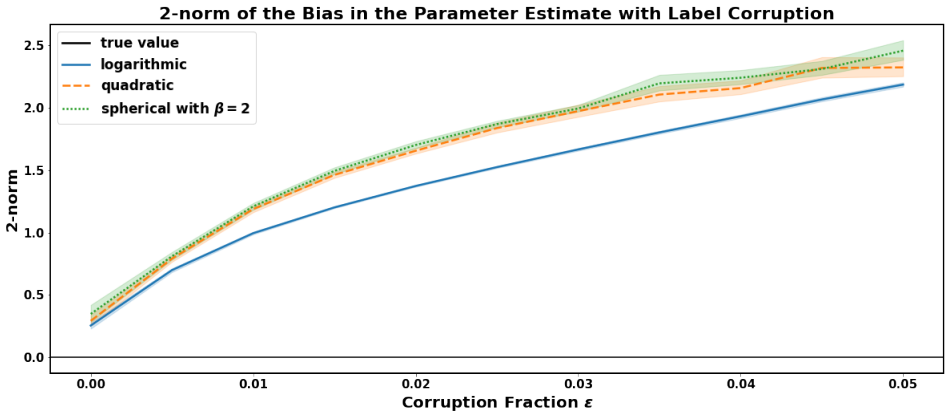

Figure 7-12: The 2-norm of the bias of the parameter estimates from the three loss functions in the multinomial logit model when the labels of ϵ fraction of the training data are corrupted. For each iteration, 100000 data points are generated from the multinomial logit model with two products with true parameter $\alpha = 3.0, c_1 = 10.0, c_2 = 5.0$. Then, ϵ fraction of the data points have labels changed from 0 (the outside option) to 2 (the unpopular product). The corruption fraction ϵ is varied from $0.00, 0.005, \ldots, 0.05$. The process is repeated for 200 iterations and the mean parameter estimates are plotted with 95% bootstrapping confidence interval.

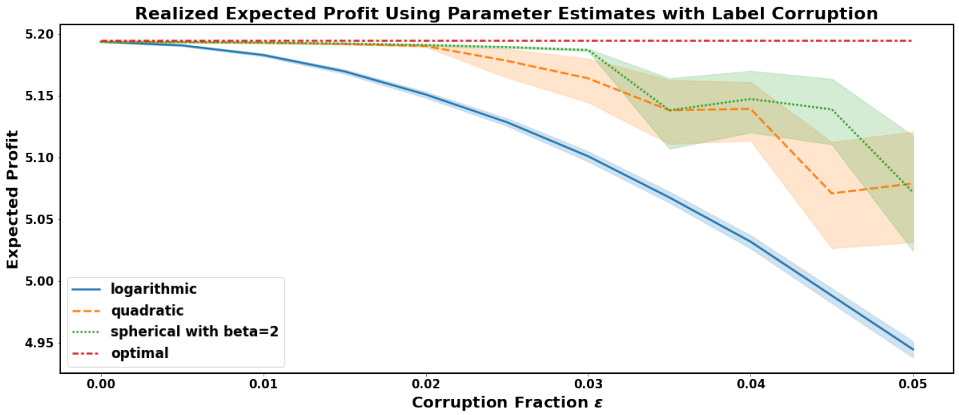

Figure 7-13: The realized expected profit of the multinomial logit model using parameter estimates from the three loss functions when the labels of ϵ fraction of the training data are corrupted. For each iteration, 100000 data points are generated from the multinomial logit model with two products with true parameter $\alpha = 3.0, c_1 = 10.0, c_2 = 5.0$. Then, ϵ fraction of the data points have labels changed from 0 (the outside option) to 2 (the unpopular product). The parameter estimates using these data points are used to estimate the optimal prices for the two products assuming that the products cost $t_1 = 20$ and $t_2 = 15$. The corruption fraction ϵ is varied from $0.00, 0.005, \ldots, 0.05$. The process is repeated for 200 iterations and the mean parameter estimates are plotted with 95% bootstrapping confidence interval.

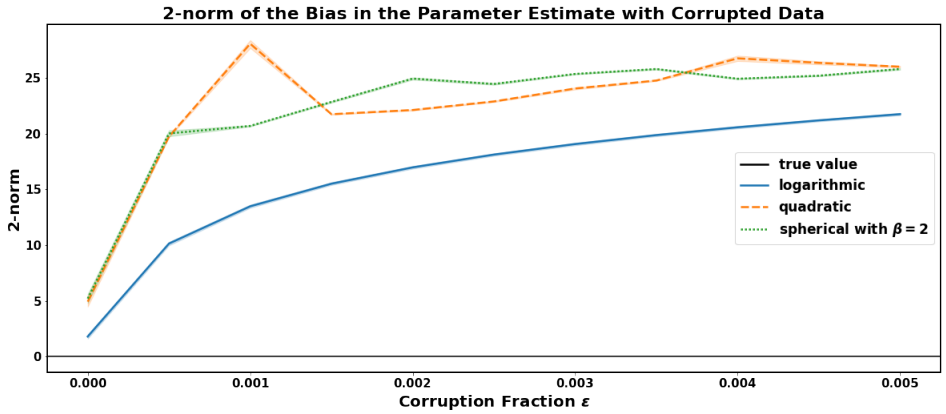

Figure 7-14: The 2-norm of the bias of the parameter estimates from the three loss functions in the multinomial logit model with extreme parameters that make the model violate the assumption for the quadratic loss function. For each iteration, 100000 data points are generated from the multinomial logit model with two products with true parameter $\alpha = 5.0, c_1 = 16.0, c_2 = 9.0$. Then, ϵ fraction of the data points are randomly chosen and have their prices set to $p_1 = 27, p_2 = 37$ and labels set to 1 (the popular product). The corruption fraction ϵ is varied from $0.00, 0.0005, \ldots, 0.005$. The process is repeated for 200 iterations and the mean parameter estimates are plotted with 95% bootstrapping confidence interval.

realized expected profits that use the parameter estimates as inputs for the price optimization problem. As seen in Figure 7-16, the performance of the quadratic loss function is worse than that of the logarithmic loss function for all corruption fraction ϵ, a complete opposite behavior compared to those in the standard setting.

Note that the spherical loss function with $\beta = 2$ happens to behave very similar to the quadratic loss function in this setting. Unlike the quadratic loss function, we do not yet have a theoretical justification as to why the spherical loss function also fails in this setting.

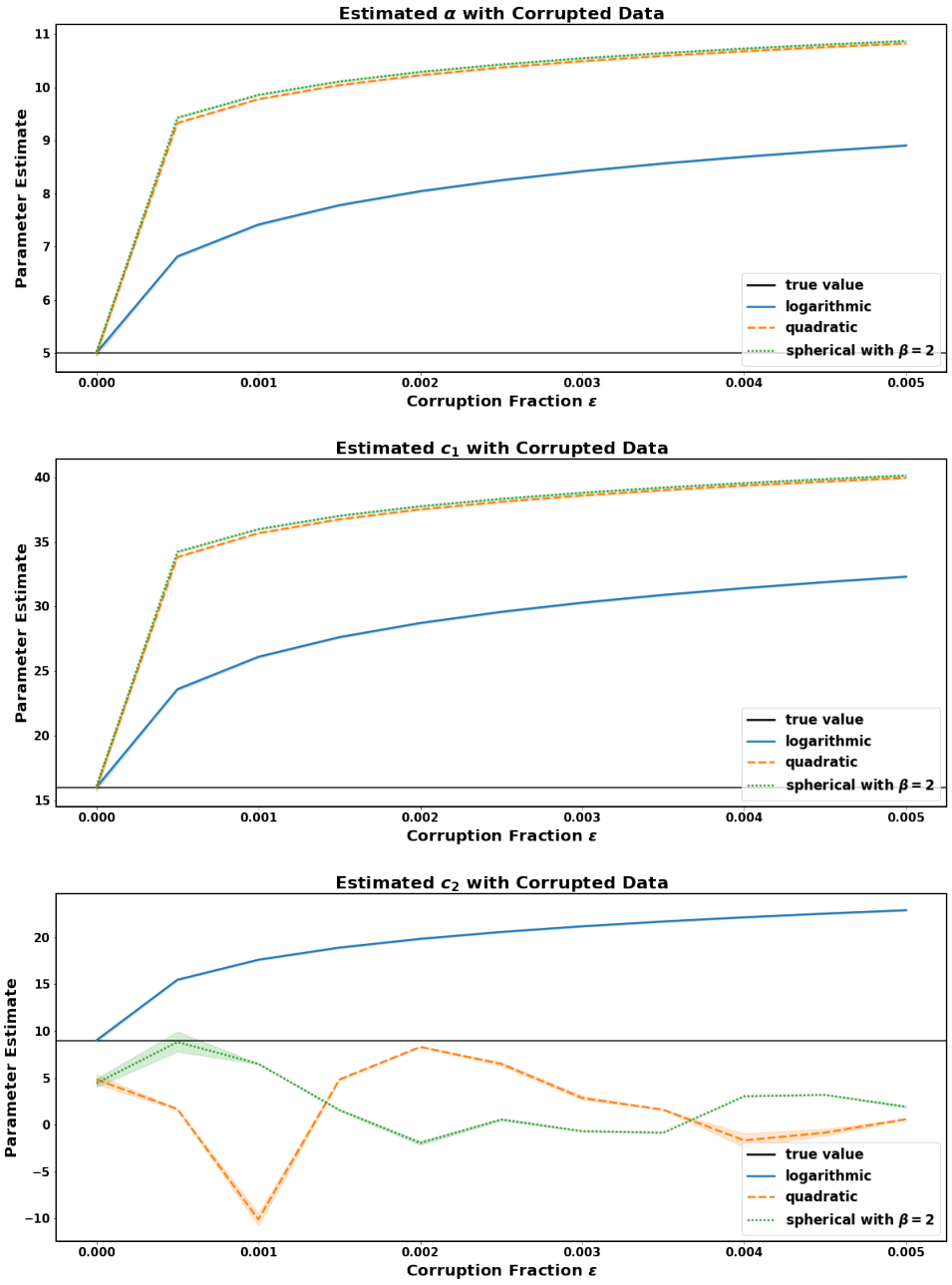

Figure 7-15: The parameter estimates of price sensitivity α and two intercepts c_1 and c_2 from the three loss functions in the multinomial logit model with extreme parameters that make the model violate the assumption for the quadratic loss function. For each iteration, 100000 data points are generated from the multinomial logit model with twp products with true parameter $\alpha = 5.0, c_1 = 16.0, c_2 = 9.0$. Then, ϵ fraction of the data points are randomly chosen and have their prices set to $p_1 = 27, p_2 = 37$ and labels set to 1 (the popular product). The corruption fraction ϵ is varied from $0.00, 0.0005, \ldots, 0.005$. The process is repeated for 200 iterations and the mean parameter estimates are plotted with 95% bootstrapping confidence interval.

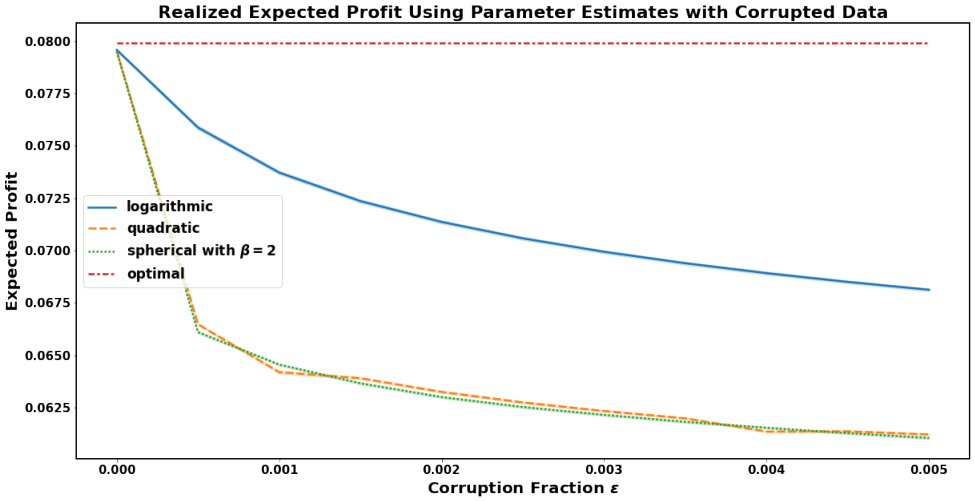

Figure 7-16: The realized expected profit of the multunomial logit model using parameter estimates from the three loss functions in the multinomial logit model with extreme parameters that make the model violate the assumption for the quadratic loss function. For each iteration, 100000 data points are generated from the multinomial logit model with two products with true parameter $\alpha = 5.0, c_1 = 16.0, c_2 = 9.0$. Then, ϵ fraction of the data points are randomly chosen and have their prices set to $p_1 = 27, p_2 = 37$ and labels set to 1 (the popular product). The parameter estimates using these data points are used to estimate the optimal prices for the two products assuming that the products cost $t_1 = 55$ and $t_2 = 30$. The corruption fraction ϵ is varied from $0.00, 0.0005, \ldots, 0.005$. The process is repeated for 200 iterations and the mean realized profits are plotted with 95% bootstrapping confidence interval.

Chapter 8

Conclusion and Further Discussion

In this thesis, we have discussed in detail the robustness property of different strictly consistent loss functions when we perform parameter estimation in multinomial outcome models. We mainly consider the logarithmic loss function, which is the equivalence of the Maximum Likelihood Estimation (MLE) method, the quadratic loss function, and sometimes the spherical loss function with parameter β. We have shown both from the theoretical perspective and the empirical perspective that the logarithmic loss function can be sensitive to even a small fraction of corruption. On the other hand, the quadratic loss function can be shown to be robust against any corruption when some mild assumptions about the model are met.

From the theoretical perspective, we have considered two approaches to study the robustness properties of loss functions. For the first approach, we study the magnitude of the influence functions corresponding to each loss function. We define a loss function to be partially robust if the 2-norm of the corresponding influence function is bounded. We have also shown that the logarithmic loss function is not partially robust, while the quadratic loss function and the spherical loss function with $\beta \geq 2$ are. In the case of the quadratic loss function, we have also provided an upper bound of the absolute value of the influence function when the parameter is one-dimenaional and an upper bound of the 2-norm of the influence function when the parameter is multi-dimensional.

In the second approach, we directly consider the magnitude of the bias of the parameter estimates. We define a loss function to be strongly robust if the 2-norm of the bias is proportional to the product of the corruption fraction ϵ and the 2-norm of the true parameter. We have shown that the logarithmic loss function is not strongly robust, whereas the quadratic loss function is strongly robust with respect to any one-dimensional logistic regression model up to a certain level of corruption fraction. It is worth to note that the assumptions used to prove the results for the quadratic loss function are identical to the ones used to prove the partial robustness results.

From the empirical perspective, we have introduced multiple interesting synthetic multinomial outcome models and demonstrated that the empirical behaviors match with the intuition from the theoretical results. We have shown that in the presence of corruption, the parameter estimates from the logarithmic loss function can be greatly biased. We have also shown that the biased parameter estimates greatly affect the subsequent optimization problems that need the parameter estimate as an input. Specifically, we study profit maximization problem that solves for the optimal prices using the parameter estimates. Under corruption, the logarithmic loss function achieves lower expected profit than those from the other two robust loss functions. The results are in agreement with the robustness properties shown from the theoretical perspective and reiterate the importance of using robust loss functions when corruption is present.

Multiple research directions naturally follow our work in this thesis. First, one can further study the spherical loss function to see if it also has properties that are similar to the partial robustness and the strong robustness results for the quadratic loss function. Second, it is desirable to generalize the strong robustness result for the quadratic loss function to a wider class of models. Third, one can study a family of robust consistent loss functions as a whole instead of inspecting specific loss functions separately. Lastly, one can study the behavior of each loss function in real-world data and confirm that the results agree with our results in theory and in the synthetic data. We visit these research directions in detail below.

Generalize results to other loss functions. We have proven the bound on the influence function only for the quadratic loss function. However, we know that the spherical loss function with $\beta \geq 2$ is also partially robust. It would thus be interesting to see if the spherical loss functions also have similar concrete bounds for their influence functions. The set of assumptions needed is likely to be different, so it would also be interesting to compare the assumptions used in the spherical loss function with those used in the quadratic loss function. Similar to when the assumptions for the quadratic loss function give us intuition about the kind of models in which the quadratic loss function is not robust, the assumptions for the spherical loss function may also provide similar useful intuition.

Extend strong robustness results to other models. We have only shown that the quadratic loss function is strongly robust with respect to a one-dimensional logistic regression model. It would be valuable to extend this robustness result to a multi-dimensional logistic regression model or other multinomial outcome models. However, to the best of our knowledge, our approach that uses the intermediate value theorem on the first-order condition for the one-dimensional case is not trivially generalizable to the multi-dimensional case. Therefore, to prove in the multi-dimensional case, one might need to utilize different techniques.

A family of robust consistent loss functions. So far, we have studied the robustness property of three specific loss functions in the family of strictly consistent loss functions $\mathcal{L}$. It would be interesting to research if we can generalize the robustness results to a broader subset of loss functions in $\mathcal{L}$. For example, one can attempt to derive a closed-form formula for the subfamily of $\mathcal{L}$ that contains only partially robust loss functions.

Study real-world data. The natural next step is to study the robustness property of loss functions in real-world empirical data. We suggest three possible ways one can approach this step. First, if there is a prior belief that the real-world dataset might contain some corruption, one can simply apply different strictly consistent loss functions to estimate the latent parameters of the multinomial outcome models and study the difference in the resulting parameter estimates. If the parameter estimates from

the loss functions that are not robust are significantly different from those from the robust loss functions, a further investigation is then needed to determine corruption in the dataset. One can also try to mitigate the effect from corruption by utilizing filtering-based algorithms with theoretical guarantee such as SEVER algorithm [Diakonikolas et al., 2019], and compare the results with our approach. Second, similar to [Broderick et al., 2020], one can check if dropping a data point significantly changes the parameter estimates from loss functions that are not robust, e.g., the logarithmic loss function. Lastly, one can artificially change the values of either the covariate or the label or both of a small fraction of data points in the real-world data and study the robustness behavior of each loss function.

Appendix A

Proofs

A.1 Proof of Theorem 1

Consider any strictly consistent loss function $\ell \in \mathcal{L}$. We have that for any $\boldsymbol{\theta} \in \Theta$,

$$\begin{aligned}
\mathbb{E}_F \left[\ell(Y, \boldsymbol{f}(X;\boldsymbol{\theta}))\right] &= \mathbb{E}_{F_X}\left[\mathbb{E}_{Y\sim\boldsymbol{f}(X;\boldsymbol{\theta}^*)}\left[\ell(Y, \boldsymbol{f}(X;\boldsymbol{\theta}))|\, X\right]\right] && \text{(A.1)}\\
&\geq \mathbb{E}_{F_X}\left[\mathbb{E}_{Y\sim\boldsymbol{f}(X;\boldsymbol{\theta}^*)}\left[\ell(Y, \boldsymbol{f}(X;\boldsymbol{\theta}^*))|\, X\right]\right] && \text{(A.2)}\\
&= \mathbb{E}_F \left[\ell(Y, \boldsymbol{f}(X;\boldsymbol{\theta}^*))\right], && \text{(A.3)}
\end{aligned}$$

with the equality happens only when $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. The equalities in (A.1) and (A.3) utilize the law of iterated expectation. The inequality (A.2) results from the property of the strictly consistent loss function ℓ. Therefore,

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_F \left[\ell(Y, \boldsymbol{f}(X;\boldsymbol{\theta}))\right],$$

as desired.

A.2 Proof of Lemma 1

From Definition 6, we have that

$$\hat{\boldsymbol{\theta}}_{\epsilon,G} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}\, \mathbb{E}_{F_{\epsilon,G}}\left[L(Z,\boldsymbol{\theta})\right].$$

Therefore, from the first order condition,

$$\begin{aligned} 0 &= \nabla_{\boldsymbol{\theta}}\, \mathbb{E}_{F_{\epsilon,G}}\left[L(Z,\hat{\boldsymbol{\theta}}_{\epsilon,G})\right] = \mathbb{E}_{F_{\epsilon,G}}\left[\nabla_{\boldsymbol{\theta}} L(Z,\hat{\boldsymbol{\theta}}_{\epsilon,G})\right] \\ &= (1-\epsilon)\mathbb{E}_F\left[\nabla_{\boldsymbol{\theta}} L(Z,\hat{\boldsymbol{\theta}}_{\epsilon,G})\right] + \epsilon\mathbb{E}_G\left[\nabla_{\boldsymbol{\theta}} L(Z,\hat{\boldsymbol{\theta}}_{\epsilon,G})\right]. \end{aligned} \tag{A.4}$$

The last equality comes from the property in Eq. (4.2). We then implicitly differentiate this identity to infer $\frac{d\,\hat{\boldsymbol{\theta}}_{\epsilon,G}}{d\,\epsilon}$.

$$\begin{aligned} 0 &= \frac{d\,(1-\epsilon)\mathbb{E}_F\left[\nabla_{\boldsymbol{\theta}} L(Z,\hat{\boldsymbol{\theta}}_{\epsilon,G})\right] + \epsilon\mathbb{E}_G\left[\nabla_{\boldsymbol{\theta}} L(Z,\hat{\boldsymbol{\theta}}_{\epsilon,G})\right]}{d\,\epsilon} \\ &= -\mathbb{E}_F\left[\nabla_{\boldsymbol{\theta}} L(Z,\hat{\boldsymbol{\theta}}_{\epsilon,G})\right] + (1-\epsilon)\mathbb{E}_F\left[\nabla^2_{\boldsymbol{\theta}} L(Z,\hat{\boldsymbol{\theta}}_{\epsilon,G})\right]\frac{d\,\hat{\boldsymbol{\theta}}_{\epsilon,G}}{d\,\epsilon} \\ &\quad + \mathbb{E}_G\left[\nabla_{\boldsymbol{\theta}} L(Z,\hat{\boldsymbol{\theta}}_{\epsilon,G})\right] + \epsilon\mathbb{E}_G\left[\nabla^2_{\boldsymbol{\theta}} L(Z,\hat{\boldsymbol{\theta}}_{\epsilon,G})\right]\frac{d\,\hat{\boldsymbol{\theta}}_{\epsilon,G}}{d\,\epsilon} \\ &= \left((1-\epsilon)\mathbb{E}_F\left[\nabla^2_{\boldsymbol{\theta}} L(Z,\hat{\boldsymbol{\theta}}_{\epsilon,G})\right] + \epsilon\mathbb{E}_G\left[\nabla^2_{\boldsymbol{\theta}} L(Z,\hat{\boldsymbol{\theta}}_{\epsilon,G})\right]\right)\frac{d\,\hat{\boldsymbol{\theta}}_{\epsilon,G}}{d\,\epsilon} \\ &\quad - \left(\mathbb{E}_F\left[\nabla_{\boldsymbol{\theta}} L(Z,\hat{\boldsymbol{\theta}}_{\epsilon,G})\right] - \mathbb{E}_G\left[\nabla_{\boldsymbol{\theta}} L(Z,\hat{\boldsymbol{\theta}}_{\epsilon,G})\right]\right) \\ &= \left((1-\epsilon)\mathbb{E}_F\left[\nabla^2_{\boldsymbol{\theta}} L(Z,\hat{\boldsymbol{\theta}}_{\epsilon,G})\right] + \epsilon\mathbb{E}_G\left[\nabla^2_{\boldsymbol{\theta}} L(Z,\hat{\boldsymbol{\theta}}_{\epsilon,G})\right]\right)\frac{d\,\hat{\boldsymbol{\theta}}_{\epsilon,G}}{d\,\epsilon} + \frac{\mathbb{E}_G\left[\nabla_{\boldsymbol{\theta}} L(Z,\hat{\boldsymbol{\theta}}_{\epsilon,G})\right]}{1-\epsilon}. \end{aligned}$$

Note that the last equality utilizes the identity in Eq. (A.4). Therefore, we achieve the desirable formula for $\frac{d\,\hat{\boldsymbol{\theta}}_{\epsilon,G}}{d\,\epsilon}$:

$$-\left((1-\epsilon)^2\mathbb{E}_F\left[\nabla^2_{\boldsymbol{\theta}} L(Z,\hat{\boldsymbol{\theta}}_{\epsilon,G})\right] + \epsilon(1-\epsilon)\mathbb{E}_G\left[\nabla^2_{\boldsymbol{\theta}} L(Z,\hat{\boldsymbol{\theta}}_{\epsilon,G})\right]\right)^{-1}\mathbb{E}_G\left[\nabla_{\boldsymbol{\theta}} L(Z,\hat{\boldsymbol{\theta}}_{\epsilon,G})\right].$$

A.3 Proof of Theorem 3

From the definition,

$$\hat{\boldsymbol{\theta}}_{\epsilon,z} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} L(z_i,\boldsymbol{\theta}) + \epsilon L(z,\boldsymbol{\theta}).$$

So, from the first order condition, we have

$$0 = \frac{1}{n}\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} L(z_i,\hat{\boldsymbol{\theta}}_{\epsilon,z}) + \epsilon \nabla_{\boldsymbol{\theta}} L(z,\hat{\boldsymbol{\theta}}_{\epsilon,z}).$$

Again, we implicitly differentiate the right-hand side of the equation.

$$\begin{aligned} 0 &= \frac{d\,\frac{1}{n}\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} L(z_i,\hat{\boldsymbol{\theta}}_{\epsilon,z}) + \epsilon \nabla_{\boldsymbol{\theta}} L(z,\hat{\boldsymbol{\theta}}_{\epsilon,z})}{d\,\epsilon} \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} L(z_i,\hat{\boldsymbol{\theta}}_{\epsilon,z}) + \epsilon \nabla^2_{\boldsymbol{\theta}} L(z,\hat{\boldsymbol{\theta}}_{\epsilon,z})\right) \frac{d\hat{\boldsymbol{\theta}}_{\epsilon,z}}{d\epsilon} + \nabla_{\boldsymbol{\theta}} L(z,\hat{\boldsymbol{\theta}}_{\epsilon,z}). \end{aligned}$$

Therefore,

$$\begin{aligned} \mathcal{I}(z) = \left.\frac{d\hat{\boldsymbol{\theta}}_{\epsilon,z}}{d\epsilon}\right|_{\epsilon=0} &= -\left(\frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} L(z_i,\hat{\boldsymbol{\theta}}_{\epsilon,z}) + \epsilon \nabla^2_{\boldsymbol{\theta}} L(z,\hat{\boldsymbol{\theta}}_{\epsilon,z})\right)^{-1} \nabla_{\boldsymbol{\theta}} L(z,\hat{\boldsymbol{\theta}}_{\epsilon,z})\Bigg|_{\epsilon=0} \\ &= -\left(\frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} L(z_i,\hat{\boldsymbol{\theta}}_{0,z})\right)^{-1} \nabla_{\boldsymbol{\theta}} L(z,\hat{\boldsymbol{\theta}}_{0,z}) \\ &= -H^{-1}_{\hat{\boldsymbol{\theta}}} \nabla_{\boldsymbol{\theta}} L\left(z,\hat{\boldsymbol{\theta}}\right) \end{aligned}$$

as desired. This comes from the fact that $\hat{\boldsymbol{\theta}}_{0,z} = \hat{\boldsymbol{\theta}}$ for any z.

A.4 Proof of Lemma 2

For $1 \leq i \leq m$ and $1 \leq j \leq n$, let a_{ij} be the element on the i-th row and j-th column of matrix A. Also for $1 \leq j \leq n$, let v_j be the j-th element of vector v. We have that

$$\begin{aligned}
\|Av\|_2^2 &= \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij} v_j \right)^2 \\
&\leq \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij}^2 \right) \left(\sum_{j=1}^{n} v_j^2 \right) \\
&= \left(\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 \right) \left(\sum_{j=1}^{n} v_j^2 \right) \\
&= \|A\|_{Frob}^2 \|v\|_2^2 ,
\end{aligned}$$

when the inequality results from Cauchy-Schwarz inequality and the rest of the equalities come from definitions and identities. Therefore, $\|Av\|_2 \leq \|A\|_{Frob} \|v\|_2$ as desired.

A.5 Proof of Theorem 5

We have that

$$\frac{\partial \ell_{sphere}^{(\beta)}(y, \boldsymbol{p})}{\partial p_i} = \begin{cases} \frac{(\beta-1) p_y^{\beta-2}}{\left(\sum_{j=0}^{C-1} p_j^{\beta}\right)^{\frac{\beta-1}{\beta}}} \left[\frac{p_y^{\beta}}{\sum_{j=0}^{C-1} p_j^{\beta}} - 1 \right] & \text{if } i = y, \\ \frac{(\beta-1) p_i^{\beta-1} p_y^{\beta-1}}{\left(\sum_{j=0}^{C-1} p_j^{\beta}\right)^{\frac{2\beta-1}{\beta}}} & \text{otherwise.} \end{cases}$$

First, we consider any $\beta \geq 2$. From Hölder's inequality

$$\left(\sum_{j=0}^{C-1} p_j^{\beta} \right)^{\frac{1}{\beta}} C^{\frac{\beta-1}{\beta}} \geq \sum_{j=0}^{C-1} p_j = 1.$$

Therefore

$$\frac{1}{\left(\sum_{j=0}^{C-1} p_j^{\beta}\right)^{\frac{1}{\beta}}} \leq C^{\frac{\beta-1}{\beta}}.$$

When $i = y$, we have that

$$\begin{aligned}\left|\frac{\partial \ell_{sphere}^{(\beta)}(y, \boldsymbol{p})}{\partial p_i}\right| &= \frac{(\beta-1)p_y^{\beta-2}}{\left(\sum_{j=0}^{C-1} p_j^{\beta}\right)^{\frac{\beta-1}{\beta}}} \left|\frac{p_y^{\beta}}{\sum_{j=0}^{C-1} p_j^{\beta}} - 1\right| \\ &\leq \frac{(\beta-1)p_y^{\beta-2}}{\left(\sum_{j=0}^{C-1} p_j^{\beta}\right)^{\frac{\beta-1}{\beta}}} \\ &\leq (\beta-1)C^{\frac{(\beta-1)^2}{\beta}}.\end{aligned}$$

When $i \neq y$, we instead have

$$\begin{aligned}\left|\frac{\partial \ell_{sphere}^{(\beta)}(y, \boldsymbol{p})}{\partial p_i}\right| &= \frac{(\beta-1)p_i^{\beta-1}p_y^{\beta-1}}{\left(\sum_{j=0}^{C-1} p_j^{\beta}\right)^{\frac{2\beta-1}{\beta}}} \\ &\leq (\beta-1)C^{\frac{(2\beta-1)(\beta-1)}{\beta}}.\end{aligned}$$

Therefore,

$$\left\|\nabla_{\boldsymbol{p}} \ell_{sphere}^{(\beta)}(y, \boldsymbol{p})\right\| \leq (\beta-1)\sqrt{C^{\frac{2(\beta-1)^2}{\beta}} + (C-1)C^{\frac{2(2\beta-1)(\beta-1)}{\beta}}},$$

and thus is bounded. From Theorem 4, we have that when $\beta \geq 2$, the spherical loss functions with parameter $\beta \geq 2$ are partially robust. Now, consider any β such that $1 < \beta \leq 2$. When $\boldsymbol{p} \to (0, 0, \ldots, 1)$, we have that at $y = i = 0$

$$\frac{\partial \ell_{sphere}^{(\beta)}(0, \boldsymbol{p})}{\partial p_0} \xrightarrow[\boldsymbol{p}\to(0,0,\ldots,1)]{} \frac{(\beta-1)0^{\beta-2}}{1}\left[\frac{0}{1} - 1\right] = -\infty.$$

Therefore, $\left\|\nabla_{\boldsymbol{p}} \ell_{sphere}^{(\beta)}(y, \boldsymbol{p})\right\|$ is not bounded, violating the condition in Eq. (5.4) in Theorem 4.

A.6 Proof of Lemma 3 and Lemma 4

Because Lemma 3 is just the one-dimensional case of Lemma 4, we only need to show the proof of Lemma 4.

Proof. For any $\boldsymbol{\theta} \in \Theta$, we have that

$$\begin{aligned}
&\mathbb{E}_F \left[\sum_{i=0}^{C-1} \nabla^2_{\boldsymbol{\theta}} f_i(X;\boldsymbol{\theta}^*) \frac{\partial \ell(Y, \boldsymbol{f}(X;\boldsymbol{\theta}))}{\partial f_i} \right] \\
&= \mathbb{E}_{F_X} \left[\mathbb{E}_{Y \sim \boldsymbol{f}(X;\boldsymbol{\theta}^*)} \left[\sum_{i=0}^{C-1} \nabla^2_{\boldsymbol{\theta}} f_i(X;\boldsymbol{\theta}^*) \frac{\partial \ell(Y, \boldsymbol{f}(X;\boldsymbol{\theta}))}{\partial f_i} \middle| X \right] \right] \\
&= \mathbb{E}_{F_X} \left[\sum_{i=0}^{C-1} \left(\nabla^2_{\boldsymbol{\theta}} f_i(X;\boldsymbol{\theta}^*) \mathbb{E}_{Y \sim \boldsymbol{f}(X;\boldsymbol{\theta}^*)} \left[\frac{\partial \ell(Y, \boldsymbol{f}(X;\boldsymbol{\theta}))}{\partial f_i} \middle| X \right] \right) \right] \\
&= \mathbb{E}_{F_X} \left[\sum_{i=0}^{C-1} \left(\nabla^2_{\boldsymbol{\theta}} f_i(X;\boldsymbol{\theta}^*) \sum_{y=0}^{C-1} f_y(X;\boldsymbol{\theta}^*) \frac{\partial \ell(y, \boldsymbol{f}(X;\boldsymbol{\theta}))}{\partial f_i} \right) \right]. \qquad \text{(A.5)}
\end{aligned}$$

The first equality comes from the law of iterated expectation. The rest of the equalities are simply algebraic manipulation.

Because ℓ is consistent, we know that $\boldsymbol{f}(X;\boldsymbol{\theta}^*)$ is the answer of the following constrained maximization,

$$\begin{aligned}
\text{minimize:} \quad & \sum_{y=0}^{C-1} \ell(y, \boldsymbol{p})\, f_y(X;\boldsymbol{\theta}^*) \\
\text{subject to:} \quad & \sum_{i=0}^{C-1} p_i = 1.
\end{aligned}$$

From Lagrange multipliers, there exists λ such that

$$\frac{\partial \sum_{y=0}^{C-1} \ell(y, \boldsymbol{p})\, f_y(X;\boldsymbol{\theta}^*)}{\partial p_i} - \lambda \frac{\partial \sum_{i=0}^{C-1} p_i - 1}{\partial p_i} \bigg|_{\boldsymbol{p} = \boldsymbol{f}(X;\boldsymbol{\theta}^*)} = 0.$$

That is,

$$\forall 0 \le i < C, \quad \sum_{y=0}^{C-1} f_y(X;\boldsymbol{\theta}^*) \frac{\partial \ell(y, \boldsymbol{f}(X;\boldsymbol{\theta}^*))}{\partial f_i} = \lambda. \qquad \text{(A.6)}$$

Therefore, using Eq. (A.5) and Eq. (A.6), we have the desired result:

$$\mathbb{E}_F\left[\sum_{i=0}^{C-1}\nabla^2_{\boldsymbol{\theta}} f_i(X;\boldsymbol{\theta}^*)\frac{\partial \ell(Y,\boldsymbol{f}(X;\boldsymbol{\theta}^*))}{\partial f_i}\right] = \mathbb{E}_{F_X}\left[\sum_{i=0}^{C-1}\nabla^2_{\boldsymbol{\theta}} f_i(X;\boldsymbol{\theta}^*)\lambda\right]$$
$$= \mathbb{E}_{F_X}\left[\lambda\nabla^2_{\boldsymbol{\theta}}\sum_{i=0}^{C-1} f_i(X;\boldsymbol{\theta}^*)\right] = \mathbb{E}_{F_X}\left[\lambda\nabla^2_{\boldsymbol{\theta}} 1\right] = \mathbf{0}.$$

□

A.7 Proof of Lemma 5

From the Definition 3 of the 2-norm of matrix, we have

$$\|Av\|_2 \leq \|A\|_2 \|v\|_2 = \lambda_{max}(A)\|v\|_2 .$$

Therefore, using the fact that if λ is an eigenvalue of A, then $\frac{1}{\lambda}$ is also an eigenvalue of A^{-1},

$$\left\|A^{-1}v\right\|_2 \leq \lambda_{max}(A^{-1})\|v\|_2 = \frac{\|v\|_2}{\lambda_{min}(A)},$$

as desired.

A.8 Proof of Lemma 6

Let

$$u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \quad \text{and} \quad v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}.$$

Therefore

$$\|D(u)v\|_2 = \sqrt{\sum_{i=1}^{n} u_i v_i} \leq \sqrt{\sum_{i=1}^{n} u_i^2}\sqrt{\sum_{i=1}^{n} v_i^2} = \|u\|_2 \|v\|_2$$

as desired. Note that the inequality is the application of Cauchy-Schwarz inequality.

A.9 Proof of Theorem 8

From the definition of $\hat{\theta}^{log}_{\epsilon,G}$, we know that

$$\hat{\theta}^{log}_{\epsilon,G} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \mathbb{E}_{F_{\epsilon,G}} \left[\ell_{log}(Y, \boldsymbol{f}(X;\boldsymbol{\theta}))\right].$$

Consider the first-order condition

$$\begin{aligned}
&A_{\epsilon,G}(\theta) \\
&= \frac{d}{d\theta}\mathbb{E}_{F_{\epsilon,G}} \left[\ell_{log}(Y, \boldsymbol{f}(X;\boldsymbol{\theta}))\right] \\
&= \frac{d}{d\theta}\mathbb{E}_{F_{\epsilon,G}} \left[\log\left(e^{\theta X}+1\right) - Y\theta X\right] && \text{(A.7)} \\
&= \mathbb{E}_{F_{\epsilon,G}} \left[X\left(\frac{e^{\theta X}}{1+e^{\theta X}} - Y\right)\right] \\
&= (1-\epsilon)\mathbb{E}_{F} \left[X\left(\frac{e^{\theta X}}{1+e^{\theta X}} - Y\right)\right] + \epsilon\mathbb{E}_{G} \left[X\left(\frac{e^{\theta X}}{1+e^{\theta X}} - Y\right)\right] && \text{(A.8)} \\
&= (1-\epsilon)\mathbb{E}_{F_X} \left[X\left(\frac{e^{\theta X}}{1+e^{\theta X}} - \frac{e^{\theta^* X}}{1+e^{\theta^* X}}\right)\right] + \epsilon\mathbb{E}_{G} \left[X\left(\frac{e^{\theta X}}{1+e^{\theta X}} - Y\right)\right]. && \text{(A.9)}
\end{aligned}$$

The eqaulity (A.7) comes from the negative log loss formula for logistic regression. The eqaulity (A.8) comes from the expansion of expectation for (ϵ, G)-corrupted distribution in Eq. (4.2). Lastly, the equality (A.9) results from the law of iterated expectation.

Let $V = |\theta^*|$ and

$$U = \sup_{\theta\in[\theta^*-V,\theta^*+V]} \left|\mathbb{E}_{F_X} \left[X\left(\frac{e^{\theta X}}{1+e^{\theta X}} - \frac{e^{\theta^* X}}{1+e^{\theta^* X}}\right)\right]\right|. \tag{A.10}$$

Note that at $Y = 0$,

$$\forall\, \theta \in \Theta, \quad X\left(\frac{e^{\theta X}}{1+e^{\theta X}} - Y\right) = \frac{Xe^{\theta X}}{1+e^{\theta X}} \geq \frac{X}{2}. \tag{A.11}$$

Therefore, if we choose G to be a Dirac distribution that concentrates at the point $X = x' > \frac{2(1-\epsilon)}{\epsilon}U$ and $Y = 0$ ($\mathbb{P}_G(X = x', Y = 0) = 1$), we have that for any $\theta \in [\theta^* - V, \theta^* + V]$

$$\begin{aligned}
|A_{\epsilon,G}(\theta)| &\geq \epsilon \left|\mathbb{E}_G\left[X\left(\frac{e^{\theta X}}{1+e^{\theta X}} - Y\right)\right]\right| \\
&\quad - (1-\epsilon)\left|\mathbb{E}_{F_X}\left[X\left(\frac{e^{\theta X}}{1+e^{\theta X}} - \frac{e^{\theta^* X}}{1+e^{\theta^* X}}\right)\right]\right| && \text{(A.12)} \\
&> (1-\epsilon)U - (1-\epsilon)\left|\mathbb{E}_{F_X}\left[X\left(\frac{e^{\theta X}}{1+e^{\theta X}} - \frac{e^{\theta^* X}}{1+e^{\theta^* X}}\right)\right]\right| && \text{(A.13)} \\
&= (1-\epsilon)\left(U - \left|\mathbb{E}_{F_X}\left[X\left(\frac{e^{\theta X}}{1+e^{\theta X}} - \frac{e^{\theta^* X}}{1+e^{\theta^* X}}\right)\right]\right|\right) \\
&\geq 0. && \text{(A.14)}
\end{aligned}$$

The inequality (A.12) results from the triangle inequality. The inequality (A.13) comes from the choice of X and Y and the property in Eq. (A.11). Lastly, the inequality (A.14) directly results from the definition of U in Eq. (A.10).

As a result, $\hat{\theta}^{log}_{\epsilon,G} \notin [\theta^* - V, \theta^* + V]$, and thus $\left|\hat{\theta}^{log}_{\epsilon,G} - \theta^*\right| \geq V = |\theta^*|$ as desired.

A.10 Proof of Theorem 9

Again, from the definition of $\hat{\theta}^{log}_{\epsilon,G}$, we know that

$$\hat{\theta}^{log}_{\epsilon,G} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{F_{\epsilon,G}} \left[\ell_{log}(Y, \boldsymbol{f}(X;\boldsymbol{\theta}))\right].$$

Consider the first-order condition

$$\begin{aligned}
A_{\epsilon,G}(\theta) &= \frac{d}{d\theta}\mathbb{E}_{F_{\epsilon,G}} \left[\ell_{log}(Y, \boldsymbol{f}(X;\boldsymbol{\theta}))\right] \\
&= \frac{d}{d\theta}\mathbb{E}_{F_{\epsilon,G}} \left[-\log f_Y(X;\theta)\right] \\
&= -\mathbb{E}_{F_{\epsilon,G}} \left[\frac{d\, f_Y(X;\theta)}{d\theta}\frac{1}{f_Y(X;\theta)}\right] \\
&= -(1-\epsilon)\mathbb{E}_F \left[\frac{d\, f_Y(X;\theta)}{d\theta}\frac{1}{f_Y(X;\theta)}\right] - \epsilon\mathbb{E}_G \left[\frac{d\, f_Y(X;\theta)}{d\theta}\frac{1}{f_Y(X;\theta)}\right],
\end{aligned}$$

when the last equality utilizes the expansion of expectation for (ϵ, G)-corrupted distribution in Eq. (4.2). Let $V = |\theta^*|$ and

$$U = \sup_{\theta\in[\theta^*-V,\theta^*+V]} \left|\mathbb{E}_F \left[\frac{d\, f_Y(X;\theta)}{d\theta}\frac{1}{f_Y(X;\theta)}\right]\right|.$$

From the condition of model F, there exists $y' \in \mathcal{Y}$ and $x' \in \mathcal{X}$ such that

$$\forall\, \theta \in \Theta, \quad \left|\frac{d\, f_{y'}(x';\theta)}{d\theta}\frac{1}{f_{y'}(x';\theta)}\right| > \frac{1-\epsilon}{\epsilon}U.$$

Therefore, if we choose G to be a Dirac distribution that concentrates at the point $X = x'$ and $Y = y'$ ($\mathbb{P}_G(X = x', Y = y') = 1$), we have that for any $\theta \in [\theta^* - V, \theta^* + V]$

$$\begin{aligned}
|A_{\epsilon,G}(\theta)| &\geq \epsilon \left|\frac{d\, f_{y'}(x';\theta)}{d\theta}\frac{1}{f_{y'}(x';\theta)}\right| - (1-\epsilon)\left|\mathbb{E}_F \left[\frac{d\, f_Y(X;\theta)}{d\theta}\frac{1}{f_Y(X;\theta)}\right]\right| \\
&> (1-\epsilon)U - (1-\epsilon)\left|\mathbb{E}_F \left[\frac{d\, f_Y(X;\theta)}{d\theta}\frac{1}{f_Y(X;\theta)}\right]\right| \\
&= (1-\epsilon)\left(U - \left|\mathbb{E}_F \left[\frac{d\, f_Y(X;\theta)}{d\theta}\frac{1}{f_Y(X;\theta)}\right]\right|\right) \\
&\geq 0.
\end{aligned}$$

Therefore, $\hat{\theta}^{log}_{\epsilon,G} \notin [\theta^* - V, \theta^* + V]$, and thus $\left|\hat{\theta}^{log}_{\epsilon,G} - \theta^*\right| > V = |\theta^*|$ as desired.

A.11 Proof of Theorem 10

Consider the first-order condition

$$
\begin{aligned}
&A_{\epsilon,G}(\theta) \\
&= \frac{d}{d\theta}\mathbb{E}_{F_{\epsilon,G}}\left[\ell_{quad}(Y, \boldsymbol{f}(X;\boldsymbol{\theta}))\right] \\
&= \frac{d}{d\theta}\mathbb{E}_{F_{\epsilon,G}}\left[2\left(Y - \frac{e^{\theta X}}{1+e^{\theta X}}\right)^2\right] \\
&= 4\mathbb{E}_{F_{\epsilon,G}}\left[\frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\left(\frac{e^{\theta X}}{1+e^{\theta X}} - Y\right)\right] \\
&= 4(1-\epsilon)\mathbb{E}_{F}\left[\frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\left(\frac{e^{\theta X}}{1+e^{\theta X}} - Y\right)\right] + 4\epsilon\mathbb{E}_{G}\left[\frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\left(\frac{e^{\theta X}}{1+e^{\theta X}} - Y\right)\right],
\end{aligned}
\tag{A.15}
$$

when the last equality comes from the expansion of expectation for (ϵ, G)-corrupted distribution in Eq. (4.2).

First, we will show that the absolute value of the second expectation term in Eq. (A.15) (the expectation with respect to G) is bounded by $\frac{1}{\theta}$. Then we will show that by changing the value of θ to be slightly above or slightly below θ^*, we can make the whole term in Eq. (A.15) positive and negative respectively.

Now, we will proof the first part. We have that for any $X > 0$ and $\theta > 0$,

$$
\frac{Xe^{\theta X}}{(1+e^{\theta X})^2} = \left(\frac{X}{1+e^{\theta X}}\right)\left(\frac{e^{\theta X}}{1+e^{\theta X}}\right) < \left(\frac{X}{\theta X}\right)(1) = \frac{1}{\theta},
$$

when the inequality comes from the fact that $e^{\theta X} + 1 \geq \theta X + 2 > \theta X$. Because both Y and $\frac{e^{\theta X}}{1+e^{\theta X}}$ are in the range of $[0, 1]$, we also have that

$$
\left|\frac{e^{\theta X}}{1+e^{\theta X}} - Y\right| \leq 1.
$$

Therefore,

$$\left|\mathbb{E}_G\left[\frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\left(\frac{e^{\theta X}}{1+e^{\theta X}}-Y\right)\right]\right| \leq \mathbb{E}_G\left[\left|\frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\left(\frac{e^{\theta X}}{1+e^{\theta X}}-Y\right)\right|\right]$$
$$\leq \mathbb{E}_G\left[\frac{1}{\theta}\right] = \frac{1}{\theta}. \tag{A.16}$$

Next, for the second part of the proof, consider that

$$\begin{aligned}
\mathbb{E}_F\left[\frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\left(\frac{e^{\theta X}}{1+e^{\theta X}}-Y\right)\right] &= \mathbb{E}_{F_X}\left[\frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\left(\frac{e^{\theta X}}{1+e^{\theta X}}-\mathbb{E}\left[Y \mid X\right]\right)\right] \\
&= \mathbb{E}_{F_X}\left[\frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\left(\frac{e^{\theta X}}{1+e^{\theta X}}-\frac{e^{\theta^* X}}{1+e^{\theta^* X}}\right)\right] \\
&= \mathbb{E}_{F_X}\left[\frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\left(\sigma(\theta X)-\sigma(\theta^* X)\right)\right].
\end{aligned}$$

We will use the following four properties to prove the second part.

1. From the mean value theorem, for any $x_1 < x_2$, we have that

$$\exists\, x_1 < \delta < x_2, \quad \sigma(x_2) = \sigma(x_1) + \sigma'(\delta)(x_2 - x_1).$$

2. For any $x > 0$, we have

$$\sigma''(x) = \frac{e^x(1-e^x)}{(1+e^x)^3} < 0.$$

   Therefore, $\sigma'(x)$ is a decreasing function.

3. Because

$$\frac{d}{d\theta}\frac{Xe^{\theta X}}{(1+e^{\theta X})^2} = \frac{X^2e^{\theta X}\left(1-e^{\theta X}\right)}{(1+e^{\theta X})^3} < 0,$$

   we also have that $\frac{Xe^{\theta X}}{\left(1+e^{\theta X}\right)^2}$ is a decreasing function on θ.

4. From Assumption 3, we have that

$$\forall\, \theta \in \Theta, \quad \mathbb{E}_{F_X}\left[\sum_{i=0}^{1}\left(\frac{d\, f_i(X;\theta)}{d\theta}\,\theta\right)^2\right] \geq M.$$

Since $\left|\frac{d\, f_0(X;\theta)}{d\theta}\right| = \left|\frac{d\, f_1(X;\theta)}{d\theta}\right| = \frac{Xe^{\theta X}}{(1+e^{\theta X})^2}$, we have that

$$\forall\, \theta \in \Theta, \quad \mathbb{E}_{F_X}\left[\left(\frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\right)^2\right] \geq \frac{M}{2\theta^2}.$$

Now, consider when $\theta = \theta^* + u$ $(u > 0)$. We have that

$$\begin{aligned}
&\frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\left(\sigma(\theta X) - \sigma(\theta X^*)\right) \\
&= \frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\left(\sigma'\left(\delta\right)(uX)\right) \text{ when } \theta^* X < \delta < \theta X && \text{(A.17)} \\
&\geq \frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\sigma'\left(\theta X\right)(uX) && \text{(A.18)} \\
&= u\left(\frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\right)^2. && \text{(A.19)}
\end{aligned}$$

The equality (A.17) comes from the property 1. The inequality (A.18) utilizes the fact that $\sigma'(x)$ is a decreasing function from the property 2. Lastly The equality (A.19) simply results from the formula of $\sigma'(x)$. Therefore,

$$\begin{aligned}
&A_{\epsilon,G}(\theta) \\
&= 4(1-\epsilon)\mathbb{E}_F\left[\frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\left(\frac{e^{\theta X}}{1+e^{\theta X}} - Y\right)\right] + 4\epsilon\mathbb{E}_G\left[\frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\left(\frac{e^{\theta X}}{1+e^{\theta X}} - Y\right)\right] \\
&\geq 4(1-\epsilon)\mathbb{E}_{F_X}\left[\frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\left(\sigma(\theta X) - \sigma(\theta^* X)\right)\right] - 4\epsilon\frac{1}{\theta} && \text{(A.20)} \\
&\geq 4(1-\epsilon)u\mathbb{E}_{F_X}\left[\left(\frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\right)^2\right] - 4\epsilon\frac{1}{\theta} && \text{(A.21)} \\
&> 4(1-\epsilon)u\mathbb{E}_{F_X}\left[\left(\frac{Xe^{\theta^* X}}{(1+e^{\theta^* X})^2}\right)^2\right] - 4\epsilon\frac{1}{\theta^*} && \text{(A.22)} \\
&\geq 4(1-\epsilon)u\frac{M}{2{\theta^*}^2} - 4\epsilon\frac{1}{\theta^*}. && \text{(A.23)}
\end{aligned}$$

The inequality (A.20) comes from Eq. (A.16) and the law of iterated expectation. The inequality (A.21) comes from Eq. (A.19). The inequality (A.22) uses the fact that $\frac{Xe^{\theta X}}{\left(1+e^{\theta X}\right)^2}$ is a decreasing function on θ in the property 3. The last inequality (A.23) utilizes the property 4. Because the last quantity is positive when $u \geq \frac{2\epsilon}{M(1-\epsilon)}\theta^*$, we have that

$$\forall\, \epsilon > 0, \forall\, u \geq \frac{2\epsilon}{M(1-\epsilon)}\theta^*, \quad A_{\epsilon,G}\left(\theta^* + u\right) > 0. \tag{A.24}$$

Similarly, consider when $\theta = \theta^* - v$ $(v > 0)$. We have that

$$\begin{aligned}
&\frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\left(\sigma(\theta^* X) - \sigma(\theta X)\right) \\
&= \frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\left(\sigma'\left(\delta\right)\left(vX\right)\right) \text{ when } \theta X < \delta < \theta^* X && \text{(A.25)} \\
&\geq \frac{Xe^{\theta^* X}}{(1+e^{\theta^* X})^2}\sigma'\left(\theta^* X\right)\left(vX\right) && \text{(A.26)} \\
&= v\left(\frac{Xe^{\theta^* X}}{(1+e^{\theta^* X})^2}\right)^2. && \text{(A.27)}
\end{aligned}$$

Again, the equality (A.25) comes from the property 1. The inequality (A.26) utilizes the fact that $\sigma'(x)$ is a decreasing function from the property 2 and that $\frac{Xe^{\theta X}}{\left(1+e^{\theta X}\right)^2}$ is a decreasing function on θ from the property 3. Lastly The equality (A.19) results from the formula of $\sigma'(x)$. Therefore,

$$\begin{aligned}
&A_{\epsilon,G}(\theta) \\
&= 4(1-\epsilon)\mathbb{E}_F\left[\frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\left(\frac{e^{\theta X}}{1+e^{\theta X}} - Y\right)\right] + 4\epsilon\mathbb{E}_G\left[\frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\left(\frac{e^{\theta X}}{1+e^{\theta X}} - Y\right)\right] \\
&\leq -4(1-\epsilon)\mathbb{E}_{F_X}\left[\frac{Xe^{\theta X}}{(1+e^{\theta X})^2}\left(\sigma(\theta^* X) - \sigma(\theta X)\right)\right] + 4\epsilon\frac{1}{\theta} \\
&\leq -4(1-\epsilon)v\mathbb{E}_{F_X}\left[\left(\frac{Xe^{\theta^* X}}{(1+e^{\theta^* X})^2}\right)^2\right] + 4\epsilon\frac{1}{\theta} \\
&\leq -4(1-\epsilon)v\frac{M}{2{\theta^*}^2} + 4\epsilon\frac{1}{\theta^* - v}.
\end{aligned}$$

For $\epsilon < \frac{M}{M+8}$, the last quantity is negative when

$$\frac{1}{2}\left(1+\sqrt{1-\frac{8\epsilon}{M(1-\epsilon)}}\right) > v \geq \frac{4\epsilon}{M(1-\epsilon)}\theta^*.$$

Therefore, we have that

$$\forall\, \epsilon < \frac{M}{M+8}, \quad A_{\epsilon,G}\left(\theta^* - \frac{4\epsilon}{M(1-\epsilon)}\theta^*\right) < 0. \tag{A.28}$$

Using the facts from Eq. (A.24) and Eq. (A.28), and from intermediate value theorem, we have that for any $\epsilon < \frac{M}{M+8}$, there exists $\hat{\theta} \in \left[\theta^* - \frac{4\epsilon}{M(1-\epsilon)}\theta^*, \theta^* + \frac{2\epsilon}{M(1-\epsilon)}\theta^*\right]$ such that $A_{\epsilon,G}\left(\hat{\theta}\right) = 0$. Therefore, we can conclude that

$$\forall\, \epsilon < \frac{M}{M+8}, \forall \,\text{corruption } G, \quad \hat{\theta}^{quad}_{\epsilon,G} \in \left[\theta^* - \frac{4\epsilon}{M(1-\epsilon)}\theta^*, \theta^* + \frac{2\epsilon}{M(1-\epsilon)}\theta^*\right],$$

and thus

$$\forall\, \epsilon < \frac{M}{M+8}, \forall \,\text{corruption } G, \quad \left|\hat{\theta}^{quad}_{\epsilon,G} - \theta^*\right| \leq \frac{4\epsilon}{M(1-\epsilon)}\left|\theta^*\right|,$$

as desired.

Bibliography

[Barreno et al., 2010] Barreno, M., Nelson, B., Joseph, A. D., and Tygar, J. D. (2010). The security of machine learning. *Machine Learning*, 81(2):121–148.

[Barron, 2019] Barron, J. T. (2019). A general and adaptive robust loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4331–4339.

[Breunig et al., 2000] Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104.

[Broderick et al., 2020] Broderick, T., Giordano, R., and Meager, R. (2020). An automatic finite-sample robustness metric: Can dropping a little data change conclusions? *arXiv preprint arXiv:2011.14999*.

[Cook and Weisberg, 1980] Cook, R. D. and Weisberg, S. (1980). Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22(4):495–508.

[Cook and Weisberg, 1982] Cook, R. D. and Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.

[de Leeuw, 2019] de Leeuw, E. (2019). Robustness of evaluation metrics for predicting probability estimates of binary outcomes. *Master Thesis Business Analytics and Quantitative Marketing*.

[Diakonikolas et al., 2019] Diakonikolas, I., Kamath, G., Kane, D., Li, J., Steinhardt, J., and Stewart, A. (2019). Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, pages 1596–1606.

[Fischler and Bolles, 1981] Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395.

[Fisher, 1922] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368.

[Fisher, 1925] Fisher, R. A. (1925). Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 22, pages 700–725. Cambridge University Press.

[Frénay and Verleysen, 2013] Frénay, B. and Verleysen, M. (2013). Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869.

[Ghosh et al., 2017] Ghosh, A., Kumar, H., and Sastry, P. (2017). Robust loss functions under label noise for deep neural networks. *arXiv preprint arXiv:1712.09482*.

[Gneiting and Raftery, 2007] Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

[Huber, 1992] Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer.

[Koh and Liang, 2017] Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1885–1894.

[Norden, 1972] Norden, R. (1972). A survey of maximum likelihood estimation. *International Statistical Review/Revue Internationale de Statistique*, pages 329–354.

[Painsky and Wornell, 2018] Painsky, A. and Wornell, G. W. (2018). Bregman divergence bounds and the universality of the logarithmic loss. *arXiv preprint arXiv:1810.07014*.

[Pregibon et al., 1981] Pregibon, D. et al. (1981). Logistic regression diagnostics. *Annals of statistics*, 9(4):705–724.

[Qiu et al., 2019] Qiu, S., Liu, Q., Zhou, S., and Wu, C. (2019). Review of artificial intelligence adversarial attack and defense technologies. *Applied Sciences*, 9(5):909.

[Savage, 1971] Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801.

[White, 1982] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society*, pages 1–25.