Deep neural networks for choice analysis: A statistical learning theory perspective

Shenhao Wang, Qingyi Wang, Nate Bailey, Jinhua Zhao

Massachusetts Institute of Technology

March 2021

Abstract

Although researchers increasingly use deep neural networks (DNN) to analyze individual choices, overfitting and interpretability issues remain obstacles in theory and practice. This study presents a statistical learning theoretical framework to examine the tradeoff between estimation and approximation errors, and between the quality of prediction and of interpretation. It provides an upper bound on the estimation error of the prediction quality in DNN, measured by zero-one and log losses, shedding light on why DNN models do not overfit. It proposes a metric for interpretation quality by formulating a function approximation loss that measures the difference between true and estimated choice probability functions. It argues that the binary logit (BNL) and multinomial logit (MNL) models are the specific cases in the model family of DNNs, since the latter always has smaller approximation errors. We explore the relative performance of DNN and classical choice models through three simulation scenarios comparing DNN, BNL, and binary mixed logit models (BXL), as well as one experiment comparing DNN to BNL, BXL, MNL, and mixed logit (MXL) in analyzing the choice of trip purposes based on the National Household Travel Survey 2017. The results indicate that DNN can be used for choice analysis beyond the current practice of demand forecasting because it has the inherent utility interpretation and the power of automatically learning utility specification. Our results suggest DNN outperforms BNL, BXL, MNL, and MXL models in both prediction and interpretation when the sample size is large ($\geq$O(10^4)), the input dimension is high, or the true data generating process is complex, while performing worse when the opposite is true. DNN outperforms BNL and BXL in zero-one, log, and approximation losses for most of the experiments, and the larger sample size leads to greater incremental value of using DNN over classical discrete choice models. Overall, this study introduces the statistical learning theory as a new foundation for high-dimensional data, complex statistical models, and non-asymptotic data regimes in choice analysis, and the experiments show the effective prediction and interpretation of DNN for its applications to policy and behavioral analysis.

Key words: deep neural networks, choice modeling, statistical learning theory, interpretability

1. Introduction

Choice modeling is a rich theoretical field widely applied throughout transportation research, and in many other contexts (Train, 1980; Ben-Akiva and Lerman, 1985; Train, 2009). While traditional discrete choice models have been used for decades, researchers have recently become increasingly interested in instead using machine learning classifiers to conduct choice analysis due to the high performance of these models in many fields (Karlaftis and Vlahogianni, 2011; Paredes et al., 2017; Hagenauer and Helbich, 2017).

Traditional discrete choice models rely on researchers' use of domain knowledge to filter through several model specifications and find the ones that best fit observed data. Machine learning classifiers can improve upon this approach owing to their automated exploration and extraordinary approximation power. By using flexible model family assumptions, the approximation power of many machine learning methods is much higher than discrete choice models, which are typically limited to a linear-in-parameter form with handcrafted features (e.g. quadratic or log forms). Among all machine learning classifiers, the deep neural network (DNN) is particularly powerful due to several factors. It has high approximation power (Hornik, Stinchcombe, and White, 1989; Hornik, 1991; Cybenko, 1989), can flexibly accommodate various types of information (Krizhevsky, Sutskever, and G. E. Hinton, 2012; LeCun, Bengio, and G. Hinton, 2015), has high predictive power as revealed in experimental studies (Fernández-Delgado et al., 2014; Karlaftis and Vlahogianni, 2011), and has been applied to numerous domains (LeCun, Bengio, and G. Hinton, 2015; Goodfellow et al., 2016; Glaeser et al., 2018). However, two unresolved issues hinder the applicability of DNN in many transportation choice analysis contexts: model overfitting in relatively small data sets, and lack of interpretability.

The first concern in applying DNN to transportation choice analysis is its potential to overfit complex models to high-dimensional data, which arises from the perspective of the classical statistical and econometric theories. An overfitted model fits the training data precisely but has poor out-of-sample performance. Classical statistical theory suggests that the Vapnik-Chervonenkis (VC) dimension (v), a measure of model complexity, must be asymptotically small relative to the sample size ($v/N \rightarrow 0$) in order to avoid overfitting (Vapnik, 1999; Vapnik, 2013). However, machine learning research increasingly uses high-dimensional data (images, natural languages, etc.) and complex models (DNN, etc.), leading to very large VC dimension and thus violating the classical asymptotic assumption. DNN is typically used in a non-asymptotic regime where the classical asymptotic assumption does not hold, implying that DNN should theoretically overfit models to data (Wainwright, 2019). While an increasing number of transportation studies use DNN to predict travel choices with high accuracy even on small data sets (Karlaftis and Vlahogianni, 2011; Hagenauer and Helbich, 2017; Cantarella and Luca, 2005; Dong et al., 2018; Mozolin, Thill, and Usery, 2000; Polson and Sokolov, 2017; Wu et al., 2018), this theoretical issue remains unresolved and there exist no practical guidelines as to what circumstances may result in overfitting issues when using DNN for choice analysis.

The second concern in transportation choice applications of DNN is its perceived lack of interpretability. Prediction is a typical focus of all modeling whether done via discrete choice models or machine learning classifiers, but many transportation applications require interpretation as well. Interpretability is important for researchers, who seek to understand findings on mode shares, elasticities, marginal rates of substitution, and social welfare, as well as the general public, among whom interpretability has been found useful in building trust (Lipton, 2016) and explaining results to users (Doshi-Velez and Kim, 2017). DNN is typically framed as a "black box" model, and it is ranked as the model with lowest interpretability among all machine learning classifiers (Kotsiantis, Zaharakis, and Pintelas, 2007; Lipton, 2016; Zhou et al., 2014). The majority of previous studies using DNN for transportation choice modeling have focused narrowly on using DNN to predict mode choice, activity choice, car ownership, or other individual choices (Hensher and Ton, 2000; Xie, Lu, and Parkany, 2003; Cantarella and Luca, 2005; Celikoglu, 2006; Omrani, 2015; Hagenauer and Helbich, 2017). Only a small number of transportation studies touched upon the interpretability of DNN in choice modeling, but do not provide explicit metrics to measure the quality of interpretability (Rao et al., 1998; Bentz and Merunka, 2000; Hagenauer and Helbich, 2017; S. Wang, Q. Wang, and Zhao, 2020a; S. Wang, Mo, and Zhao, 2020; S. Wang, Q. Wang, and Zhao, 2020b). The interpretability of DNN models, particularly in comparison to discrete choice models, will be a key factor in determining whether these approaches can be extended to transportation contexts beyond demand prediction and have practical implications on our understanding of individual decision-making behavior.

This paper seeks to address both of these issues through the development of a statistical learning theoretical framework consisting of two dimensions. The first dimension is the decomposition of estimation and approximation errors. We demonstrate that the estimation error of DNN architectures used in choice models is not very large, addressing the first overfitting issue. We present a proof that illustrates that the magnitude of the parameters in DNN is more important than the number of parameters in upper bounding the estimation error in a non-asymptotic way. The second dimension focuses on the relationship between the quality of prediction and of interpretation. Particularly, we substantiate the concept of interpretation by formulating the function approximation loss to measure interpretation quality as a counterpart to prediction, which we measure through zero-one and log loss. With our formulation, interpretation quality is measured by the difference between the true and the estimated choice probability functions, drawing on the fact that all valuable economic information can be derived from this function. Model interpretability in DNN relies on the full choice probability function based on automatically learned utility specification. This is in sharp contrast to traditional choice models, which are interpreted through the individual parameters chosen for the utility functions. One limitation of the interpretation quality metric is that it can only be measured in a simulated context, in which the true choice probability functions are known. Nonetheless, through our new conceptualization, we can evaluate models in terms of both prediction and interpretation qualities, allowing us to evaluate and demonstrate the potential for DNN to serve as a powerful predictive and interpretable tool for choice analysis research.

To illustrate this theoretical framework, we compare DNN to binary logit model (BNL), binary mixed logit model (BXL), multinomial logit model (MNL), and mixed logit model (MXL), representative discrete choice modeling approaches through four experiments. Three of these experiments use synthetic data with binary outputs in combination with Monte Carlo simulation, illustrating the tradeoffs between approximation and estimation errors as well as those between interpretation and prediction qualities under different sample sizes and input dimensions. The last experiment uses data from the National Household Travel Survey 2017 (NHTS 2017) in order to shed light on the practical relevance of this new theoretical framework, allowing us to provide practical suggestions for future DNN applications in choice modeling research. For now, we focus on the comparison of BNL, BXL, MNL, and MXL with DNN, but future research can investigate how other discrete choice models (DCMs) such as nested logit model and hybrid choice model compare to DNN.

This study makes contributions in three main areas. First, it introduces statistical learning theory to build novel theoretical foundations for DNN-based choice analysis in the transportation field. As opposed to classical statistical and econometric tools, these foundations enable choice modeling to tackle non-asymptotic data regimes, high-dimensional data, and complex models such as DNN. Second, we describe the two dimensions of approximation vs. estimation errors and predictive vs. interpretation qualities to evaluate DNNs, shedding light on the two issues of overfitting and interpretability. We define a metric of function approximation loss to measure the capacity of DNN-based choice models in capturing economic information, enabling direct comparison with BNL and MNL. The metrics we develop provide important benchmarks for future DNN choice modeling research to use and improve upon. Lastly, our experiments demonstrate the tradeoffs between DNNs' and DCMs' performance in terms of sample size, input dimensions, model complexity, and data generating process, providing specific modeling suggestions for future studies. Overall, our work illustrates a non-asymptotic statistical theory foundation, demonstrates the predictive and interpretable potential of DNNs, and informs their use beyond demand forecasting and in domains typically reserved for discrete choice models, such as policy and behavioral analysis.

The paper is organized as follows. In section 2, we describe in more detail the theoretical background and relevant past studies for our framework. In this section, we formulate evaluation metrics for interpretation which can be used for both DCM and DNN and then use statistical learning theory to characterize the four quadrants resulting from the dual tradeoffs between approximation error and estimation error and between prediction and interpretation qualities. The introduction of each quadrant is followed by the review of the previous studies most relevant to them. In section 3, we describe our three simulation experiments on synthetic data, illustrating the dynamics of the tradeoffs between the four quadrants. Then we apply our framework to the NHTS data and discuss the resulting findings. Section 4 concludes the paper with remarks on implications and future research.

2. Theory and Literature Review

2.1. Setup of DNN-Based Choice Modeling with Statistical Learning Theory

Let $s(x_i)$ denote the probability of individual i choosing alternative 1 out of $\{0, 1\}$ alternatives, and x_i the inputs including alternative- and individual-specific variables: $s(x_i) : R^d \rightarrow [0, 1]$. Individual choice $y_i \in \{0, 1\}$ is a Bernoulli random variable with $s(x_i)$ probability of choosing the alternative 1. This soft decision rule is a common assumption in choice analysis, and it is more generic than the hard decision rule that does not involve probabilistic decisions.[1] Let $f(x_i) : R^d \rightarrow \{0, 1\}$ represent the hard decision rule mapping. Let $\mathcal{F}_1$ denote the model class represented by a feedforward DNN, with the layer-by-layer feature transformation $\Phi_1(x_i, w) = (g_m \circ ...g_2 \circ g_1)(x_i)$, in which $g_j(x) = ReLU(\langle W_j, x\rangle)$ representing one standard module in the DNN consisting of ReLU activation and linear transformation. When DNN is applied to a binary choice case, the choice probability $s(x_i, w)$ becomes

$$s(x_i, w) = \sigma(\Phi_1(x_i, w)) = \frac{1}{1 + e^{-\Phi_1(x_i,w)}} \tag{1}$$

where σ is the Sigmoid activation function, and w represents all the coefficients in the DNN. Note that Φ_1 is similar to the deterministic utility difference $V_1 - V_0$ in choice models. With larger Φ_1, individual i is more likely to choose alternative 1 over 0. Let $\mathcal{F}_0$ represent the model class of BNL and $\Phi_0(x_i, w) = \langle w, x_i\rangle$ represent the linear feature mapping in BNL. It can be shown that BNL is a special case of DNN (shown in Appendix I): $\mathcal{F}_0 \subset \mathcal{F}_1$. The choice probability of $s(x_i)$ in BNL is similar to Equation 1, except for replacing Φ_1 with Φ_0. Let $S = \{x_i, y_i\}_{i=1}^N$ denote the sample; N the sample size; $x \sim P_x(x)$ the data generating process of x; and $s^*(x)$, $f^*(x)$, and w^* the true models and parameters. Empirical risk minimization is used to obtain their estimators: $\hat{s}(x)$, $\hat{f}(x)$, and $\hat{w}$.

Definition 1. *Empirical risk minimization (ERM) is defined as*

$$\min_{f\in\mathcal{F}} \hat{L}(f) = \min_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} l(y_i, f(x_i)) \tag{2}$$

Estimator based on ERM is defined as

$$\hat{f} = \operatorname*{argmin}_{f\in\mathcal{F}} \hat{L}(f) = \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} l(y_i, f(x_i)) \tag{3}$$

In training ERM, it is critical to choose a specific *expected loss function* $L(y, x) = \mathbb{E}_{x,y}[l(y, f(x))]$. The loss function is generally defined and chosen as log-loss, zero-one loss, hinge loss, mean squared errors, mean absolute errors, or generally defined L_p norm losses. To understand the out-of-sample performance of any estimator, we need to examine the *excess error*:

[1] An asymptotically soft decision rule with Softmax or Sigmoid activation function becomes a hard decision rule.

Definition 2. *Excess error of $\hat{f}$ is defined as*

$$\mathbb{E}_S[L(\hat{f}) - L(f^*)] \tag{4}$$

that of $\hat{s}$ is defined as

$$\mathbb{E}_S[L(\hat{s}) - L(s^*)] \tag{5}$$

$L(\hat{f})$ and $L(\hat{s})$ are the population error of the estimator, while $L(f^*)$ and $L(s^*)$ are the population error of the true model. Excess error measures to what extent the error of the estimator deviates from the true model, averaged over random sampling S. A tight upper bound of excess error can guarantee reliable out-of-sample performance. In the following discussions, we will mainly use f^* and $\hat{f}$ as the running examples, but all the following arguments apply to s^* and $\hat{s}$. Excess error can be decomposed into estimation error and approximation error as following.

$$\mathbb{E}_S[L(\hat{f}) - L(f^*)] = \mathbb{E}_S[L(\hat{f}) - L(f^*_F)] + \mathbb{E}_S[L(f^*_F) - L(f^*)] \tag{6}$$

where $f^*_F = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\, L(f)$, the best function in function class $\mathcal{F}$ to approximate f^*.

Definition 3. *Estimation error refers to*

$$\mathbb{E}_S[L(\hat{f}) - L(f^*_F)] \tag{7}$$

Definition 4. *Approximation error refers to*

$$\mathbb{E}_S[L(f^*_F) - L(f^*)] \tag{8}$$

Estimation error is the first term in Equation 6 and approximation error is the second term. Estimation error is a quantity that measures whether $\hat{f}$ overfits: very large $L(\hat{f}) - L(f^*_F)$ implies serious overfitting. Since the estimation error has a $\hat{f}$ term, it captures the randomness from sampling and training. Approximation error is more deterministic and captures only the difference between the best function f^*_F in function class $\mathcal{F}$ and the true function f^*. The following four subsections will introduce the zero-one and log losses measuring prediction quality, the function approximation loss measuring interpretation quality, the approximation error, and the estimation error of DNNs in order.

2.2. Prediction Quality

Prediction quality can be measured by zero-one and log losses.

Definition 5. *Zero-one loss is defined as*

$$L_{0/1}(f) = \mathbb{E}_{x,y}[\mathbb{1}\{y \neq f(x)\}] \tag{9}$$

Empirical zero-one loss is defined as

$$\hat{L}_{0/1}(f) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{y_i \neq f(x_i)\} \tag{10}$$

Definition 6. *Log loss is defined as*

$$L_{log}(f) = \mathbb{E}_{x,y}[y \log s(x)] \tag{11}$$

Empirical log loss is defined as

$$\hat{L}_{log}(f) = \frac{1}{N}\sum_{i=1}^{N} y_i \log s(x_i) \tag{12}$$

Both zero-one and log losses are common metrics to evaluate prediction performance. The zero-one loss is widely used in the machine learning community (Cheng et al., 2019; Tang, Xiong, and Zhang, 2015; Paredes et al., 2017; Allahviranloo and Recker, 2013; Hagenauer and Helbich, 2017; Cantarella and Luca, 2005; Hensher and Ton, 2000), and the log loss, which is the same as the negative log-likelihood and the cross-entropy loss, is predominant in the classical discrete choice models (Ben-Akiva and Lerman, 1985; Train, 1980). Besides these two, prediction quality can also be measured by other metrics, such as Brier score (Harrell, 2015) or pseudo R square, which are not discussed in this work.

This study does not intend to normatively judge between the two metrics, since each has their own pros and cons. From a choice modeling perspective, log loss is recommended because it respects the probabilistic nature of individual decision-making (Train, 2009), constitutes a convex objective function to guarantee the unique identification of parameters (Harrell, 2015; Hillel, Bierlaire, et al., 2020), and connects to information and statistical theories (Gneiting and Raftery, 2007). In the machine learning field, the zero-one loss dominates (Krizhevsky, Sutskever, and G. E. Hinton, 2012; LeCun, Bengio, and G. Hinton, 2015; He et al., 2016). A practical reason is that the zero-one loss enables a wide comparison between probabilistic and deterministic models. For example, the zero-one loss can be used to compare DNNs to support vector machines (SVM) and K-nearest neighbors (KNN), while the log loss enables the comparison across probabilistic models (Hillel, Bierlaire, et al., 2019). The probabilistic models enable researchers to compute elasticities for behavioral analysis, while the deterministic classifiers cannot generate continuous gradients. The choice between deterministic and probabilistic models is also context-specific. The probabilistic models (e.g. discrete choice models) can be used for population and sub-population forecasting, since the aggregate choice probabilities are unbiased (Train, 2009). The deterministic classifiers (e.g. KNN, SVM) can fit individual-level forecasting, but the aggregate market share can be severely biased. Theoretically, the zero-one loss is bounded between zero and one, which facilitates the derivation of its estimation error bounds, while the unbounded log loss, due to the $\log s(x)$ term, creates difficulty. A comparison of these two metrics would involve deeper discussions about the classical and the modern statistics, which is beyond the scope of this work. Therefore, this work uses both metrics, thus keeping the discussion open for future researchers.

2.3. Interpretation Quality

Definition 7. *Interpretation quality is measured by the function approximation loss, which is defined as the difference between true and estimated choice probability functions*

$$L_e(s) = ||s^* - s||^2_{L^2(P_x)} = \int_x (s^*(x) - s(x))^2 dP(x) \tag{13}$$

Empirical function approximation loss is defined as

$$\hat{L}_e(s) = \frac{1}{N}\sum_{i=1}^{N}(s^*(x_i) - s(x_i))^2 \tag{14}$$

Interpretation quality is measured by the difference between true and estimated choice probability, integrated over domain $\mathcal{X}$ and weighted by $P_x(x)$. Note the function approximation loss is categorically different from the metrics measuring the prediction quality, since the true choice probability function $s^*(x_i)$ in Definition 7 can only be obtained in simulated contexts. We choose to use this measurement because researchers can obtain complete economic information through the choice probability function $s(x)$, as argued in S. Wang, Q. Wang, and Zhao (2020a). For example, the probability derivatives of choosing alternative 1 with respect to price x_j can be computed as the derivative $\frac{ds(x)}{dx_j}$; its associated elasticity is $\frac{d\log s(x)}{d\log x_j}$; value of travel time savings (VTTS) can be computed as ratio of two derivatives $\frac{ds(x)/dx_{j1}}{ds(x)/dx_{j2}}$; the utility difference can be computed by using inversed Sigmoid function $V_1 - V_0 = \sigma^{-1}(s)$; or the empirical market share of alternative 1 can be computed by $\sum_{i=1}^{N} s(x_i)$. Therefore, an accurate function estimator $\hat{s}(x)$ could help recover elasticity values, marginal rate of substitution (such as VTTS), market share, utility values, and social welfare, which provide most of the economic information needed in practice.

It is crucial to see that we focus on *function* estimation $\hat{s}(x)$ rather than *parameter* estimation $\hat{w}$, which is the traditional focus of the majority of the econometric models. The focus on parameter estimation is nearly impossible for DNN for at least three reasons. First, a simple feedforward DNN could easily have tens of thousands parameters, and this large number renders it impossible for researchers to discuss individual parameters. Second, DNN has the property called symmetry of parameter space (Bishop, 2006), implying that different parameters could lead to the same choice probability function $s(x)$. Therefore, interpreting individual parameters w is vacuous in DNN. Third, studies have shown that semantic information cannot be revealed from individual neurons, but from the space of each layer in DNN (Szegedy et al., 2014). A large number of studies used the function estimators in DNN for interpretation, while none used individual neurons/parameters (Montavon, Samek, and Muller, 2018; G. Hinton, Vinyals, and Dean, 2015; Baehrens et al., 2010; Ross and Doshi-Velez, 2018). Mullainathan and Spiess (2017) argued that ML classifiers (including

DNN) are categorically different from econometric models since the ML classifiers focus on $\hat{y}$ while the econometric models focus on $\hat{w}$. This is generally true; however, in the case of DNN, an accurate estimator of the choice probability function $\hat{s}(x)$ could satisfy most of our interpretation purposes traditionally achieved through using $\hat{w}$. In fact, several studies in the transportation field have visualized or computed the gradient information of the choice probability functions to interpret the ML classifiers, supporting our definition of the function approximation loss based on the choice probability functions (Rao et al., 1998; Bentz and Merunka, 2000; Hagenauer and Helbich, 2017). Moreover, the process of interpreting elasticity $\frac{ds}{dx_j}$ is the same as the discussion of using input gradients in the ML community (Baehrens et al., 2010; Montavon, Samek, and Muller, 2018). Therefore, the shifting focus from parameter to function estimation enables researchers to interpret DNN results in choice analysis context, and this shift is both inevitable and desirable.

Whereas our definition for interpretation quality captures the key economic information through the choice probability function, it is not the only way to define model interpretability. Lipton (2016) discussed multiple aspects of interpretability, including simulatability, decomposability, algorithmic transparency, and post-hoc interpretability. Our definition is focused on the post-hoc interpretability restricted to only economic information, and does not address the other aspects of interpretability and other types of information obtained by using post-hoc interpretation methods (Ribeiro, Singh, and Guestrin, 2016; Montavon, Samek, and Muller, 2018; G. Hinton, Vinyals, and Dean, 2015). Whereas our approach aligns with the long tradition of choice modeling, it is possible to define interpretability in other ways, as shown in a recent working paper by Bertsimas et al. (2019).

2.4. Approximation Error

Since BNL is one subset of DNN ($\mathcal{F}_0 \subset \mathcal{F}_1$) (shown in Figure 1), the approximation error of DNN is always smaller than BNL (Vapnik, 1999). Intuitively, the best model ($f^*_{\mathcal{F}_0}$) in $\mathcal{F}_0$ is also in $\mathcal{F}_1$, so it is generally true that $f^*_{\mathcal{F}_1}$ can approximate f^* better than $f^*_{\mathcal{F}_0}$. Formally,

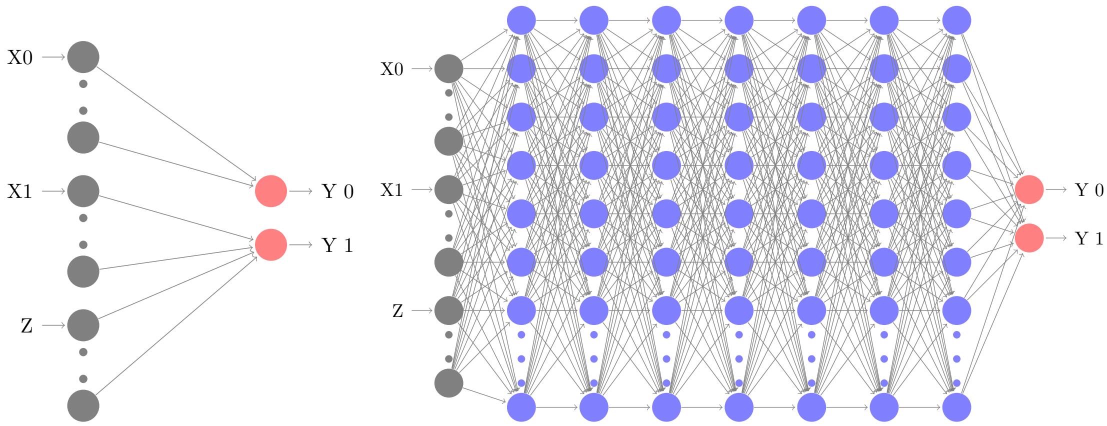
Proposition 1. *The approximation error of the zero-one loss in DNN is always smaller than that in BNL*

$$\mathbb{E}_S[L_{0/1}(f^*_{\mathcal{F}_1}) - L_{0/1}(f^*)] \leq \mathbb{E}_S[L_{0/1}(f^*_{\mathcal{F}_0}) - L_{0/1}(f^*)] \tag{15}$$

Similarly, the approximation error of the function approximation loss in DNN is also smaller than that in BNL:

$$\mathbb{E}_S[L_e(s^*_{\mathcal{F}_1}) - L_e(s^*)] \leq \mathbb{E}_S[L_e(s^*_{\mathcal{F}_0}) - L_e(s^*)] \tag{16}$$

While these results are not difficult to see, they can be understood from various mathematical perspectives. The first perspective is the *universal approximator theorem* of DNN, developed in the 1990s. The studies suggest that even a shallow neural network (SNN) is asymptotically a universal approximator when the width becomes infinite (Cybenko, 1989; Hornik, Stinchcombe, and White, 1989; Hornik, 1991). Recently, this asymptotic perspective leads to a more non-asymptotic question, asking why depth is necessary for SNN to be powerful enough for practical use cases. Research has demonstrated that DNN can approximate functions with an exponentially smaller number of neurons than SNN in many settings (Cohen et al., 2016; Rolnick and Tegmark, 2017; Poggio, Mhaskar, et al., 2017). This perspective is quite relevant to our focus, since BNL is one type of SNN (Bentz and Merunka, 2000). The choice between DNN and BNL can equivalently be framed as the choice between DNN and SNN.



(a) F0 One-Layer Sparse NN (BNL)

(b) F1 Deep Dense Feedforward NN (DNN)

Fig. 1. Two architectures of BNL and DNN; first graph represents BNL with linear specification, second graph is DNN. Visually, DNN is an extension of BNL, as is its function class. The red neurons in both graphs visualize utility values, and the blue neurons in DNN can be seen as the process of specifying utility.

In addition to these mathematical perspectives, we highlight the economic perspective that describes the similarity between BNL and DNN, as well as their difference between *automatic* and *handcrafted* utility specification. BNL and DNN are categorically similar since both involve the process of specifying and comparing utility values. The notion of utility specification and comparison in choice modeling setting is widely known (Train, 2009; Ben-Akiva and Lerman, 1985), but they can also be applied to DNN. In fact, the last layer of DNN could be named as utilities and the layers before the last can be seen as utility specifications. However, their key difference is that traditional BNL approaches use handcrafted utility specification based on experts' prior knowledge, while DNN automatically learns utility specification based on a complex model assumption. The automatic learning in a standard DNN architecture can capture a nonlinear relationship, and the handcrafted linear BNL can detect only the linear relationship, although it can always be enriched with feature transformation to capture a nonlinear relationship. Automatic feature learning is nearly inevitable in many tasks such as face recognition, in which handcrafting features of human faces seem nearly impossible (Mullainathan and Spiess, 2017). Studies in the ML community typically praise the power of this automatic feature learning, although it is still a heated debate whether researchers should rely on only the automatic feature learning or a mixture of automatic and handcrafted feature learning (LeCun, Bengio, and G. Hinton, 2015; Bengio, Courville, and Vincent, 2013; Liao and Poggio, 2018). However, the bottom line is that a pure handcrafted utility specification will not be able to maximize the predictive and interpretable power of the data, and using or at least augmenting the power of the automatic feature learning in DNN could greatly add to the future of choice modeling practice.

2.5. Estimation Error

The more challenging question is about the estimation error of DNN, particularly because the smaller approximation error is always associated with larger estimation errors. Specifically, the question is whether DNN has well-bounded estimation error, when the number of its parameters is so large. To address this question, we will present two proofs. While both rely on empirical process theory, the first uses contraction inequality, which provides a tighter upper bound than the second proof, which is based on VC dimension. The proof based on empirical process theory shows that the estimation error of both zero-one and function approximation losses in DNN can be bounded or at least controlled by l_1 and l_2 regularizations. We believe this part is critical since the empirical process theory provides a new foundation for future studies that rely on high-dimensional data and complex statistical models. We put only the key propositions in the following section, with detailed proofs provided in Appendix II.

Definition 8. *Empirical Rademacher complexity is defined as*

$$\hat{\mathcal{R}}_n(\mathcal{F}|_S) = \mathbb{E}_\epsilon \Big[\sup_{f \in \mathcal{F}} \Big| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i f(x_i) \Big| \Big] \tag{17}$$

$\epsilon_i \in \{+1, -1\}$ *with probabilties* $[0.5, 0.5]$*;* $\mathcal{F}|_S$ *denotes the function class* $\mathcal{F}$ *projected to sample* S*.*

Proposition 2. *The upper bound of the estimation error of* $\hat{f}$ *can be given by the Rademacher complexity*

$$\mathbb{E}_S[L(\hat{f}) - L(f_F^*)] \leq 2\mathbb{E}_S \hat{\mathcal{R}}_n(l \circ \mathcal{F}|_S) \tag{18}$$

Proof of Proposition 2 is available in Appendix II.A. Rademacher complexity measures the complexity of function class $\mathcal{F}$ conditioning on the dataset S. Proposition 2 shows that the upper bound of the estimation error is provided by the complexity of the function class $l \circ \mathcal{F}$, defined as $l \circ \mathcal{F} = \{l \circ f(x) | f(x) \in \mathcal{F}\}$. Intuitively, it is harder to search for the best model $\hat{f}$, as the function class $\mathcal{F}$ becomes larger. It also aligns with traditional statistics, as higher VC dimension or more parameters (more complexity of function class) leads to larger estimation errors. Specifically, Proposition 2 can be used as an upper bound for the estimation errors of zero-one and function approximation losses:

Proposition 3. *The upper bound of the estimation error of the zero-one loss is found as*

$$\mathbb{E}_S[L_{0/1}(\hat{f}) - \hat{L}_\gamma(\hat{f})] \leq \frac{2}{\gamma}\mathbb{E}_S\hat{\mathcal{R}}_n(\mathcal{F}|_S) \tag{19}$$

Proposition 4. *The upper bound of the estimation error of the function approximation loss is given by*

$$\mathbb{E}_S[L_e(\hat{s}) - L_e(s_F^*)] \leq 4\mathbb{E}_S\hat{\mathcal{R}}_n(\mathcal{F}|_S) \tag{20}$$

Proof of Propositions 3 and 4 is available in Appendix II.B and II.C. Proposition 3 provides an upper bound on $\mathbb{E}_S[L_{0/1}(\hat{f})]$ by using γ-margin error (definition in Appendix II.B). While the left hand side is not exactly the same as $\mathbb{E}_S[L_{0/1}(\hat{f}) - L_{0/1}(f_F^*)]$, both $\hat{L}_\gamma(\hat{f})$ and $\frac{2}{\gamma}\mathbb{E}_S\hat{\mathcal{R}}_n(\mathcal{F}|_S)$ can be computed in practice. Compared to the estimation error of the zero-one loss, the interpretation part is easier, and Proposition 4 demonstrate that Rademacher complexity provides an upper bound on the estimation error of the function approximation loss up to a constant. One remaining question is how to provide an effective upper bound on Rademacher complexity of DNN.

Proposition 5. *Let H_d be the class of neural network with depth D over the domain $\mathcal{X}$ ($x \in B_1^{(d_0)}$), where each parameter matrix W_j has Frobenius norm at most $M_F(j)$ and its one-infinity norm at most $M(j)$, and with ReLU activation functions. Then by using contraction inequality, the upper bound of the Rademacher complexity of DNN ($\mathcal{F}_1$) is given by*

$$\hat{\mathcal{R}}_n(\mathcal{F}_1|_S) \lesssim O(\frac{\sqrt{\log d_0} \times \prod_{j=1}^{D} 2M(j)}{\sqrt{N}}) \tag{21}$$

The tightest bound found in the literature Golowich, Rakhlin, and Shamir, 2017 is:

$$\hat{\mathcal{R}}_n(\mathcal{F}_1|_S) \lesssim \frac{\sqrt{\log d_0} \times (\sqrt{2\log D} + 1) \times \prod_{j=1}^{D} M_F(j)}{\sqrt{N}} \tag{22}$$

Proposition 6. *Rademacher complexity of DNN with* 0/1 *loss has an upper bound determined by VC dimension*

$$\hat{\mathcal{R}}_n(l \circ \mathcal{F}_1) \lesssim 4\sqrt{\frac{v\log(N+1)}{N}} \lesssim 4\sqrt{\frac{TD\log(T) \times \log(N+1)}{N}} \tag{23}$$

with T denoting the total number of parameters and D the depth of DNN Bartlett, Harvey, et al., 2017.

Proposition 5 describes the important factors that influence the upper bound on estimation error, including the input dimension d_0, norm of parameters in each layer $M(j)$ or $M_F(j)$, and sample size. The result is intuitive: with larger sample size, and smaller input dimension and norm of parameters, the estimation error of DNN is more likely to be bounded. Proof of Propositions 5 and 6 is available in Appendix II.D and II.E.

The most important message about estimation error is revealed by the difference between Propositions 5 and 6: instead of computing the ratio of v and N as in Proposition 6, researchers could compute the *norms* of the coefficients in each layer to upper bound estimation error, as in Proposition 5. The total number of parameters is fixed when researchers choose one specific DNN architecture, so it is hard to control the Rademacher complexity through VC dimension. On the contrary, the norms of the weights in each layer $M(j)$ can be controlled by l_1 or l_2 regularization. Therefore, Proposition 5 along with Propositions 3 and 4 provide valid and much tighter upper bounds on estimation error than the traditional VC dimension perspective.

The results above heavily rely on the progresses in non-asymptotic statistical learning theory and particularly the empirical process theory in the recent two decades. Readers should refer to (Bousquet, Boucheron, and Lugosi, 2004; Von Luxburg and Schölkopf, 2011; Anthony and Bartlett, 2009; Wainwright, 2019; Vapnik, 2013) for general introductions; to (Vapnik, 1999; Vapnik, 2013; Sontag, 1998; Bartlett, Harvey, et al., 2017) for the proof about Rademacher complexity bound of DNN based on VC dimension; to (Golowich, Rakhlin, and Shamir, 2017; Neyshabur, Tomioka, and Srebro, 2015; Bartlett and Mendelson, 2002; Bartlett, Jordan, and McAuliffe, 2006) for the proof about Rademacher complexity bound of DNN based on contraction inequality.

2.6. Summary

So far we have provided concrete mathematical formulation and theoretical discussions for the two dimensions and four quadrants that define our theoretical framework, as summarized in Table 1. Both dimensions are important from a historical view. The tradeoff between estimation and approximation error is the first order decomposition in statistical learning theory (Vapnik, 1999; Vapnik, 2013; Von Luxburg and Schölkopf, 2011). Prediction vs. interpretation marks the difference of two statistical cultures, as pointed out by Breiman (2001), and is recently remarked again by Mullainathan and Spiess (2017). For the purpose of our study, the two dimensions can be used to bridge the classical low-dimensional DCMs and the new high-dimensional DNN models from a theoretical perspective.

	Approximation Error	**Estimation Error**
Prediction Quality	Approximation error of zero-one and log losses $\mathbb{E}_S[L_{0/1}(f^*_F) - L_{0/1}(f^*)]$ $\mathbb{E}_S[L_{log}(f^*_F) - L_{log}(f^*)]$	Estimation error of zero-one and log losses $\mathbb{E}_S[L_{0/1}(\hat{f}) - L_{0/1}(f^*_F)]$ $\mathbb{E}_S[L_{log}(\hat{f}) - L_{log}(f^*_F)]$
Interpretation Quality	Approximation error of function approximation loss $\mathbb{E}_S[L_e(s^*_F) - L_e(s^*)]$	Estimation error of function approximation loss $\mathbb{E}_S[L_e(\hat{s}) - L_e(s^*_F)]$

Table 1: Two dimensions of the theoretical framework

3. Experiments

3.1. Design of Experiments

The experiments consist of two parts: one with three simulated datasets and one with the NHTS dataset. The experiments with simulated and real datasets are complementary in terms of their purposes. With Monte Carlo simulation, the underlying true data generating process (DGP; e.g. $s^*(x)$ or $f^*(x)$) is known, so we can compute both the approximation and estimation errors related to $s^*(x)$ and $f^*(x)$, which cannot be done in the experiment with real datasets. On the other hand, real datasets reveal the real decision making process, which has to be presumed, sometimes arbitrarily, in Monte Carlo simulations.

In the experiments with synthetic data, we compare one DNN architecture with fixed hyperparameters to one BNL model and one BXL model with linear utility specification. The DNN architecture has 5 layers, 100 neurons in each layer, and ReLU activation functions. The DNN training uses standard ERM procedure, with He initialization (He et al., 2015), Adam optimization (Kingma and Ba, 2014), and mild regularizations. The BNL and BXL in all our simulations uses only linear specification. Again, this linear specification does not limit the generality of our discussion since any domain knowledge based utility specification could always be provided to DNN as inputs. DNN's theoretical properties do not vary much with the specific choice of parameters and hyperparameters. The following discussions focus on comparing BNL and DNN, since their properties are derived in Section 2, but we will also incorporate the BXL's empirical results.

The experiment with Monte Carlo simulation consists of three scenarios, representing three typical cases researchers face in reality. The three scenarios are differentiated by the "location" of the true DGP with respect to $\mathcal{F}_0$ and $\mathcal{F}_1$: (1) $f^* \in F_0$ and $f^* \in F_1$; (2) $f^* \notin F_0$ and $f^* \in F_1$; (3) $f^* \notin F_0$ and $f^* \notin F_1$. Scenario 1 represents the case in which a simple BNL is the true DGP, which belongs to both model classes of BNL and DNN, so the approximation errors of both BNL and DNN are zero. Secnario 2 represents the case in which the true DGP is more complicated than BNL, so the approximation error of BNL is larger than zero while that of DNN is still zero. Scenario 2 commonly happens when information is complete while the function used in model training is misspecified in choice modeling. Scenario 3 represents the case in which both BNL and DNN have strictly positive approximation errors, which happens when important variables are omitted, traditionally called omitted variable bias. In terms of function relationship between $\mathcal{F}_0$, $\mathcal{F}_1$ and f^*, the three scenarios are exhaustive. Our simulation also varies sample size and number of input variables to demonstrate how estimation error changes, based on our theory about estimation error of DNN (Proposition 5). Sample size in the Monte Carlo simulations ranges from 100, the smallest possible one in a survey, to 1 million, the largest number observed in existing transportation questionnaire-based or observational surveys. The number of input variables is either 20 or 50, typical in choice analysis. For each experiment, we analyze the four quadrants, estimation and approximation errors of prediction and interpretation qualities, mapping back to our theoretical framework in Table 1. More details of the simulation are attached in Appendix III.

In the experiment with the NHTS dataset, we compare DNN, BNL, BXL, MNL, MXL, and three other machine learning classifiers in predicting trip purpose choices, a prevalent travel behavior analyzed in the past studies (De Dios Ortuzar and Willumsen, 2011; Zegras, 2010; Cervero and Kockelman, 1997), with varying sample size from 100 to $500,000$. The NHTS dataset is chosen since it covers the whole United States and it is one of the only datasets that has a sample size on the order of magnitude of 1 million. Because of the absence of a true data-generating process, the decomposition of estimation and approximation errors is impossible for the experiment with the real datasets, but we discuss both the prediction and interpretation of DNN-based choice models.

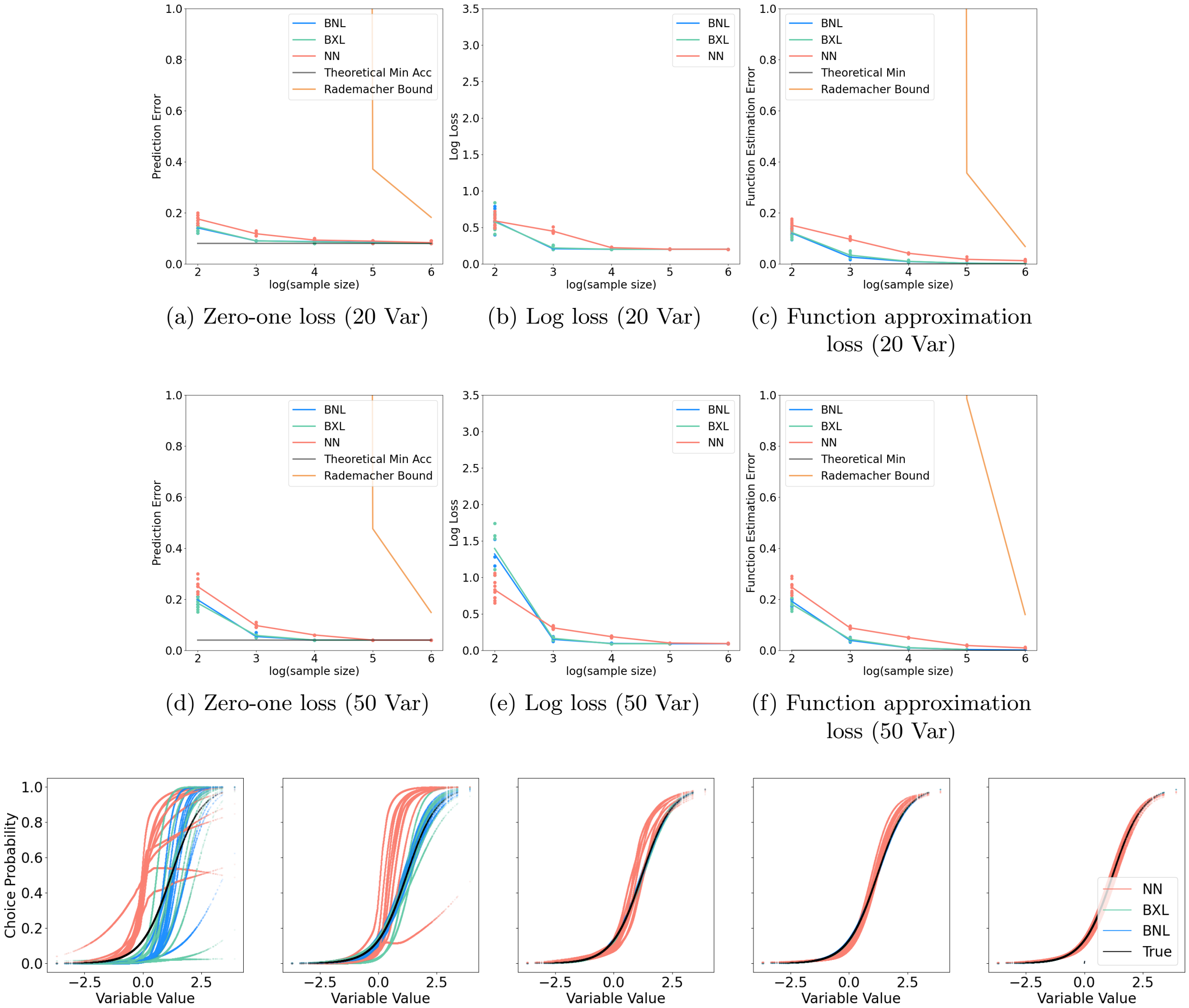
3.2. Three Experiments with Simulated Datasets

3.2.1. Scenario 1

In scenario 1, $s^*(x) = \sigma(\langle w, x \rangle)$, in which σ is the Sigmoid function, w is randomly generated variables taking $\{-1, +1\}$ values with equal probabilities, and x is generated as multivariate Gaussian distribution. In Figure 2, the two upper rows show the zero-one, log, and the function approximation losses of simulations with 20 and 50 input variables. In each subfigure, y-axis represents the values of the losses; x-axis represents sample size; each dot is a training result with the red ones representing DNN, blue ones representing BNL, and green ones representing BXL; red, blue, and green curves are the average values of the losses conditioning on a sample size. The black dashed line represents the minimum possible loss, which measures the amount of randomness in each DGP. In scenario 1, the gap between the red curve and the dashed black line is the estimation error, since it is exactly $\mathbb{E}_S[L(\hat{f}) - L(f_F^*)]$.[2] The yellow curves represent the theoretical upper bound on estimation error, based on Proposition 5. The lower row of Figure 2 shows the relationship between choice probabilities and an input variable with varying sample sizes from 100 to $1,000,000$. In each subfigure, the black curve represents the true $s^*(x)$; each red curve represents the estimated function $\hat{s}(x)$ from DNN, each blue one represents that from BNL, and each green one represents that from BXL. Since BNL and BXL perform highly similarly throughout the three scenarios, the following discussion focuses on only the comparison of DNN and BNL.

The estimation error of zero-one, log, and function approximation losses in both DNN and BNL converges to zero as sample size increases, and the convergence of DNN's estimation error is only slightly slower than that of BNL, as shown in Figures from 2a to 2f. It is not surprising that estimation errors always decrease as sample sizes increase since Equations 19 and 20 imply that larger sample size leads to smaller out-of-sample zero-one and function approximation losses. What is surprising is that the convergence of DNN is only *marginally* slower than BNL, particularly as examined from the classical statistical perspective since the number of parameters in the DNN is about $2,000$ times more than the parsimonious BNL model. More precisely, the VC dimension of our DNN architecture $v = 50,000 \times 5 \times \log(50,000) \simeq 3Million$ (Equation 23), which is larger than any sample size we use and far-off from the classical asymptotic data regime. On the contrary, the

[2] In scenario 1, f_F^* is the same as f^*. Hence $L(f_F^*)$ is represented by the black dash line



(a) Zero-one loss (20 Var)

(b) Log loss (20 Var)

(c) Function approximation loss (20 Var)

(d) Zero-one loss (50 Var)

(e) Log loss (50 Var)

(f) Function approximation loss (50 Var)

(g) Choice probability curves (20 Var); Sample size = 100, 1000, 10000, 100000, 1000000

Fig. 2. Scenario 1. Two upper rows: comparison of DNN, BNL, and BXL for prediction and interpretation qualities; Third row: visualizing how choice probabilities change with inputs; red curves: DNN, blue curves: BNL, green curves: BXL, black curves: true models. The figures in the two upper rows map to the theoretical framework in Table 1: the difference between red and black curves is the prediction/interpretation qualities of DNN, which equal to only their estimation errors in this scenario since the approximation errors are zero.

theoretical upper bound based on contraction inequality (Propositions 5 and 6) is represented by the yellow curve, which is much tighter than that based on the VC dimension, although it is still quite loose compared to the simulation results. Therefore, the simulation results resonate with our theoretical discussion that number of parameters in DNN should not be used to measure its estimation error bound. Empirically, DNN and BNL need roughly the same amount of data for accurate interpretation and prediction. With 20 or 50 variables, at least about 10^4 samples are needed so that the prediction and interpretation qualities of DNN become close to the theoretical optimum.

While this 10^4 sample size is larger than the sample sizes commonly obtained by questionnaire surveys, it is not unattainable; for instance, NHTS dataset has about $700,000$ observations, which is much larger than 10^4.

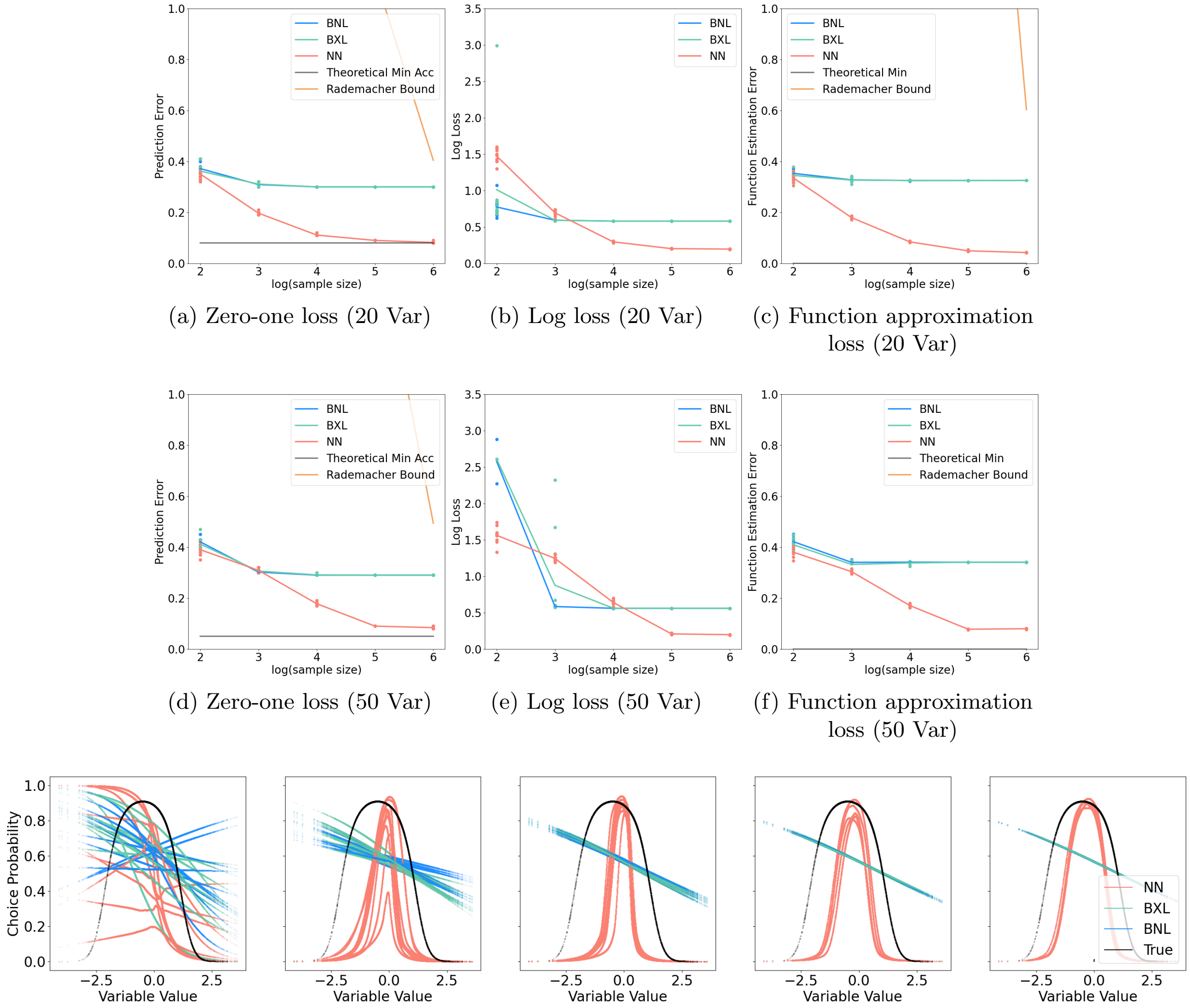
To interpret DNN results, we visualize the relationship between $\hat{s}(x)$ and one input variable x_j, as shown in Figure 2g. This method of visualizing sensitivity of $\hat{s}(x)$ with respect to x_j has been used for interpreting DNN results in several studies (Rao et al., 1998; Bentz and Merunka, 2000; Montavon, Samek, and Muller, 2018; S. Wang, Q. Wang, and Zhao, 2020a; S. Wang, Q. Wang, and Zhao, 2020b). Again, the estimated $\hat{s}(x)$ from DNN converges very quickly towards the true $s(x)$, and it captures the S-shaped choice probability curve and the linear utility specification, even when it is not *a priori* specified as linear. Overall, when researchers are very confident that prior expert knowledge has captured *every* piece of information, the BNL with handcrafted features perform better in prediction and interpretation, although DNN is only slightly worse.

3.2.2. Scenario 2

A more realistic case is the scenario in which researchers cannot correctly specify the utility function. In scenario 2, $s^*(x) = \sigma(w'\phi(x))$, in which $\phi(x)$ takes the quadratic transformation: $\phi([x_1, x_2, ..., x_d]) = [x_1, x_2, ...x_d, x_1^2, x_2^2, ...x_d^2])$. Then BNL $\mathcal{F}_0$ has the misspecification error, while DNN $\mathcal{F}_1$ does not. The results are visualized in Figure 3, and the formats of Figure 3 is exactly the same as Figure 2.

In scenario 2, DNN dominates BNL in terms of both prediction and interpretation qualities, even at a relatively small sample size, as shown in Figures from 3a to 3f. Except for the evaluation with the log loss using a small sample size ($\leq 10^3$), DNN can outperform BNL in terms of the zero-one, log, and function approximation losses. The key reason of DNN's dominance is its zero approximation error, in contrast to the large approximation error of BNL, measured by the gap between the theoretical minimum and the loss value that the blue curve converges to. Sample size is still a critical factor for both BNL and DNN, although it differs in terms of the critical magnitude for each. BNL converges to its optimum value ($f^*_{\mathcal{F}_0}$) with only about 10^3 observations, while DNN converges to its optimum ($f^*_{\mathcal{F}_1} = f^*$) when sample size reaches about 10^5 or 10^6. This result is very consistent with our theoretical discussion. BNL aligns with classical statistics, and as v/N is small, its estimation error is small. This result also implies that low dimensional statistical tools such as BNL cannot unleash the predictive power of a dataset with a large sample size. Only very complicated models such as DNN can fully unleash the predictive and interpretative power of a large sample.

Figure 3g visualizes the relationship between $\hat{s}(x)$ and an input variable x with varying sample sizes. With function misspecification, it is impossible for BNL to recover the true pattern even if the sample size becomes very large, while DNN with the power of automatic utility specification can gradually learn the underlying true utility specification, even without prior domain knowledge. Consistent with Figures 3c and 3f, DNN needs the sample size at the scale of about 10^5 and 10^6 to recover the true pattern of choice probability functions. Due to the misspecification and its



(a) Zero-one loss (20 Var)

(b) Log loss (20 Var)

(c) Function approximation loss (20 Var)

(d) Zero-one loss (50 Var)

(e) Log loss (50 Var)

(f) Function approximation loss (50 Var)

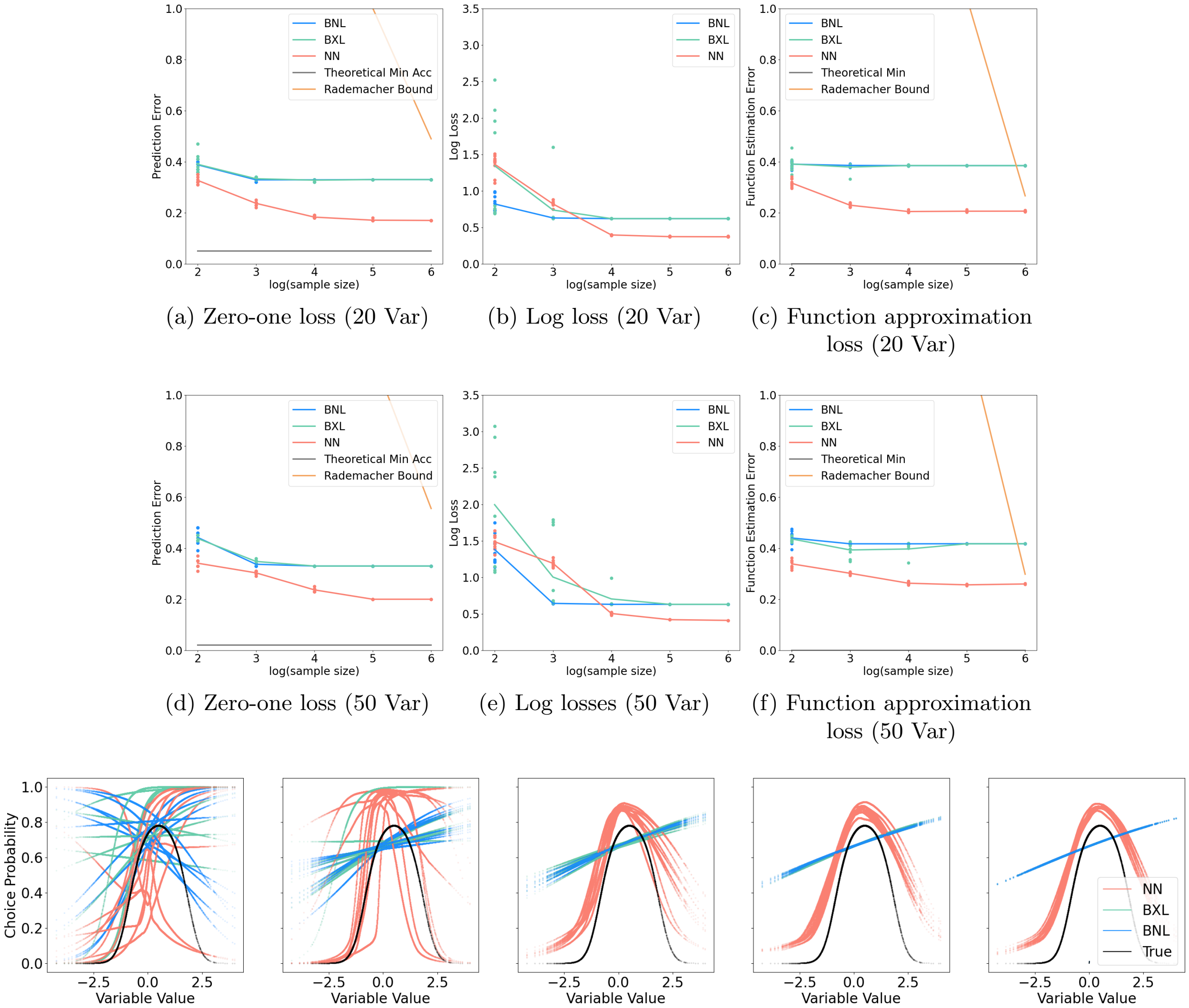
(g) Choice probability curves (20 Var); sample size = 100, 1000, 10000, 100000, 1000000

Fig. 3. Scenario 2. Two upper rows: comparison of DNN, BNL, and BXL for zero-one, log and function approximation losses; lower row: visualizing how choice probabilities change with inputs; red curves: DNN, blue curves: BNL, green curves: BXL, black curves: true models. The figures in two upper rows map to the theoretical framework in Table 1. Different from Scenario 1, BNL has approximation errors since the blue curves cannot converge to the theoretical minimum values, whereas DNN has no approximation errors.

corresponding approximation error in BNL, it is possible for DNN to outperform BNL in terms of both prediction and interpretation even when sample size is very small.

3.2.3. Scenario 3

A even more realistic case is the scenario in which researchers can neither collect the full information nor correctly specify the utility function ($f^* \notin F_0$ and $f^* \notin F_1$). In scenario 3, $s^*(x) = \sigma(w'\phi(x))$, where $\phi(x) = [1, x_1, x_2, ..., x_d, {x_1}^2, {x_2}^2, ...{x_d}^2, x_1x_2, ...x_{d-1}x_d]$, which includes both quadratic trans-



(a) Zero-one loss (20 Var)

(b) Log loss (20 Var)

(c) Function approximation loss (20 Var)

(d) Zero-one loss (50 Var)

(e) Log losses (50 Var)

(f) Function approximation loss (50 Var)

(g) Choice probability curves (20 Var); sample size = 100, 1000, 10000, 100000, 1000000

Fig. 4. Scenario 3. Two upper rows: comparison of DNN, BNL, and BXL for zero-one, log, and function approximation losses; lower row: visualizing how choice probabilities change with inputs; red curves: DNN, blue curves: BNL, green curves: BXL, black curves: true models. Different from scenario 1 and 2, both BNL and DNN have approximation errors since the red curves cannot converge to the theoretical minimum values.

formation and interaction terms. To make $f^* \notin F_1$, we randomly drop 5 variables out of 20 and 20 variables out of 50 in training, so that even $f^*_{\mathcal{F}_1}$ cannot approximate f^* well. Results are visualized in Figure 4, with the same format as previous ones.

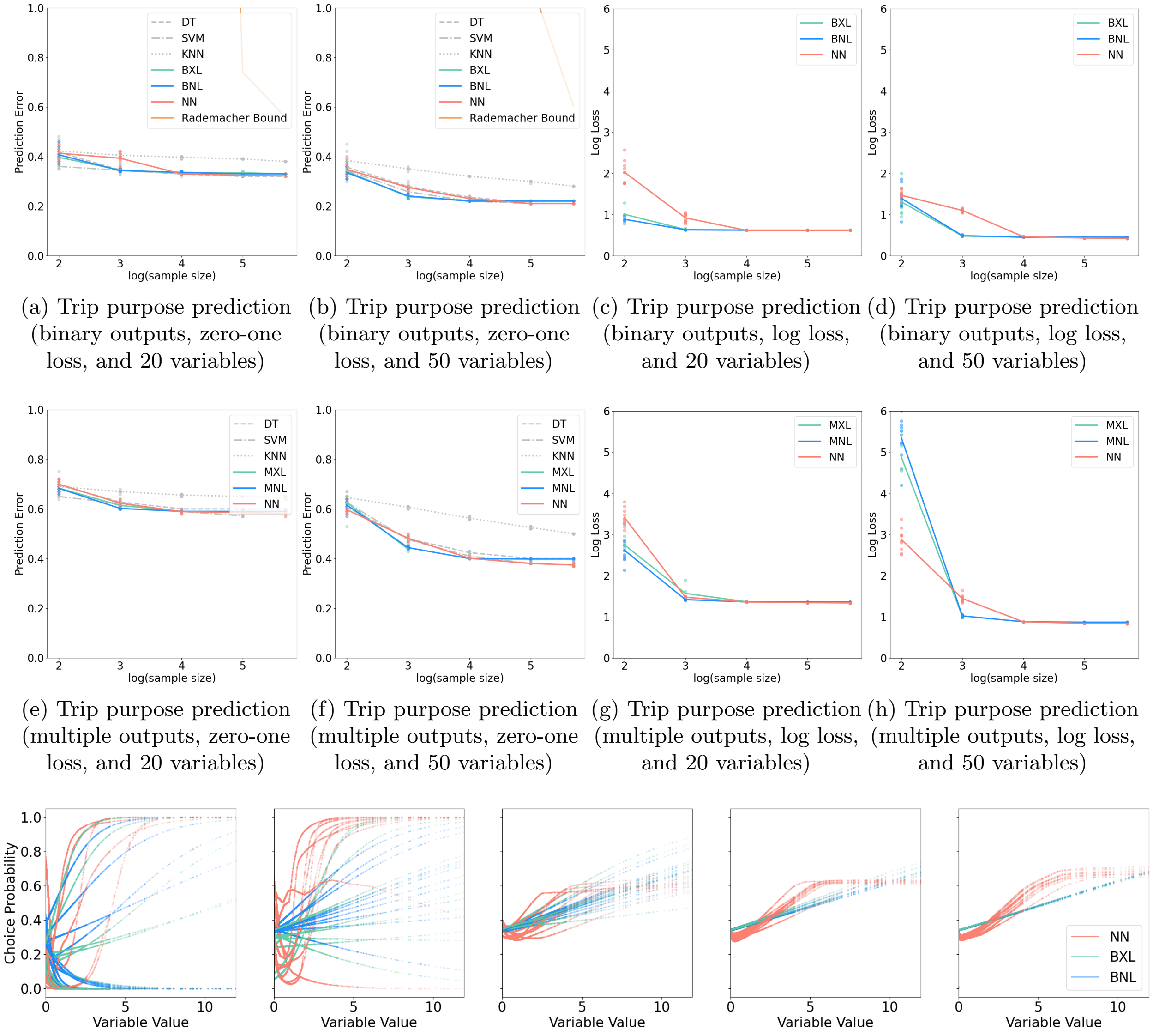
As shown in Figure 4, the results are very similar to that in scenario 2, with only one critical difference that DNN also has the approximation error here. The approximation error of DNN is measured by the difference between the theoretical minimum and the values of the three minimum loss values that DNNs converge to: the red curves no longer converge to theoretical minimum due to the existence of approximation errors, as shown in Figures 4a-4f. It is also an important message that DNN, although frequently referred to as a universal approximator, still suffers from the threat such as omitting variables. Without the completeness of information, it is unlikely for DNN to approximate the underlying $s^*(x)$ well. However, Figure 4g suggests that DNN could still well capture the choice probability function with respect to observed variables, even with omitted variables. The red curves (DNN) could approximate the true bell-shaped choice probability functions when the sample size reaches 10^4 or 10^5.

Overall, this scenario demonstrates that DNN cannot solve all the problems. The "universal approximator" statement only applies to the functional forms of observed information, therefore only holds when all the information is observed in the model. However, even with omitted information, DNN still performs better than BNL in terms of both prediction and interpretation, owing to its power of stretching the observed information for the unobserved ones and the power of automatically learning utility specification.

3.3. Experiment with NHTS Dataset

The NHTS dataset is chosen owing to its broad geographical coverage (full U.S.), the large sample size ($780,000$ trips), and the large number of input variables, enabling us to observe the variation of prediction accuracy with varying sample size and input variables. Based on the study from Hillel (2020), the NHTS data is resampled at the households' level and split into training and testing sets with a ratio of 9:1 (Hillel, 2020). To form a parallel discussion with our simulations, the NHTS experiment varies sample size (from 100 to $500,000$) and the number of input variables (20 and 50). The input variables are selected to be the most important ones that determine trip purposes. The summary statistics of the NHTS dataset is attached in Appendix III.A. The DNN models use a simple feedforward architecture, and BNL, BXL, MNL and MXL use the linear utility specifications. More specification details can be found in Appendix III.B. Three machine learning classifiers - decision tree (DT), support vector machine (SVM), and K nearest neighbors (KNN) - are also included for comparison. Among these models, only DCMs and DNN can output log loss, while the other three classifiers can only be evaluated by zero-one loss. The results are visualized in Figure 5, with the format similar to previous ones but two differences: interpretation quality can no longer be examined since $s^*(x)$ is no longer known and approximation error is no longer available because the theoretical minimum value is unknown either.

Interestingly, Figures 5a through 5d show a pattern that mixes scenarios 1 and 2: BNL outperforms DNN when sample size is around 10^3, while DNN starts to outperform BNL when sample size is larger than 10^4. The convergence of BNL is very quick, and it stops at around 10^3 sample size, while the convergence of DNN is still unclear given that the red curves still have decreasing trend when sample size reaches $500,000$. Figures 5e through 5h compare DNN and MNL, showing a pattern highly similar to the previous BNL case: DNN starts to outperform MNL when sample size reaches at least $O(10^4)$. The comparison of DNN, BNL, and MNL again demonstrates that only in cases with very large sample sizes can DNN's improved predictive power result in better performance. These results also suggest that handcrafted utility specification has captured certain



(a) Trip purpose prediction (binary outputs, zero-one loss, and 20 variables)

(b) Trip purpose prediction (binary outputs, zero-one loss, and 50 variables)

(c) Trip purpose prediction (binary outputs, log loss, and 20 variables)

(d) Trip purpose prediction (binary outputs, log loss, and 50 variables)

(e) Trip purpose prediction (multiple outputs, zero-one loss, and 20 variables)

(f) Trip purpose prediction (multiple outputs, zero-one loss, and 50 variables)

(g) Trip purpose prediction (multiple outputs, log loss, and 20 variables)

(h) Trip purpose prediction (multiple outputs, log loss, and 50 variables)

(i) Choice probabilility change w.r.t. trip distance (from left to right: sample size 100, 1000, 10000, 100000, 500000)

Fig. 5. NHTS Results. Upper row: comparison of DNN, BNL, and BXL in predicting trip purposes with binary outputs; middle row: comparison of DNN, MNL, and MXL in predicting trip purposes with multiple outputs; lower row: visualizing how choice probabilities change with inputs.

information, given its better performance when sample size is relatively small, although it does not capture all possible information in the dataset, given its worse performance when sample size is large. Obviously, the approximation error of DNN is smaller than BNL, but the estimation error of DNN does not appear large either.[3]

Figure 5i visualizes how probability of choosing home-based trips varies with trip distance. The results are quite similar to our findings in scenario 2 and 3 in that DNN starts to converge when

[3] Other models are not our main focus, but still provide valuable benchmark information. All the models, including DT, KNN, and SVM, demonstrate the pattern of convergence with larger sample size, although the rate of convergence varies. BXL and MXL models perform highly similar to the BNL and MNL models.

sample size reaches 10^4 and its pattern becomes quite stable when sample size equals to 10^5 or 10^6. The difference between DNN and BNL implies again that the approximation error exists in BNL. The probability functions of DNN and BNL are similar and intuitive in that both are monotonically increasing, while DNN seems to capture more subtlety than BNL: BNL suggests a nearly linear relationship, while DNN describes a relationship with roughly decreasing sensitivity to trip distance changes. This decreasing sensitivity is intuitive since people are less likely to change their travel behavior as trip distance is already large enough.

4. Conclusion and Discussions

This study discusses when and why DNN can be applied to choice analysis, with focuses on answering the non-overfitting and interpretability challenges faced by DNN. The statistical learning theoretical framework is presented to describe the tradeoff between estimation and approximation errors, and the balance between prediction and interpretation qualities. The theory is further demonstrated by using three simulated scenarios and the NHTS dataset, yielding these major findings.

First, the non-overfitting issue can be at least partially addressed by recent progresses in statistical learning theory and demonstrated in our experiments. The estimation error of both prediction and interpretation qualities can be bounded by Rademacher complexity of DNN. It is still challenging to provide a clear-cut rule about the correct sample size, since the theoretical development is still on-going and the theory suggests a subtle dynamics between sample size, input dimensions and scale, DNN depth, and norms of each layer in DNN. However, the bottom-line is that researchers do not need to count the number of parameters to bound estimation error of DNN, since the VC dimension based upper bound is too loose for DNN applications. Although sample size requirement is not as large as expected from classical statistical theory, a relatively large sample is still critical for generalizable results from DNN. Our experiments suggests that sample size needs to reach at least 10^4 for DNN to outperform BNL for typical travel behavior analysis. This requirement of sample size is slightly larger than the common size of questionnaire surveys, but still attainable in practice. In fact, several studies that found DNNs outperforming MNL have sample sizes with a similar magnitude to 10^4 (Xie, Lu, and Parkany, 2003; Omrani, 2015). However, even when sample size is less than 10^4, it does not imply that DNN cannot work. In this case, careful regularization methods can and should be used to improve model performance (S. Wang, Mo, and Zhao, 2020), although we do not focus on regularization much in this study.

Second, interpretability can be operationalized by using the function approximation loss of choice probability functions, owing to the fact that utility comparison and specification naturally exist in DNN and that an accurate estimator $\hat{s}(x)$ of choice probability function enables researchers to extract all the necessary economic information commonly obtained from traditional choice modeling. Our model interpretation is discussed in a way quite different from traditional methods for at least three reasons. (1) The process can be named as prediction-driven interpretation,[4] implying that researchers extract information from DNN after model training even though DNN is designed to maximize prediction accuracy in the first place. This prediction-driven interpretation is intuitive since "some structure must have been found in DNN, when predictive quality is consistently high" (Mullainathan and Spiess, 2017). (2) Our interpretation is based on function estimation rather than parameter estimation. It is nearly impossible to evaluate each individual parameter in DNN, so function estimation that focuses on the whole space of the transformed feature in DNN is a more viable way for interpretation. (3) This prediction-driven interpretation approach could automatically learn the underlying utility specification, as shown in our Monte Carlo simulations and the NHTS application. This approach is effective since handcrafted utility specification can rarely capture the full information, and correspondingly, certain power of automatic learning utility specification should always be involved in choice analysis.

These findings can be generalized to most applications involving MNL models, in which more than two alternatives exist. Choice analysis can be categorized into three cases according to the number of alternatives: (1) binary alternatives, (2) multiple but few alternatives, and (3) many alternatives.[5] Although this study mainly targets BNL, the theoretical formulation can extend to the case of few alternatives, because the estimation errors of DNN are still bounded by the formula in Proposition 5 and 6. The case of few alternatives is the most common type of empirical research, such as modeling the travel mode choice among automobiles, public transit, bicycle, and walking. The NHTS experiments also suggest that the case of few alternatives is empirically similar to that of binary alternatives. However for the third case, such as modeling the choice among thousands of residential locations, the answer is unclear, since the norm of the last layer in the DNN with thousands of outputs might not be bounded well. Future studies should focus on designing specific DNN architectures to address this issue; one potential solution is to design DNNs resembling the nested logit model that can condense the large number of alternatives with a tree structure. Since most of the travel behavioral analysis falls into the first and second cases, our work can serve a broad range of research for the research community.

This research suggests that DNNs are an appropriate modeling tool when sample size is large, input dimension is high, or the true DGP is complex; DCMs may be more appropriate when all of these three are false. Specifically, our simulation results favor the baseline DNNs more when sample size exceeds 10^4 and the true DGP deviates from the linear-in-parameter structure. When sample size is around or smaller than $O(10^3)$ observations, the number of input variables is around the order of $O(10)$, and the DGP is not very complex, the classical DCMs or their close variants appear to outperform DNN with the benefits of parameter-based interpretation and potentially higher prediction accuracy. However, these guidelines are somewhat specific to contexts, models, and data. While a naive DNN architecture might perform worse than a DCM in the case of small

[4] It can also be named as post-hoc interpretation, implying that researchers extract information from prediction-driven models after model training. It is debatable whether this approach is the best, since many other alternative approaches exist (Doshi-Velez and Kim, 2017; Ribeiro, Singh, and Guestrin, 2016; Lipton, 2016)

[5] Here, many alternatives indicate that the number of alternatives is of a magnitude similar to the sample size.

sample, DNNs with effective regularization and architecture design can be better than DCMs. In addition, when researchers use unstructured data such as images and natural language, DNNs so far appear to be the only viable modeling approach.

We believe these insights contribute to understanding when and why DNN can be used for choice analysis, and they are of both theoretical and practical importance. The theoretical framework can serve as a new foundation for future investigation in choice analysis, since it extends the classical asymptotic data regime to the non-asymptotic data regime, or equivalently, from low-dimensional statistical to high-dimensional statistical learning theory. This extension is important since high-dimensional data and complex models are becoming increasingly common in practice. Meanwhile, researchers can generate economic information from DNN-based choice models using our findings on interpretation quality, allowing these models to serve behavioral and policy analysis purposes.

However, many important questions still remain, since each one of the four quadrants can be explored to a greater extent than our overview in this paper. In the case of approximation error, the complexity of the underlying DGP influences the relative effectiveness of DNN and DCM. Further studies simulating true decision rules will be critical to understand how powerful DNNs are in approximating individual decision-making mechanisms. As for estimation error, while our study has demonstrated a state-of-the-art upper bound, the Rademacher complexity is still a loose bound relative to the empirical results. This result marks the gap between theoretical and empirical results in the deep learning community. To solve this problem, revolutionary theoretical work or novel regularization methods will need to be found to mitigate the estimation errors of DNNs in practice. To more thoroughly understand prediction quality, future studies should empirically investigate DNN and DCM under more contexts with different geographical locations, sample sizes, and targeting tasks, so that more conclusive results on the relative effectiveness of DNN and DCM can be obtained. As to the interpretation quality, while we present the function approximation loss that measures the interpretability of DNN models in a specific aspect, there remain many other aspects of interpretability that may remain issues for DNN applications to choice modeling research moving forward (Lipton, 2016; Bertsimas et al., 2019). While the broad scope of this research touches on each of these areas, more thorough exploration is left to future research in the rich intersection of machine learning models and individual decision-making theories.

Contributions of the Authors

S.W. and J.Z. conceived of the presented idea; S.W. developed the theory and reviewed previous studies; S.W. and Q.W. designed and conducted the experiments; S.W. and N.B. drafted the manuscripts; S.W. derived the analytical proofs. J.Z. supervised this work. All authors discussed the results and contributed to the final manuscript.

Acknowledgement

We thank Singapore-MIT Alliance for Research and Technology (SMART) for partially funding this research. We thank Sun Rui for his help in verifying the math proofs.

References

Train, Kenneth (1980). "A structured logit model of auto ownership and mode choice". In: *The Review of Economic Studies* 47.2, pp. 357–370.

Ben-Akiva, Moshe E and Steven R Lerman (1985). *Discrete choice analysis: theory and application to travel demand*. Vol. 9. MIT press.

Train, Kenneth (2009). *Discrete choice methods with simulation*. Cambridge university press.

Karlaftis, Matthew G and Eleni I Vlahogianni (2011). "Statistical methods versus neural networks in transportation research: Differences, similarities and some insights". In: *Transportation Research Part C: Emerging Technologies* 19.3, pp. 387–399.

Paredes, Miguel et al. (2017). "Machine learning or discrete choice models for car ownership demand estimation and prediction?" In: *Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2017 5th IEEE International Conference on*. IEEE, pp. 780–785.

Hagenauer, Julian and Marco Helbich (2017). "A comparative study of machine learning classifiers for modeling travel mode choice". In: *Expert Systems with Applications* 78, pp. 273–282.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989). "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5, pp. 359–366.

Hornik, Kurt (1991). "Approximation capabilities of multilayer feedforward networks". In: *Neural networks* 4.2, pp. 251–257.

Cybenko, George (1989). "Approximation by superpositions of a sigmoidal function". In: *Mathematics of control, signals and systems* 2.4, pp. 303–314.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*, pp. 1097–1105.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *Nature* 521.7553, pp. 436–444.

Fernández-Delgado, Manuel et al. (2014). "Do we need hundreds of classifiers to solve real world classification problems". In: *Journal of Machine Learning Research* 15.1, pp. 3133–3181.

Goodfellow, Ian et al. (2016). *Deep learning*. Vol. 1. MIT press Cambridge.

Glaeser, Edward L et al. (2018). "Big data and big cities: The promises and limitations of improved measures of urban life". In: *Economic Inquiry* 56.1, pp. 114–137.

Vapnik, Vladimir (1999). "An overview of statistical learning theory". In: *IEEE transactions on neural networks* 10.5, pp. 988–999.

— (2013). *The nature of statistical learning theory*. Springer science and business media.

Wainwright, Martin J (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press.

Cantarella, Giulio Erberto and Stefano de Luca (2005). "Multilayer feedforward networks for transportation mode choice analysis: An analysis and a comparison with random utility models". In: *Transportation Research Part C: Emerging Technologies* 13.2, pp. 121–155.

Dong, Chunjiao et al. (2018). “An innovative approach for traffic crash estimation and prediction on accommodating unobserved heterogeneities”. In: *Transportation research part B: methodological* 118, pp. 407–428.

Mozolin, Mikhail, J-C Thill, and E Lynn Usery (2000). “Trip distribution forecasting with multilayer perceptron neural networks: A critical evaluation”. In: *Transportation Research Part B: Methodological* 34.1, pp. 53–73.

Polson, Nicholas G and Vadim O Sokolov (2017). “Deep learning for short-term traffic flow prediction”. In: *Transportation Research Part C: Emerging Technologies* 79, pp. 1–17.

Wu, Yuankai et al. (2018). “A hybrid deep learning based traffic flow prediction method and its understanding”. In: *Transportation Research Part C: Emerging Technologies* 90, pp. 166–180.

Lipton, Zachary C (2016). “The mythos of model interpretability”. In: *arXiv preprint arXiv:1606.03490*.

Doshi-Velez, Finale and Been Kim (2017). “Towards a rigorous science of interpretable machine learning”. In:

Kotsiantis, Sotiris B, I Zaharakis, and P Pintelas (2007). “Supervised machine learning: A review of classification techniques”. In: *Emerging artificial intelligence applications in computer engineering* 160, pp. 3–24.

Zhou, Bolei et al. (2014). “Object detectors emerge in deep scene cnns”. In: *arXiv preprint arXiv:1412.6856*.

Hensher, David A and Tu T Ton (2000). “A comparison of the predictive potential of artificial neural networks and nested logit models for commuter mode choice”. In: *Transportation Research Part E: Logistics and Transportation Review* 36.3, pp. 155–172.

Xie, Chi, Jinyang Lu, and Emily Parkany (2003). “Work travel mode choice modeling with data mining: decision trees and neural networks”. In: *Transportation Research Record: Journal of the Transportation Research Board* 1854, pp. 50–61.

Celikoglu, Hilmi Berk (2006). “Application of radial basis function and generalized regression neural networks in non-linear utility function specification for travel mode choice modelling”. In: *Mathematical and Computer Modelling* 44.7, pp. 640–658.

Omrani, Hichem (2015). “Predicting travel mode of individuals by machine learning”. In: *Transportation Research Procedia* 10, pp. 840–849.

Rao, PV Subba et al. (1998). “Another insight into artificial neural networks through behavioural analysis of access mode choice”. In: *Computers, environment and urban systems* 22.5, pp. 485–496.

Bentz, Yves and Dwight Merunka (2000). “Neural networks and the multinomial logit for brand choice modelling: a hybrid approach”. In: *Journal of Forecasting* 19.3, pp. 177–200.

Wang, Shenhao, Qingyi Wang, and Jinhua Zhao (2020a). “Deep neural networks for choice analysis: Extracting complete economic information for interpretation”. In: *Transportation Research Part C: Emerging Technologies* 118, p. 102701. issn: 0968-090X.

Wang, Shenhao, Baichuan Mo, and Jinhua Zhao (2020). “Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions”. In: *Transportation Research Part C: Emerging Technologies* 112, pp. 234–251. issn: 0968-090X.

Wang, Shenhao, Qingyi Wang, and Jinhua Zhao (2020b). "Multitask learning deep neural networks to combine revealed and stated preference data". In: *Journal of Choice Modelling*, p. 100236. ISSN: 1755-5345.

Cheng, Long et al. (2019). "Applying a random forest method approach to model travel mode choice behavior". In: *Travel behaviour and society* 14, pp. 1–10.

Tang, Liang, Chenfeng Xiong, and Lei Zhang (2015). "Decision tree method for modeling travel mode switching in a dynamic behavioral process". In: *Transportation Planning and Technology* 38.8, pp. 833–850.

Allahviranloo, Mahdieh and Will Recker (2013). "Daily activity pattern recognition by using support vector machines with multiple classes". In: *Transportation Research Part B: Methodological* 58, pp. 16–43.

Harrell, Frank E (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis.* Springer. ISBN: 3319194259.

Hillel, Tim, Michel Bierlaire, et al. (2020). "A systematic review of machine learning classification methodologies for modelling passenger mode choice". In: *Journal of Choice Modelling*, p. 100221. ISSN: 1755-5345.

Gneiting, Tilmann and Adrian E Raftery (2007). "Strictly proper scoring rules, prediction, and estimation". In: *Journal of the American statistical Association* 102.477, pp. 359–378. ISSN: 0162-1459.

He, Kaiming et al. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hillel, Tim, Michel Bierlaire, et al. (2019). "Weak teachers: Assisted specification of discrete choice models using ensemble learning". In: *8th Symposium of the European Association for Research in Transportation, Budapest.*

Bishop, Christopher M (2006). *Pattern recognition and machine learning.* springer.

Szegedy, Christian et al. (2014). "Intriguing properties of neural networks". In: *arXiv preprint arXiv:1312.6199.*

Montavon, Gregoire, Wojciech Samek, and Klaus-Robert Muller (2018). "Methods for interpreting and understanding deep neural networks". In: *Digital Signal Processing* 73, pp. 1–15.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean (2015). "Distilling the knowledge in a neural network". In: *arXiv preprint arXiv:1503.02531.*

Baehrens, David et al. (2010). "How to explain individual classification decisions". In: *Journal of Machine Learning Research* 11.Jun, pp. 1803–1831.

Ross, Andrew Slavin and Finale Doshi-Velez (2018). "Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients". In: *Thirty-second AAAI conference on artificial intelligence.*

Mullainathan, Sendhil and Jann Spiess (2017). "Machine learning: an applied econometric approach". In: *Journal of Economic Perspectives* 31.2, pp. 87–106.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). “Why should i trust you?: Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 1135–1144.

Bertsimas, Bimitris et al. (2019). “The Price of Interpretability”. In: *Arxiv preprint*.

Cohen, Jonathan D et al. (2016). *Measuring time preferences*. Tech. rep. National Bureau of Economic Research.

Rolnick, David and Max Tegmark (2017). “The power of deeper networks for expressing natural functions”. In: *arXiv preprint arXiv:1705.05502*.

Poggio, Tomaso, Hrushikesh Mhaskar, et al. (2017). “Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review”. In: *International Journal of Automation and Computing* 14.5, pp. 503–519.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent (2013). “Representation learning: A review and new perspectives”. In: *IEEE transactions on pattern analysis and machine intelligence* 35.8, pp. 1798–1828.

Liao, Qianli and Tomaso Poggio (2018). *When Is Handcrafting Not a Curse?* Tech. rep.

Golowich, Noah, Alexander Rakhlin, and Ohad Shamir (2017). “Size-independent sample complexity of neural networks”. In: *arXiv preprint arXiv:1712.06541*.

Bartlett, Peter L, Nick Harvey, et al. (2017). “Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks”. In: *arXiv preprint arXiv:1703.02930*.

Bousquet, Olivier, Stéphane Boucheron, and Gábor Lugosi (2004). “Introduction to statistical learning theory”. In: *Advanced lectures on machine learning*. Springer, pp. 169–207.

Von Luxburg, Ulrike and Bernhard Schölkopf (2011). “Statistical learning theory: Models, concepts, and results”. In: *Handbook of the History of Logic*. Vol. 10. Elsevier, pp. 651–706.

Anthony, Martin and Peter L Bartlett (2009). *Neural network learning: Theoretical foundations*. cambridge university press.

Sontag, Eduardo D (1998). “VC dimension of neural networks”. In: *NATO ASI Series F Computer and Systems Sciences* 168, pp. 69–96.

Neyshabur, Behnam, Ryota Tomioka, and Nathan Srebro (2015). “Norm-based capacity control in neural networks”. In: *Conference on Learning Theory*, pp. 1376–1401.

Bartlett, Peter L and Shahar Mendelson (2002). “Rademacher and Gaussian complexities: Risk bounds and structural results”. In: *Journal of Machine Learning Research* 3.Nov, pp. 463–482.

Bartlett, Peter L, Michael I Jordan, and Jon D McAuliffe (2006). “Convexity, classification, and risk bounds”. In: *Journal of the American Statistical Association* 101.473, pp. 138–156.

Breiman, Leo (2001). “Statistical modeling: The two cultures (with comments and a rejoinder by the author)”. In: *Statistical science* 16.3, pp. 199–231.

He, Kaiming et al. (2015). “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification”. In: *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034.

Kingma, Diederik P and Jimmy Ba (2014). “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980*.

De Dios Ortuzar, Juan and Luis G Willumsen (2011). *Modelling transport*. John Wiley and Sons.

Zegras, Christopher (2010). “The built environment and motor vehicle ownership and use: Evidence from Santiago de Chile”. In: *Urban Studies* 47.8, pp. 1793–1817.

Cervero, Robert and Kara Kockelman (1997). “Travel demand and the 3Ds: density, diversity, and design”. In: *Transportation Research Part D: Transport and Environment* 2.3, pp. 199–219.

Hillel, Tim (2020). “New perspectives on the performance of machine learning classifiers for mode choice prediction”. In:

McFadden, Daniel (1974). “Conditional logit analysis of qualitative choice behavior”. In:

Ledoux, Michel and Michel Talagrand (2013). *Probability in Banach Spaces: isoperimetry and processes*. Springer Science Business Media.

Poggio, Tomaso, Qianli Liao, et al. (2018). “Theory IIIb: Generalization in deep networks”. In: *arXiv preprint arXiv:1806.11379*.

Poggio, Tomaso, Kenji Kawaguchi, et al. (2018). *Theory of deep learning iii: the non-overfitting puzzle*. Tech. rep. Technical report, Technical report, CBMM memo 073.

Soudry, Daniel et al. (2018). “The implicit bias of gradient descent on separable data”. In: *The Journal of Machine Learning Research* 19.1, pp. 2822–2878.

Appendix I: BNL as One Special Case of DNN

Suppose individuals choose between two alternatives 0 and 1, which have the utility specifications: $U_{i0} = V_{i0} + \epsilon_{i0}$; $U_{i1} = V_{i1} + \epsilon_{i1}$, in which V is the deterministic utility and ϵ is the random utility term. Choice modeling assumes that individuals seek to maximize utility, so the choice probability functions are given by Equation 24 when ϵ follows extreme value distribution $EV(0, 1)$.

$$\begin{aligned} P(y_i = 1|x_i, w) &= \frac{1}{1 + e^{-(V_{i1} - V_{i0})}} \\ P(y_i = 0|x_i, w) &= \frac{1}{1 + e^{+(V_{i1} - V_{i0})}} \end{aligned} \tag{24}$$

Assuming that attributes relevant to alternatives 0 and 1 are x_{i0} and x_{i1}, the deterministic utility function with linear specification is

$$\begin{aligned} V_{i0}(x_{i0}) &= \langle w_0, x_{i0} \rangle \\ V_{i1}(x_{i1}) &= \langle w_1, x_{i1} \rangle \end{aligned} \tag{25}$$

This specification could be more involved by using some transformation $\phi(x)$ (quadratic or log) based on prior knowledge. Hence often the real utility specification could be denoted as

$$\begin{aligned} V_{i0}(x_{i0}) &= \langle w_0, \phi(x_{i0}) \rangle \\ V_{i1}(x_{i1}) &= \langle w_1, \phi(x_{i1}) \rangle \end{aligned} \tag{26}$$

This specification is quite close to that in DNN, which is:

$$\begin{aligned} V_{i0}(x_{i0}) &= \langle w_0, (g_{m-1} ... \circ g_2 \circ g_1)(x_{i0}) \rangle \\ V_{i1}(x_{i1}) &= \langle w_1, (g_{m-1} ... \circ g_2 \circ g_1)(x_{i1}) \rangle \end{aligned} \tag{27}$$

in which $g_j(x) = ReLU(\langle W_j, x \rangle)$. Comparing Equations 25, 26, and 27, it is not hard to see that DNN specification is more general than previous two. A more formal way to demonstrate this point is to use the results from McFadden (1974) McFadden, 1974, which proved that Softmax activation function implicitly implies a random utility maximization with a random utility term that follows Gumbel distribution. By changing the notation of Equation 27,

$$\Phi_1(x_i, w) = V_{i1}(x_{i1}) - V_{i0}(x_{i0}) = (g_m \circ ... \circ g_2 \circ g_1)(x_i) \tag{28}$$

in which g_m is $w_1 - w_0$ and x_i includes all input information. Then Equation 28 implies the choice probability of in DNN is:

$$\sigma(\Phi_1(x_i, w)) = \frac{1}{1 + e^{-\Phi_1(x_i, w)}} \tag{29}$$

which is the same as Equation 1.

Appendix II.A: Proof of Proposition 2

Estimation error can be decomposed:

$$
\begin{aligned}
\mathbb{E}_S[L(\hat{f}) - L(f_F^*)] &= \mathbb{E}_S[L(\hat{f}) - \hat{L}(\hat{f}) + \hat{L}(\hat{f}) - \hat{L}(f_F^*) + \hat{L}(f_F^*) - L(f_F^*)] && (30)\\
&\leq \mathbb{E}_S[L(\hat{f}) - \hat{L}(\hat{f})] && (31)\\
&\leq \mathbb{E}_S[\sup_{f\in F} |L(f) - \hat{L}(f)|] && (32)
\end{aligned}
$$

The first inequality holds since (1) $\hat{L}(\hat{f}) - \hat{L}(f_F^*) \leq 0$ due to the definition of $\hat{f}$ and (2) $\mathbb{E}_S[\hat{L}(f_F^*) - L(f_F^*)] = 0$ due to law of large numbers; the second inequality holds since $\hat{f}$ is only one function in F ($\mathcal{F}_0$ or $\mathcal{F}_1$ in this study). The right hand side of Equation 32 above can be further upper bounded by using a technique called symmetrization. Formally, suppose another set of $\{x_i'\}_1^N$ is also generated, following the same distribution as $\{x_i\}_1^N$. Then

$$
\begin{aligned}
\mathbb{E}_S\Big[\sup_{f\in F}\Big|L(f) - \hat{L}(f)\Big|\Big] &= \mathbb{E}_S\Big[\sup_{f\in F}\Big|\mathbb{E}_{x,y}[l(y, f(x))] - \frac{1}{N}\sum_{i=1}^{N} l(y_i, f(x_i))\Big|\Big] && (33)\\
&= \mathbb{E}_S\Big[\sup_{f\in F}\Big|\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{x'} l(y, f(x_i')) - \frac{1}{N}\sum_{i=1}^{N} l(y_i, f(x_i))\Big|\Big] && (34)\\
&\leq \mathbb{E}_{S,S'}\Big[\sup_{f\in F}\Big|\frac{1}{N}\sum_{i=1}^{N} l(y, f(x_i')) - \frac{1}{N}\sum_{i=1}^{N} l(y_i, f(x_i))\Big|\Big] && (35)\\
&= \mathbb{E}_{S,S'}\Big[\sup_{f\in F}\Big|\frac{1}{N}\sum_{i=1}^{N} \epsilon_i(l(y, f(x_i')) - l(y_i, f(x_i))\Big|\Big] && (36)\\
&\leq \mathbb{E}_{S,S'}\Big[\sup_{f\in F}\Big|\frac{1}{N}\sum_{i=1}^{N} \epsilon_i l(y, f(x_i')\Big| + \Big|\frac{1}{N}\sum_{i=1}^{N} \epsilon_i l(y_i, f(x_i)\Big|\Big] && (37)\\
&\leq 2\mathbb{E}_S\hat{\mathcal{R}}_n(l \circ \mathcal{F}|_S) && (38)
\end{aligned}
$$

The first line uses the definition of L and $\hat{L}$; the second line uses the symmetrization technique by which $\mathbb{E}_{x,y}$ is replaced by an average of another sample $\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{x'} l(y, f(x_i'))$; the third line uses $\mathbb{E}\ \sup \geq \sup\ \mathbb{E}$ and uses S' to denote the new sample $\{x'\}_1^N$; the fourth line adds the Rademacher random variable ϵ_i due to the symmetry of S and S'; the fifth line uses the fact $\sup|A + B| \leq \sup|A| + \sup|B|$; and the last line is the definition of Rademacher complexity.

Appendix II.B: Proof of Proposition 3

Definition 9. *Ramp loss function is defined as*

$$\phi(s) = \begin{cases} 1 & s \leq 0 \\ 1 - s/\gamma & 0 < s < \gamma \\ 0 & s \geq \gamma \end{cases} \tag{39}$$

Associated error function is

$$L_\phi = \mathbb{E}[\phi(s)] \tag{40}$$

Definition 10. *γ-margin loss function is defined as*

$$\mathbb{1}\{y\Phi(x) \leq \gamma\} \tag{41}$$

Associated error function is

$$L_\gamma = \mathbb{E}[\mathbb{1}\{y\Phi(x) \leq \gamma\}] \tag{42}$$

L_ϕ is an example of *surrogate loss* functions for $L_{0/1}$. It is a surrogate loss since L_ϕ is designed to (1) upper bound $L_{0/1}$ and (2) be L-Lipschitz so that the contraction inequality can be applied. The Lipschitz constant of L_ϕ is $1/\gamma$. By design, three error functions are related:

$$L_{0/1} \leq L_\phi \leq L_\gamma \tag{43}$$

Therefore, the estimation error measured by prediction error $L_{0/1}$ has an upper bound

$$\mathbb{E}_S[L_{0/1}(\hat{f}) - \hat{L}_\gamma(\hat{f})] \leq \mathbb{E}_S[L_\phi(\hat{f}) - \hat{L}_\phi(\hat{f})] \tag{44}$$

An upper bound for the right hand side of Equation 44 can be found by using Proposition 2 and contraction inequality.

$$\begin{aligned}
\mathbb{E}_S[L_\phi(\hat{f}) - \hat{L}_\phi(\hat{f})] &\leq \mathbb{E}_S \sup_{f \in \mathcal{F}_1} |L_\phi(f) - \hat{L}_\phi(f)| & (45)\\
&= \mathbb{E}_S \sup_{f \in \mathcal{F}_1} |\,\mathbb{E}[\phi(f)] - \frac{1}{N}\sum_{i=1}^{N} \phi(f(x_i))| & (46)\\
&= \mathbb{E}_{S,\epsilon} \sup_{f \in \mathcal{F}_1} \frac{2}{N}\sum_{i=1}^{N} |\epsilon_i \phi(f(x_i))| & (47)\\
&\leq \frac{2}{\gamma} \times \mathbb{E}_{S,\epsilon} \sup_{f \in \mathcal{F}_1} \frac{1}{N}\sum_{i=1}^{N} |\epsilon_i f(x_i)| & (48)\\
&= \frac{2}{\gamma}\mathbb{E}_{S,\epsilon}\hat{\mathcal{R}}_n(\mathcal{F}_1|_S) & (49)
\end{aligned}$$

The first inequality holds due to the sup operator; the second line uses the definitions of ramp cost functions; the third line used Proposition 2; the fourth line used contraction inequality Ledoux and Talagrand, 2013; and the last line used the definition of empirical Rademacher complexity. Using Equation 44, it implies

$$\mathbb{E}_S[L_{0/1}(\hat{f}) - \hat{L}_\gamma(\hat{f})] \leq \frac{2}{\gamma}\mathbb{E}_{S,\epsilon}\hat{\mathcal{R}}_n(\mathcal{F}_1|_S) \tag{50}$$

Therefore, the upper bound of $L_{0/1}(\hat{f})$ can be given by empirical γ-margin loss plus Rademacher complexity. $\hat{L}_\gamma(\hat{f})$ can be empirically computed, so a valid upper bound exists for $L_{0/1}(\hat{f})$. However, the unresolved question is whether DNN automatically finds a maximum margin similar to SVM, so that the $L_{0/1}(\hat{f})$ is bounded well. It is still an on-going research field Poggio, Liao, et al., 2018; Poggio, Kawaguchi, et al., 2018; Soudry et al., 2018. □

Appendix II.C: Proof of Proposition 4

Definition 11. *Mean squared error (MSE) is defined as*

$$L_{mse}(s) = \mathbb{E}_{x,y}[(y - s(x))^2] \tag{51}$$

The corresponding empirical mean squared error is defined as

$$\hat{L}_{mse}(s) = \frac{1}{N}\sum_{i=1}^{N}(y_i - s(x_i))^2 \tag{52}$$

Lemma 4.1. *Estimation error for interpretation equals to that of MSE.*

$$\mathbb{E}_S[L_{mse}(\hat{s}) - L_{mse}(s_F^*))] = \mathbb{E}_S[L_e(\hat{s}) - L_e(s_F^*))] \tag{53}$$

Proof of Lemma 4.1. Since y is sampled as a Bernoulli random variable with probability $s^*(x)$, $E[y|x] = s^*(x)$.

$$\begin{aligned}
\mathbb{E}_{S,x,y}[(\hat{s}(x) - y)^2] &= \mathbb{E}_{S,x,y}((\hat{s}(x) - s^*(x) + s^*(x) - y)^2) && (54)\\
&= \mathbb{E}_{S,x,y}[((\hat{s}(x) - s^*(x))^2 + 2(\hat{s}(x) - s^*(x))(s^*(x) - y) + (s^*(x) - y)^2)] && (55)\\
&= \mathbb{E}_{S,x,y}[(\hat{s}(x) - s^*(x))^2] + \mathbb{E}_{x,y}[(s^*(x) - y)^2)] + 2\mathbb{E}_{S,x,y}[(\hat{s}(x) - s^*(x))(s^*(x) - y)] && (56)\\
&= \mathbb{E}_{S,x,y}[(\hat{s}(x) - s^*(x))^2] + \mathbb{E}_{x,y}[(s^*(x) - y)^2)] + 2\mathbb{E}_x\left[\mathbb{E}_{S,y}[(\hat{s}(x) - s^*(x))(s^*(x) - y)|x]\right] && (57)\\
&= \mathbb{E}_{S,x,y}[(\hat{s}(x) - s^*(x))^2] + \mathbb{E}_{x,y}[(s^*(x) - y)^2)] + 2\mathbb{E}_x\left[\mathbb{E}_S[(\hat{s}(x) - s^*(x))|x]\mathbb{E}_y[(s^*(x) - y)|x]\right] && (58)\\
&= \mathbb{E}_{S,x,y}[(\hat{s}(x) - s^*(x))^2] + \mathbb{E}_{x,y}[(s^*(x) - y)^2)] && (59)
\end{aligned}$$

The fourth equality uses Law of Iterated Expectation; the fifth uses the conditional independence $S \perp y|x$; the lase one uses $E[y|x] = s^*(x)$. With very similar process, we could show

$$\begin{aligned}
\mathbb{E}_{x,y}[(y - s_F^*(x))^2] &= \mathbb{E}_{x,y}[(y - s^*(x) + s^*(x) - s_F^*(x))^2] && (60)\\
&= \mathbb{E}_{x,y}[(y - s^*(x))^2] + \mathbb{E}_{x,y}[(s^*(x) - s_F^*(x))^2] + 2\mathbb{E}_{x,y}[(y - s^*(x))(s^*(x) - s_F^*(x))] && (61)\\
&= \mathbb{E}_{x,y}[(y - s^*(x))^2] + \mathbb{E}_{x,y}[(s^*(x) - s_F^*(x))^2] + 2\mathbb{E}_x\left[(s^*(x) - s_F^*(x))\mathbb{E}_y[y - s^*(x)|x]\right] && (62)\\
&= \mathbb{E}_{x,y}[(y - s^*(x))^2] + \mathbb{E}_{x,y}[(s^*(x) - s_F^*(x))^2] && (63)
\end{aligned}$$

Combining the two equations above implies

$$\begin{aligned}
\mathbb{E}_{x,y}[(s^*(x) - y)^2)] &= \mathbb{E}_{S,x,y}[(\hat{s}(x) - y)^2] - \mathbb{E}_{S,x,y}[(\hat{s}(x) - s^*(x))^2] && (64)\\
&= \mathbb{E}_{x,y}[(y - s_F^*(x))^2] - \mathbb{E}_{x,y}[(s^*(x) - s_F^*(x))^2] && (65)
\end{aligned}$$

By changing the notation, it implies

$$\mathbb{E}_S[L_{mse}(\hat{s}) - L_{mse}(s_F^*))] = \mathbb{E}_S[L_e(\hat{s}) - L_e(s_F^*))] \tag{66}$$

Proof of Proposition 4. Lemma 4.1 shows that the estimation error on function estimation is the same as the one on MSE. Hence we will provide an upper bound on the MSE by using Proposition 2. Formally,

$$\mathbb{E}_S[L_{mse}(\hat{s}) - L_{mse}(s^*_F))] \leq 2\mathbb{E}_S[\hat{R}_n(l \circ \mathcal{F}|_S)] \tag{67}$$

$$\leq 4\mathbb{E}_S[\hat{R}_n(\mathcal{F}|_S)] \tag{68}$$

The first inequality uses Proposition 2; the second uses contraction inequality and the fact that squared loss here is bounded between $[0, 1]$ and that its Lipschitz constant is at most two. □

Appendix II.D: Proof of Proposition 5

The proof is an iterative process going through layer by layer. Suppose for layer j of DNN, the mapping is

$$F_j = \{f : x \to \sum_{t=1}^{d_{j-1}} w_t \sigma(f_t(x)); f_t \in F_{j-1}, ||w||_1 \leq M(j)\}$$

Then the Rademacher complexity of F_j can be represented by that of F_{j-1}.

$$N\hat{\mathcal{R}}_n(F_j|_S) = \mathop{\mathbb{E}}_{\epsilon} \sup_{f_j \in F_j} \Big| \sum_{i=1}^{N} \epsilon_i f(x_i) \Big| \tag{69}$$

$$= \mathop{\mathbb{E}}_{\epsilon} \sup_{f_j \in F_j} \Big| \sum_{i=1}^{N} \epsilon_i \sum_{t=1}^{d_{j-1}} w_t \sigma(f_t(x_i)) \Big| \tag{70}$$

$$= \mathbb{E}_\epsilon \sup_{\substack{||w||_1 \leq M(j) \\ f_t \in F_{j-1}}} \Big| \sum_{t=1}^{d_{j-1}} w_t \sum_{i=1}^{N} \epsilon_i \sigma(f_t(x_i)) \Big| \tag{71}$$

$$= 2\mathbb{E}_\epsilon \sup_{\substack{||w||_1 \leq M(j) \\ f_t \in F_{j-1}}} \sum_{t=1}^{d_{j-1}} w_t \sum_{i=1}^{N} \epsilon_i \sigma(f_t(x_i)) \tag{72}$$

$$= 2M(j)\mathbb{E}_\epsilon \sup_{f_t \in F_{j-1}} \max_t \Big| \sum_{i=1}^{N} \epsilon_i \sigma(f_t(x_i)) \Big| \tag{73}$$

$$= 2M(j)\mathbb{E}_\epsilon \sup_{f_t \in F_{j-1}} \Big| \sum_{i=1}^{N} \epsilon_i \sigma(f_t(x_i)) \Big| \tag{74}$$

$$\leq 2M(j)\mathbb{E}_\epsilon \sup_{f_t \in F_{j-1}} \Big| \sum_{i=1}^{N} \epsilon_i f_t(x_i) \Big| \tag{75}$$

$$\leq 2M(j)N\hat{\mathcal{R}}_n(F_{j-1}|_S) \tag{76}$$

which implies this iterative formula for DNN:

$$\hat{\mathcal{R}}_n(F_j|_S) = 2M(j)\hat{\mathcal{R}}_n(F_{j-1}|_S) \tag{77}$$

The remaining question is about the Rademacher complexity of layer 0, which is a linear transformation $F_0 = \{x \to \langle w, x\rangle : w \in B_1^d\}$ with normalized input X.

$$\hat{\mathcal{R}}_n(F_0|_S) \leq \sqrt{\frac{\log d_0}{N}} \tag{78}$$

Combining the equations above, Rademacher complexity of DNN can be proved as:

$$\hat{\mathcal{R}}_n(\mathcal{F}_1|_S) \lesssim \frac{\sqrt{\log d_0} \times \prod_{j=1}^{D} 2M(j)}{\sqrt{N}} \tag{79}$$

Note that here the Rademacher complexity has the 2^D factor. With more involved technique, a tighter upper bound could be proved as

$$\hat{\mathcal{R}}_n(\mathcal{F}_1|_S) \lesssim \frac{\sqrt{\log d_0} \times (\sqrt{2\log(D)} + 1)\prod_{j=1}^{D} M_F(j)}{\sqrt{N}} \tag{80}$$

This result can be found in in Golowich et al. (2017) Golowich, Rakhlin, and Shamir, 2017, with slight differences. The key steps of the proof we presented here can be found in Bartlett and Mendelson (2002) Bartlett and Mendelson, 2002. Other relevant work can be found in Anthony and Bartlett, 2009 and Neyshabur, Tomioka, and Srebro, 2015.

Appendix II.E. Proof of Proposition 6

Since VC dimension is only used as a benchmark, we will demonstrate a simple proof that upper bound the estimation error by $O(\sqrt{\frac{v\log(N+1)}{N}})$ for binary output. Using Lemma 4.14 from Wainwright (2019) Wainwright, 2019

$$\hat{\mathcal{R}}_n(l \circ \mathcal{F}_1|_S) \leq 4\sqrt{\frac{v\log(N+1)}{N}} \tag{81}$$

Note that $\log(N+1)$ is much smaller than v and N. This upper bound can be simplified to

$$O(\sqrt{\frac{v}{N}}) \tag{82}$$

which is similar to the traditional wisdom of examining the ratio between number of parameters and number of observations, since v is the same as parameter numbers in generalized linear models. For DNN, the tightest possible VC dimension can be found in Bartlett, Harvey, et al., 2017, which is

$v = O(TD\log(T))$ with T denoting the total number of coefficients and D the depth of DNN. This $O(\sqrt{\frac{v}{N}})$ can also be used for the $\hat{s}(x)$ case. But we won't discuss details here. Readers could refer to Vapnik, 1999; Vapnik, 2013; Von Luxburg and Schölkopf, 2011; Wainwright, 2019 for details.

Appendix III.A: Summary Statistics of NHTS Dataset

The summary statistics are summarized in Table 2.

Appendix III.B: Formulation of DNN, BNL, and MNL in NHTS Experiments

The DNN, BNL, and MNL models are specified as the simplest forms. Their forms are also consistent with the theoretical discussion. Specifically, the choice probability function of DNN is:

$$s(x_i, w) = \sigma(\Phi_1(x_i, w)) = \frac{1}{1 + e^{-\Phi_1(x_i,w)}} \tag{83}$$

with $\Phi_1(x_i, w) = (g_m \circ ...g_2 \circ g_1)(x_i)$, in which $g_j(x) = ReLU(\langle W_j, x\rangle)$. x_i represents the vector that combines all the alternative-specific variables x_{ik} and individual-specific variable z_i.

The utility functions of BNL and MNL are:

$$V_{ik} = \beta_{0,k} + \beta_{x,k}^T x_{ik} + \beta_{z,k}^T z_i, \quad as \quad k \neq ref \tag{84}$$

$$V_{ik} = \beta_{x,k}^T x_{ik}, \quad as \quad k = ref \tag{85}$$

in which V_{ik} is the deterministic utility value for alternative k; $\beta_{0,k}$ represents the alternative-specific constant for alternative k; $\beta_{x,k}$ represents the parameters for the alternative-specific variables x_{ik}; $\beta_{z,k}$ represents the parameters for the individual-specific variables z_i; ref represents the reference alternative. This formulation is the simplest specification that guarantees the parameter identification in choice modeling. This study uses the linear specification for two reasons. First for fairness, both BNL/MNL and DNNs use the linear inputs, so their comparison is not biased. Second for simplicity, while we use only linear specification, future studies can compare DNN to the BNL/MNL with feature transformations.

In our experiments with binary alternatives, the choice set of the travel mode choice models includes automobiles and non-automobile; that of the trip purpose models includes home-based and non-home-based trips. The travel mode choice models with multiple alternatives use six alternatives, including walk/bicycle, automobiles, SUV, trucks, public transit, and others. The trip purpose models use five alternatives, including home-based others, home-based shopping, home-based social and recreational trips, home-based working, and non-home based trips.

Description	mean	std	min	25%	50%	75%	max
NBIKETRP: count of bike trips	0.31	1.64	0.0	0.00	0.00	0.00	99.00
NWALKTRP: count of walk trips	5.99	8.70	0.0	1.00	4.00	7.00	200.00
RIDESHARE: count of rideshare app usage	0.34	1.81	0.0	0.00	0.00	0.00	99.00
PTUSED: count of public transit usage	0.91	4.32	0.0	0.00	0.00	0.00	240.00
TRPACCMP: count of people on trip	0.71	1.86	0.0	0.00	0.00	1.00	400.00
TRPHHACC: count of household members on trip	0.42	0.73	0.0	0.00	0.00	1.00	10.00
WRKCOUNT: number of workers in the household	1.24	0.97	0.0	0.00	1.00	2.00	7.00
GASPRICE: gas price in cents	240.02	22.90	201.3	223.00	235.60	259.40	295.90
DRVRCNT: number of drivers in the household	1.94	0.81	0.0	1.00	2.00	2.00	9.00
PHYACT: level of physical activity	2.20	0.58	1.0	2.00	2.00	3.00	3.00
HHSIZE: household size	2.50	1.27	1.0	2.00	2.00	3.00	13.00
TRVLCMIN: trip duration (minutes)	21.32	31.42	0.0	8.00	15.00	25.00	1140.00
TRPMILES: trip distance (miles)	11.41	68.85	0.0	1.41	3.62	9.18	9621.05
HTEEMPDN: category of workers per square mile in the census tract of the household	1384	1503	25	150	750	3000	5000
DELIVER: number of times purchased online for delivery in the past 30 days	2.90	4.51	0.0	0.00	1.00	4.00	99.00
CNTTDTR: count of person trips on travel day	5.73	3.02	1.0	4.00	5.00	7.00	50.00
R_AGE_IMP: age	52.45	17.36	6.0	39.00	55.00	66.00	92.00
HHVEHCNT: household vehicle count	2.23	1.20	0.0	1.00	2.00	3.00	12.00
HEALTH: level of health condtion	2.16	0.95	1.0	1.00	2.00	3.00	5.00
LPACT: count of light/moderate physical activity in the past week	2.52	3.11	0.0	0.00	2.00	4.00	25.00
MSASIZE_NEW: population of the Metropolitan Statistical Area of the household	1.62e+6	1.63e+6	0	1.25e+5	7.5e+5	4.0e+6	4.0e+6

Table 2: Summary statistics of the NHTS dataset