Deep neural networks for choice analysis: Extracting complete economic information for interpretation

Deep Neural Networks for Choice Analysis: Extracting Complete Economic Information for Interpretation

Shenhao Wang
Qingyi Wang
Jinhua Zhao

Massachusetts Institute of Technology
77 Mass Ave, Cambridge, Massachusetts, U.S.

January 2020

Abstract

While deep neural networks (DNNs) have been increasingly applied to choice analysis showing high predictive power, it is unclear to what extent researchers can interpret economic information from DNNs. This paper demonstrates that DNNs can provide economic information as **complete** as classical discrete choice models (DCMs). The economic information includes choice predictions, choice probabilities, market shares, substitution patterns of alternatives, social welfare, probability derivatives, elasticities, marginal rates of substitution, and heterogeneous values of time. Unlike DCMs, DNNs can automatically learn utility functions and reveal behavioral patterns that are not prespecified by domain experts, particularly when the sample size is large. However, the economic information obtained from DNNs can be unreliable when the sample size is small, because of three challenges associated with the automatic learning capacity: high sensitivity to hyperparameters, model non-identification, and local irregularity. The first challenge is related to the statistical challenge of balancing approximation and estimation errors of DNNs, the second to the optimization challenge of identifying the global optimum in the DNN training, and the third to the robustness challenge of mitigating locally irregular patterns of estimated functions. To demonstrate the strength and challenges, we estimated the DNNs using a stated preference survey from Singapore and a revealed preference data from London, extracted the full list of economic information from the DNNs, and compared them with those from the DCMs. We found that the economic information either aggregated over trainings or population is more reliable than the disaggregate information of the individual observations or trainings, and that larger sample size, hyperparameter searching, model ensemble, and effective regularization can significantly improve the reliability of the economic information extracted from the DNNs. Future studies should investigate the requirement of sample size, better ensemble mechanisms, other regularizations and DNN architectures, better optimization algorithms, and robust DNN training methods to address DNNs' three challenges to provide more reliable economic information for DNN-based choice models.
Keywords: Deep Neural Network; Machine Learning; Choice Analysis; Interpretability.

1. Introduction

Discrete choice models (DCMs) have been used to examine individual decision making for decades with wide applications to economics, marketing, and transportation [9, 101]. Recently, there is an emerging trend of using machine learning models, particularly deep neural networks (DNNs), to analyze individual decisions. DNNs have shown its extraordinary predictive power across various academic disciplines [63], and in the transportation field, DNNs achieve higher prediction accuracy than DCMs in predicting travel mode choice, automobile ownership, route choice, and many other tasks [77, 84, 111, 20, 21, 54]. However, the interpretability of DNNs is relatively understudied despite the recent progresses [85, 29, 120]: it remains unclear how to obtain reliable economic information from the DNNs in the context of choice analysis.

This study demonstrates that DNNs can provide economic information as *complete* as the classical DCMs, including choice predictions, choice probabilities, market share, substitution patterns of alternatives, social welfare, probability derivatives, elasticities, marginal rates of substitution (MRS), and heterogeneous values of time (VOT). The full list of economic information can be computed by using either the estimated utility and choice probability functions or the input gradients of these functions in DNNs. The process of interpreting DNNs for economic information is different from that of interpreting classical DCMs. The DNN interpretation has to be based on the full *function* of choice probabilities, rather than the *individual parameters* as in classical DCMs. With thousands of individual parameters existing in DNNs, it proves meaningless and unnecessary to delve into individual parameters to extract economic information. We compare the DNNs to the multinomial logit (MNL) model by applying them to a stated preference dataset of travel mode choice in Singapore and a revealed preference dataset in London, showing the robustness of our approach to diverse contexts. This process of interpreting DNNs for economic information can be applied to any choice analysis scenario.

While DNNs can automatically reveal utility functions and behavioral patterns, this power of automatic utility learning comes with three challenges: (1) high sensitivity to hyperparameters, (2) model non-identification, and (3) local irregularity. The first refers to the fact that the estimated DNNs are highly sensitive to the selection of hyperparameters that control the DNN complexity. The second refers to the fact that the optimization in the DNN training often identifies the local minima or saddle points rather than the global optimum, depending on the initialization of the DNN parameters. The third refers to the fact that DNNs have locally irregular patterns such as exploding gradients and the lack of monotonicity to the extent that certain choice behavior revealed by DNNs is not reasonable. The three challenges are embedded respectively in the statistical, optimization, and robustness discussions about DNNs. They become more severe when the sample size is small, but are somewhat mitigated when the sample size is large. While all three challenges create difficulties in interpreting DNN models for economic information, our empirical experiments show that simple random hyperparameter searching, common regularization methods, model ensemble, and information aggregation can partially mitigate these issues.

This study makes the following contributions. This is the first study that systematically dis-

cusses the interpretation of DNNs for economic information in choice analysis, and shows that DNNs can provide economic information as complete as classical DCMs. At the same time, we point out the three challenges of interpreting DNNs for reliable economic information, as well as their theoretical roots. The challenges are different from those in the classical DCMs: DCMs do not even have the notion of hyperparameters, and model non-identification and local irregularity are typically not problems in the DCMs. While this study cannot fully address the challenges in DNN-based choice models, we demonstrate the importance of using large samples, hyperparameter searching, model ensemble, and regularization methods to improve the reliability of the economic information extracted from the DNNs. The paper provides a practical guide for transportation modelers and methodological benchmarks for future researchers to compare to and improve upon. For future researchers to replicate our work, we uploaded our codes to a Github repository: `https://github.com/cjsyzwsh/dnn-for-economic-information.git`.

The paper is structured as follows. Section 2 reviews the studies about DCMs and DNNs concerning prediction, interpretability, sensitivity to hyperparameters, model non-identification, and local irregularity. Section 3 introduces the theory, models, and methods of computing economic information. Section 4 sets up the experiments, and Section 5 discusses the list of economic information obtained from the DNNs. Section 6 concludes the study and discusses the limitations, challenges, and future research.

2. Literature Review

DCMs have been used for decades to analyze the choice of travel modes, travel frequency, travel scheduling, destination and origin, travel route, activities, location, car ownership, and many other decisions in the transportation field [10, 20, 11, 93, 27, 2]. While demand forecasting is important in these applications, all the economic information provides insights to guide policy interventions. For example, market shares can be computed from the DCMs to understand the market power of competing industries [101]. Elasticities of travel demand describe how effective it is to influence travel behavior through the change of tolls or subsidies [94, 45]. VOT, as one important instance of MRS, can be used to measure the monetary gain of saved time after the improvement of a transportation system in a benefit-cost analysis [94, 93].

Recently researchers started to use machine learning models to analyze individual decisions. Karlaftis and Vlahogianni (2011) [55] summarized 86 studies in six transportation fields in which DNNs were applied. Researchers used DNNs to predict travel mode choice [20], car ownership [80], travel accidents [118], travelers' decision rules [24], driving behaviors [52], trip distribution [73], hierarchical demand structure [108], queue lengths [65], parking occupancy [112], metro passenger flows [41], and traffic flows [82, 68, 109, 117, 28, 69]. DNNs are also used to complement the smartphone-based survey [110], improve survey efficiency [91], synthesize new population [17], and impute survey data [30]. In the studies that focus on prediction accuracy, researchers often compare many classifiers, including DNNs, support vector machines, decision trees, random forests, and

DCMs, and typically find that DNNs and RF perform better than the classical DCMs [83, 78, 89, 39, 20]. In other fields, researchers also found the superior performance of DNNs in prediction compared to all the other machine learning (ML) classifiers [33, 58].

Since DNNs are often criticized as a "black-box" model, many resent studies have investigated how to improve its interpretability [29]. Researchers distilled knowledge from DNNs by re-training an interpretable model to fit the predicted soft labels of a DNN [48], visualizing hidden layers in convolutional neural networks [120, 115], using salience or attention maps to identify important inputs [67], computing input gradients with sensitivity analysis [4, 90, 95, 31], using instance-based methods to identify representative individuals for each class [1, 31, 92], or locally approximating functions to make models more interpretable [85]. In the transportation field, only a very small number of studies touched upon the interpretability issue of DNNs for the choice analysis. For example, researchers extracted the elasticity values from DNNs [84], ranked the importance of DNN input variables [39], or visualized the input-output relationship to improve the understanding of DNN models [12]. However, no study has discussed systematically how to compute all the economic information from DNNs, and none have demonstrated the specific practical and theoretical challenges in the process of interpreting DNNs for economic information. The challenges include at least three types, and they are specific to DNNs but not DCMs.

First, DNN performance is highly sensitive to the choice of hyperparameters and model complexity, which is essentially a statistical challenge of balancing approximation and estimation errors. Mathematically, model performance is evaluated by the excess error [1], defined as $\mathbb{E}_S[L(\hat{f}) - L(f^*)]$, in which $L = \mathbb{E}_{x,y}[l(y, f(x))]$ and $l(y, f(x))$ is the loss function; f^* is the true data generating function; S is the sample $\{(x_i, y_i)_1^N\}$. This excess error can be further decomposed: $\mathbb{E}_S[L(\hat{f}) - L(f^*)] = \mathbb{E}_S[L(\hat{f}) - L(f_F) + L(f_F) - L(f^*)]$, in which $\mathbb{E}_S[L(\hat{f}) - L(f_F)]$ is referred to as the estimation error and $\mathbb{E}_S[L(f_F) - L(f^*)]$ as the approximation error; f_F is the best function to approximate true model f^* in the hypothesis function space F. A complex model tends to have larger estimation errors and smaller approximation errors, and a simple model is the opposite. DNNs have very small approximation errors because it has been proven to be a universal approximator [51, 50, 25], which also leads to the large estimation error as an issue. The large estimation error in DNNs can be examined by using statistical learning theory [18, 106, 103, 107, 104]. Formally, the model complexity can be measured by the Vapnik-Chervonenkis (VC) dimension (v), which provides an upper bound on DNNs' estimation error (proof is available in Appendix I). Recently, progress has been made to provide a tighter upper bound on the estimation error of DNNs by using other methods [7, 3, 75, 36]. It is important to note that to select DNNs' hyperparameters is to control DNNs' model complexity, which balances between approximation and estimation errors. When either the approximation errors or the estimation errors are high, the overall DNN performance is low.

Researchers can control the model complexity of DNNs by using architectural and regularization hyperparameters, hyperparameter optimization methods, and model ensemble. In a standard feed-

[1] This excess error can also be referred to as generalization error, since it measures the capacity of estimated $\hat{f}$ being generalized to other contexts.

forward DNN, the architectural hyperparameters include depth and width, and the regularization hyperparameters include the L_1 and L_2 penalty constants, training iterations, minibatch sizes, data augmentation, dropouts, early stopping, and others [38, 16, 59, 105, 116]. To choose from the large number of hyperparameters, researchers used random search [14, 13], grid search [13, 14], Gaussian process [96], multi-fidelity optimization [32, 66], or even reinforecement learning [121]. Besides the common regularization hyperparameters, model ensemble is a particularly useful way to reduce the excess error of DNNs. Hansen and Salamon (1990) formally introduced neural network ensemble to improve DNNs' generalizability [40]. Researchers proved that the generalization error of an ensemble model is always smaller than the average of individual models [60]. Recently, researchers used neural network ensembles to predict financial behavior and human activities, showing the superior power of model ensemble over individual models [102, 53]. In our study, while it is impossible to incorporate all these methods, we will demonstrate that several baseline methods are effective in improving the reliability of economic information in DNNs.

Second, DNN models are not identifiable, because the empirical risk minimization (ERM) is non-convex with high dimensionality. Given the ERM being non-convex, the DNN training is highly sensitive to the initialization [44, 35]. With different initializations, the DNN model can end with local minima or saddle points, rather than the global optimum [38, 26]. This issue does not happen in the classical MNL models, because the ERM of the MNL models is globally convex [19]. Decades ago, model non-identification was one reason why DNNs were discarded [63]. Nowadays, however, researchers argue that some high quality local minima are acceptable, and the global minimum in the training may be irrelevant since the global minimum tends to overfit [22]. This problem of model non-identification indicates that each training of DNNs can lead to very different models, even conditioned on the fixed hyperparameters and training samples. Importantly these trained DNNs may have very similar prediction performance, creating difficulties for researchers to choose the final model for interpretation.

Third, the choice probability functions in DNNs can be locally irregular because their gradients can be exploding or the functions themselves are non-monotonic, both of which are discussed under the robust DNN framework. When the gradients of choice probability functions are exploding, it is very simple to find an adversarial input x', which is ϵ-close to the initial x ($||x' - x||_p \leq \epsilon$) but is wrongly predicted to be a label different from the initial x with high confidence. This type of system is not robust because they can be easily fooled by the adversarial example x'. In fact, it has been found that DNNs lack robustness [76, 100]. With even a small ϵ perturbation introduced to an input image x, DNNs label newly generated image x' to the wrong category with extremely high confidence, when the correct label should be the same as the initial input image x [100, 37]. Therefore, the lack of robustness in DNNs implies the locally irregular patterns of the choice probability functions and the gradients, which are the key information for DNN interpretation.

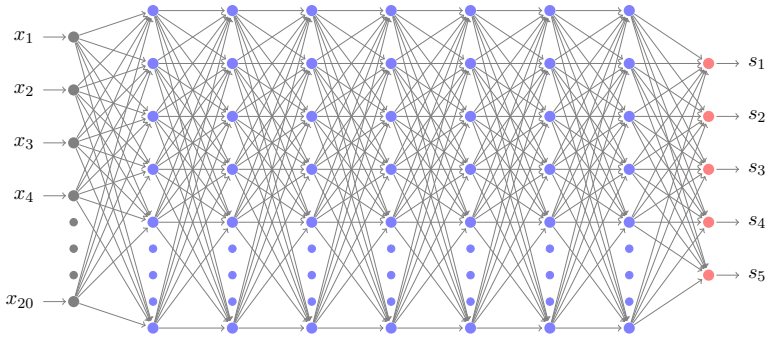


Fig. 1. A feedforward DNN architecture (7 hidden layers * 100 neurons)

3. Model

3.1. DNNs for Choice Analysis

DNNs can be applied to choice analysis. Let $s_k^*(x_i)$ denote the true probability of individual i choosing alternative k out of $[1, 2, ..., K]$ alternatives, with x_i denoting the input variables: $s_k^*(x_i) : R^d \rightarrow [0, 1]$. Individual i's choice $y_i \in \{0, 1\}^K$ is sampled from a multinomial random variable with $s_k^*(x_i)$ probability of choosing k. With DNNs applied to choice analysis, the choice probability function is:

$$s_k(x_i) = \frac{e^{V_{ik}}}{\sum_j e^{V_{ij}}} \tag{1}$$

in which V_{ij} and V_{ik} are the jth and kth inputs into the Softmax activation function of DNNs. V_{ik} takes the layer-by-layer form:

$$V_{ik} = (g_m^k \circ g_{m-1} ... \circ g_2 \circ g_1)(x_i) \tag{2}$$

where each $g_l(x) = ReLU(W_l x + b_l)$ is the composition of linear and rectified linear unit (ReLU) transformation; g_m^k represents the transformation of the last hidden layer into the utility of alternative k; and m is the total number of layers in a DNN. Figure 1 visualizes a feedforward DNN architecture with 20 input variables, 5 output alternatives, and 7 hidden layers. The grey nodes represent the input variables; the blue ones represent the hidden layers; and the red ones represent the Softmax activation function. The layer-by-layer architecture in Figure 1 reflects the compositional structure of Equation 2.

The inputs into the Softmax layers in DNNs can be treated as utilities, the same as those in the classical DCMs. This utility interpretation in DNNs is shown by the Lemma 2 in McFadden (1974) [71], which implies that the Softmax activation function is equivalent to a random utility term with Gumbel distribution under the random utility maximization (RUM) framework. Hence DNNs and MNL models are both under the RUM framework, and their difference only resides in the utility specifications. In other words, the inputs into the last Softmax activation function of DNNs can be

interpreted as utilities; the outputs from the Softmax activation function are choice probabilities; the transformation before this Softmax function can be seen as the specification of utility functions; and the Softmax activation function can be seen as the comparison of utility values.

Despite their similarity, DNNs are a much more generic model family than MNL models, and this relationship can be understood from various perspectives. The universal approximator theorem developed in the 1990s indicates that a neural network with only one hidden layer is asymptotically a universal approximator when the width becomes infinite [25, 51, 50]. Recently this asymptotic perspective leads to a more non-asymptotic question, asking why depth is necessary when a wide and shallow neural network is already powerful enough. It can be shown that DNNs can approximate functions with an exponentially smaller number of neurons than a shallow neural network in many settings [23, 86, 81]. In other words, DNNs can be treated as an efficient universal approximator, thus being much more generic than the MNL model, which is a shallow neural network with zero hidden layers. However, a more generic model family leads to both smaller approximation errors and large estimation errors. Since the out-of-sample prediction error equals to the sum of the approximation and estimation errors, DNNs do not necessarily outperform MNL models from a theoretical perspective. The major challenge of DNNs is its large estimation error, which is associated with its extraordinary approximation power. To find the best balance between the approximation and estimation errors, the procedure of hyperparameter searching needs to be used since the hyperparameters, such as the DNNs' depth and width, control the model complexity. A brief theoretical proof about the large estimation error of DNNs is available in Appendix I. More detailed discussions are available in the recent studies from statistical learning theory [104, 107, 36, 75, 7, 64, 6].

3.2. Computing Economic Information From DNNs

The utility interpretation in DNNs enables us to derive all the economic information traditionally obtained from DCMs. With $\hat{V}_k(x_i)$ denoting the estimated utility of alternative k and $\hat{s}_k(x_i)$ the estimated choice probability function, Table 1 summarizes the formula of computing the economic information, which is sorted into two categories. Choice probabilities, choice predictions, market share, substitution patterns, and social welfare are derived by using functions (either choice probability or utility functions). Probability derivatives, elasticities, MRS, and VOTs are derived from the gradients of choice probability functions. This differentiation is owing to the the different theoretical properties between functions and their gradients [2]. The formula in Table 1 can be applied to both DNNs and MNLs, but the MNLs have a further explicit parametric form for each piece of economic information while DNNs don't [101].

This process of interpreting economic information from DNNs is significantly different from the classical DCMs for the following reasons. In DNNs, the economic information is directly computed by using the *full functions* $\hat{s}_k(x_i)$ and $\hat{V}_k(x_i)$, rather than *individual parameters* $\hat{w}$. This focus

[2]The uniform convergence proof is possible for the estimated functions, while it is much harder for the gradients because the estimated functions may not be even differentiable.

Table 1: Formula to compute economic information in both DNNs and DCMs; F stands for function, GF stands for the gradients of functions.

Economic Information	**Formula in DNNs**	**Categories**
Choice probability	$\hat{s}_k(x_i)$	F
Choice prediction	$\underset{k}{\text{argmax}}\ \hat{s}_k(x_i)$	F
Market share	$\sum_i \hat{s}_k(x_i)$	F
Substitution pattern between alternatives k_1 and k_2	$\hat{s}_{k_1}(x_i)/\hat{s}_{k_2}(x_i)$	F
Social welfare	$\sum_i \frac{1}{\alpha_i} \log(\sum_{k=1}^{K} e^{\hat{V}_{ik}}) + C$	F
Change of social welfare	$\sum_i \frac{1}{\alpha_i} \big[\log(\sum_{k=1}^{K} e^{\hat{V}^1_{ik}}) - \log(\sum_{k=1}^{K} e^{\hat{V}^0_{ik}}) \big]$	F
Probability derivative of alternative k w.r.t. x_{ij}	$\partial \hat{s}_k(x_i)/\partial x_{ij}$	GF
Elasticity of alternative k w.r.t. x_{ij}	$\partial \hat{s}_k(x_i)/\partial x_{ij} \times x_{ij}/\hat{s}_k(x_i)$	GF
Marginal rate of substitution between x_{ij_1} and x_{ij_2}	$-\frac{\partial \hat{s}_k(x_i)/\partial x_{ij_1}}{\partial \hat{s}_k(x_i)/\partial x_{ij_2}}$	GF
VOT (x_{ij_1} is time and x_{ij_2} is monetary value)	$-\frac{\partial \hat{s}_k(x_i)/\partial x_{ij_1}}{\partial \hat{s}_k(x_i)/\partial x_{ij_2}}$	GF

on functions rather than individual parameters in DNNs is inevitable because the non-convex and high-dimensional DNN training leads to unstable parameter estimates, while MNL has the same estimate in every training owing to the convexity of its empirical risk minimization. This focus on the full functions is also consistent with other studies concerning the interpretation of DNNs: a large number of recent studies focused on the full functions of DNNs for interpretation, while none focused on individual neurons/parameters [72, 48, 4, 87]. Hence the DNN interpretation can be seen as an end-to-end mechanism without involving the individual parameters as an intermediate process. In addition, the interpretation of DNNs is a prediction-driven process: the economic information is generated in a post-hoc manner after a model is trained to be highly predictive. This prediction-driven interpretation takes advantage of DNNs' capacity of automatic feature learning, and it is also in contrast to the classical DCMs that rely on handcrafted utility functions. This prediction-driven interpretation is based on the belief that "when predictive quality is (consistently) high, some structure must have been found" [74].

3.3. MNLs for Choice Analysis

The classical MNL with linear specification is used as the reference point to the DNNs in our empirical experiments. The utility function in the MNL models is shown as the following:

$$V_{ik} = \beta_{0,k} + \beta_{x,k}^T x_{ik} + \beta_{z,k}^T z_i, \quad as \quad k \neq ref \tag{3}$$

$$V_{ik} = \beta_{x,k}^T x_{ik}, \quad as \quad k = ref \tag{4}$$

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. For parameter identification, the utility specification is different depending on whether the alternative is used as the reference. This formulation is the simplest specification that guarantees the parameter identification in choice modeling. A more generic form is $V_{ik} = \beta_{0,k} + \beta_{x,k}^T \phi_x(x_{ik}) + \beta_{z,k}^T \phi_z(z_i)$, in which ϕ_x and ϕ_z represent the functions for feature transformation, such as quadratic and log transformation. This study uses the linear specification for two reasons. First for fairness, both 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 DNNs to the MNLs with feature transformations.

4. Setup of Experiments

The experiments use two groups of DNN models, referred to as Random-DNNs and Opt-DNNs. Random-DNNs are those DNNs trained with varying hyperparameters *randomly* chosen within a prespecified hyperparameter space, and Opt-DNNs refer to the repeated trainings of the DNNs with the *fixed* hyperparameters that perform the best in the Random-DNNs.

4.1. Random-DNNs: Hyperparameter Training

The group of Random-DNNs is constructed by randomly exploring a pre-specified hyperparameter space and using the sampled set of hyperparameters for each DNN training [14]. The hyperparameter space consists of the architectural hyperparameters, including the depth and width of DNNs; and the regularization hyperparameters, including L_1 and L_2 penalty constants, and dropout rates. 100 sets of hyperparameters are randomly generated for comparison. The details of the hyperparameter space is available in Appendix II. Besides the hyperparameters varying across the 100 models, all the DNN models share certain fixed components, including ReLU activation functions in the hidden layers, Softmax activation function in the last layer, Gloret initialization, and Adam optimization, following the standard practice [38, 34]. Formally, the hyperparameter searching is formulated as

$$\hat{w}_h = \underset{w_h \in \{w_h^{(1)}, w_h^{(2)}, \ldots, w_h^{(S)}\}}{\operatorname{argmin}} \; \underset{w}{\operatorname{argmin}} \; L(w, w_h) \tag{5}$$

where $L(w, w_h)$ is the empirical risk function that the DNN training aims to minimize, w represents the parameters in a DNN architecture, w_h represents the hyperparameter, $w_h^{(s)}$ represents one set of hyperparameters randomly sampled from the hyperparameter space, and $\hat{w}_h$ is the chosen hyperparameter with the highest prediction accuracy. Besides this baseline random searching, other approaches can be used for hyperparameter training, such as reinforcement learning or Bayesian methods [97, 121], which are beyond the scope of our study.

4.2. Opt-DNNs: Training with Fixed Hyperparameters

After the hyperparameter searching, we examine a set of optimum hyperparameters that lead to the highest prediction accuracy. By using the same training set and the fixed set of optimum hyperparameters, we train the DNN models another 100 times to construct the group of Opt-DNNs. Each training seeks to minimize the empirical risk conditioned on the fixed hyperparameters, formulated as following.

$$\min_{w} L(w, \hat{w}_h) = \min_{w} \frac{1}{N} \sum_{i=1}^{N} l(y_i, s_k(x_i; w, \hat{w}_h)) + \gamma ||w||_p \tag{6}$$

where w represents the parameters; $\hat{w}_h$ represents the best hyperparameters; $l()$ is the loss function, typically the cross-entropy loss; N is the sample size. $\gamma||w||_p$ represents L_p penalty ($||w||_p = (\sum_j (w_j)^p)^{\frac{1}{p}}$), and L_1 (LASSO) and L_2 (Ridge) penalties are the two specific cases of L_p penalties. Note that DNNs have the model non-identification challenge because the objective function in Equation 6 is not globally convex. DNNs have the local irregularity challenge because this optimization over the *global* prediction risks is insufficient to guarantee the *local* fidelity. The two issues will be demonstrated in more details in Section 5.

4.3. Datasets

Our experiments rely on two datasets: a stated preference survey conducted in Singapore and a revealed preference data in London. The Singapore data set was collected by the authors, with the help of a professional survey company Qualtrics in July 2017. The survey started with asking the respondents to report their postal codes of their home and working locations and their current travel mode. From respondents' home (origin) and work (destination) locations, we computed the walking time, waiting time, in-vehicle travel time, and travel cost of each travel mode for each individual's commute trip using Google Maps API. The information was then used to automatically generate the stated preference section, which was the bulk of the questionnaire. The respondents were asked to choose among five travel modes: walking, public transit, driving, ride sharing, and shared autonomous vehicles, with varying values of price and travel time. In the end, respondents reported socioeconomic information such as gender, education, and income. The London data set was publicly provided in Hillel, et al. (2018) [46], in which the authors constructed a new data set based on the London Travel Demand Survey (LTDS) by combining the individual trip records and the trip trajectories along the mode alternatives. The authors started with LTDS, removed the trips that had the same origin-destination post codes, assigned each trip to one of four main travel modes (walking, cycling, public transport, and driving), and simplified the trip purposes to five purposes (B, HBW, HBE, HBO, and NHBO). The authors then augmented travel time and price information to the initial LTDS by using Google Map API and Oyster cards.

To understand the impacts of different sample sizes and contexts, we constructed three datasets from these two initial data sets for our experiments: (1) the SGP dataset with the full $8,418$

observations (8K-SGP), (2) the LD dataset with the full 81,086 observations (80K-LD), and (3) the LD dataset with randomly sampled 8,000 observations (8K-LD). The comparison between the 8K-SGP and 8K-LD datasets reveals the effect of two contexts, and the comparison between the 80K-LD and 8K-LD datasets reveals the effect of the sample size. All the datasets are split into training and testing sets by using the default random seed in the Numpy module in Python. The sample sizes of training and testing sets in 8K-SGP are 7,015 and 1,403, and that in 80K-LD are 72,977 and 8,109. In both datasets, the choice variable y is travel mode choice, including five alternatives (walking, taking public transit, ride sharing, using an autonomous vehicle, and driving) in the SGP dataset and four alternatives (walking, cycling, driving, and using public transit) in the LD dataset. The explanatory variables include 20 individual-specific and alternative-specific variables in the SGP dataset and 14 variables in the LD dataset. Please refer to Appendix III for the summary statistics of the two data sets.

5. Experimental Results

This section shows that it is feasible to extract all the economic information from DNNs without using individual parameters, and that by using large sample, hyperparameter searching, model ensemble, and regularization methods, it is possible to extract reliable economic information. We will first present prediction accuracy, then the function-based interpretation for choice probabilities, substitution patterns of alternatives, market share, and social welfare, and lastly the gradient-based interpretation for probability derivatives, elasticities, VOT, and heterogeneous preferences. This section summarizes two groups of DNN models (Opt-DNNs and Random-DNNs) and the linear MNL model, applied to the 8K-SGP, 80K-LD, and 8K-LD datasets.

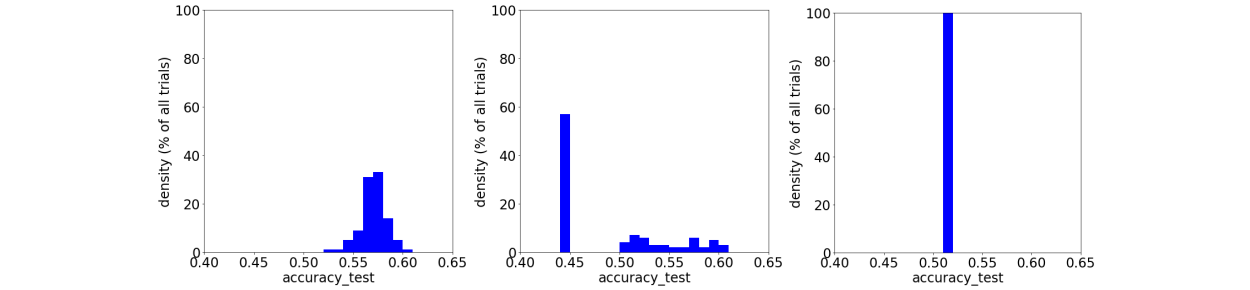
5.1. Prediction Accuracy

Figures 2a-2g reveal three findings. First, Opt-DNNs on average outperform the MNL models by about 2 to 8 percentage points prediction accuracy, which is consistent with the previous studies that found the outperformance of DNNs over MNL [84, 77, 55]. Second, choosing the correct hyperparameter plays a critical role in improving the model performance of DNNs, as shown by the higher prediction accuracy of the Opt-DNNs than the Random-DNNs in both 8K-SGP and 80K-LD datasets. Third, larger sample size improves the prediction accuracy of DNNs, as shown by comparing Figures 2d and 2e.

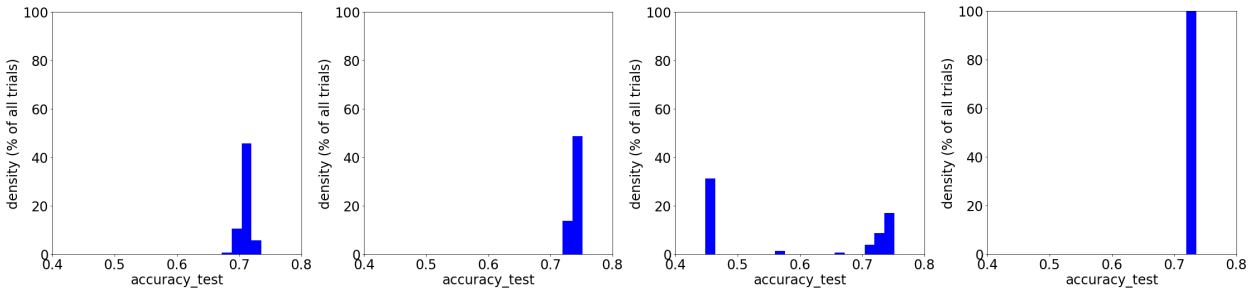
5.2. Function-Based Interpretation

5.2.1. Choice Probability Functions

The choice probability functions are visualized in Figure 3. Since the inputs of the choice probability functions $s(x)$ have high dimensions, the $s(x)$ is visualized by computing the driving probability with varying only the driving cost, holding all the other variables constant at the sample mean.



(a) Opt-DNNs (8K-SGP) (b) Random-DNNs (8K-SGP) (c) MNL (8K-SGP)



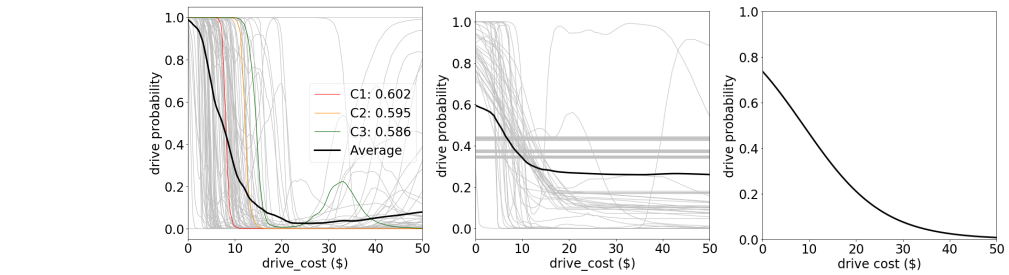
(d) Opt-DNNs (8K-LD) (e) Opt-DNNs (80K-LD) (f) Random-DNNs (80K-LD) (g) MNL (80K-LD)

Fig. 2. Histograms of the prediction accuracy in seven scenarios (100 trainings for each model group)

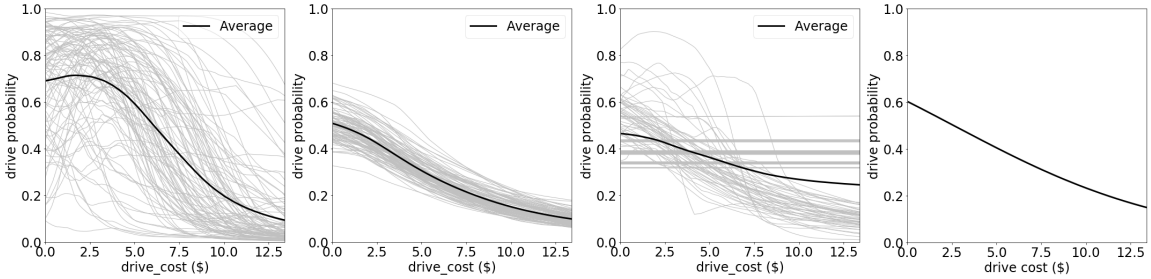
Each light grey curve in Figures 3a-3g represents one individual training result, and the dark curve is the ensemble of all 100 models. In Figures 3c and 3g, only one training result is visualized because the MNL training has no variation.

Figure 3 demonstrates the power of DNNs being able to automatically learn the choice probability functions in both large and small sample sizes. With large sample size, the choice probability functions in Figure 3e are highly concentrated and are quite reasonable, similar to the pattern in Figure 3g. Even with small sample size, the majority of the choice probability functions in Figures 3a and 3d are roughly decreasing. The choice probabilities are high when the driving cost is close to zero, while they become much smaller when the driving cost increases. This roughly decreasing pattern is reasonable from a behavioral perspective. In comparison to the choice probability functions of MNL (Figure 3c), the choice probability functions of the Opt-DNNs in Figure 3a are richer and more flexible. However, the caveat is that the DNN choice probability functions may be too flexible to reflect the true behavioral mechanism in the small sample, owing to three theoretical challenges.

First, the large variation of the Random-DNNs in Figures 3b and 3f reveal that DNN models are sensitive to the choice of hyperparameters. With different hyperparameters, some of Random-DNNs' choice probability functions are simply flat without revealing any useful information, while others are similar to Opt-DNNs with reasonable patterns. This challenge can be mitigated by



(a) Opt-DNNs (8K-SGP) (b) Random-DNNs (8K-SGP) (c) MNL (8K-SGP)



(d) Opt-DNNs (8K-LD) (e) Opt-DNNs (80K-LD) (f) Random-DNNs (80K-LD) (g) MNL (80K-LD)

Fig. 3. Driving probability functions with driving costs (100 trainings for each model group)

hyperparameter searching. For example, the Opt-DNNs can reveal more reasonable economic information than the Random-DNNs because the Opt-DNNs use specific architectural and regularization hyperparameters, chosen from the results of hyperparameter searching based on their high prediction accuracy.

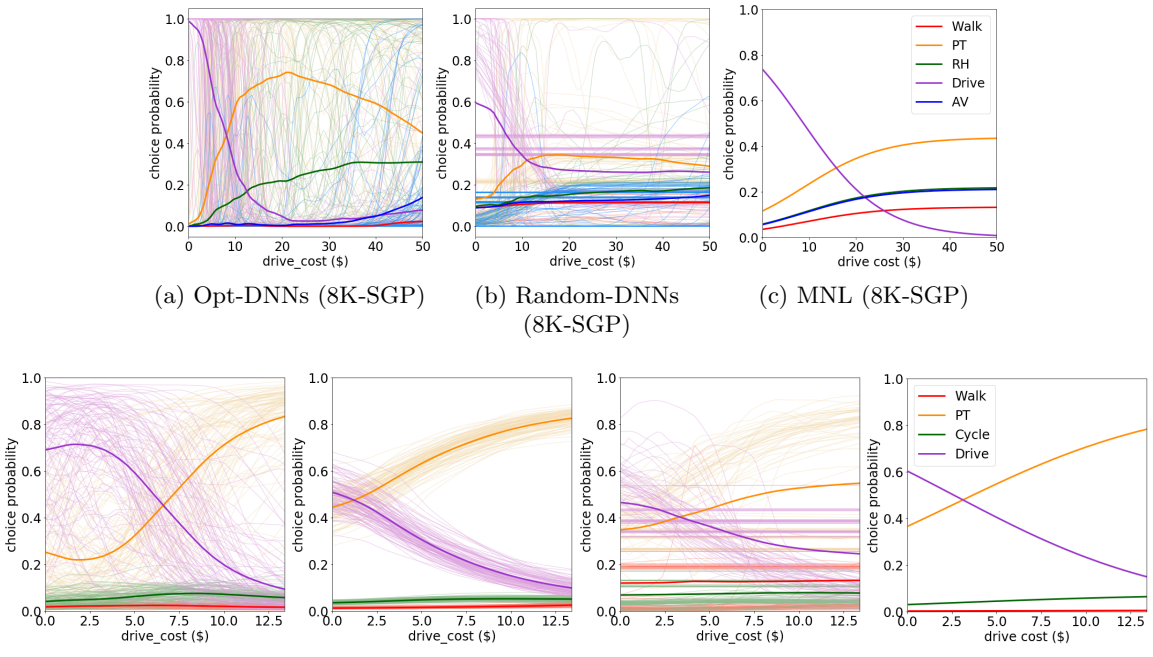
Second, the large variation of the individual Opt-DNN trainings (Figures 3a, 3d, and 3e) reveal the challenge of model non-identification. Given that the 100 trainings are conditioned on the same training data and the same set of hyperparameters, the variation across the Opt-DNNs can only be attributed to the model non-identification issue, or more specifically, the optimization difficulties in minimizing the non-convex risk function of DNNs. As DNNs' risk function is non-convex, different model trainings can converge to very different local minima or saddle points. Whereas these local minima have similar prediction accuracy, it brings difficulties to the model interpretation since the functions learnt from different local minima are different. For example in Figure 3a, the three individual training results (C1, C2, and C3) have very similar out-of-sample prediction accuracy (60.2%, 59.5%, and 58.6%); however, their corresponding choice probability functions are very different. In fact, the majority of the 100 individual trainings have quite similarly high prediction accuracy, whereas their choice probability functions differ from each other. On the other side, the choice probability function averaged over the 100 trainings of the Opt-DNNs is more stable than individual ones, demonstrating the importance of model ensemble in controlling generalization errors and smoothing choice probability functions.

Third, the shapes of the individual choice probability curves show the local irregularity of the

choice probability functions. Let's take Figure 3a as an example. Some choice probability functions can be sensitive to the small change of input values: the probability of choosing driving in C1 drops from 96.6% to 7.8% as the driving cost increases from \$7 to \$9, indicating a locally exploding gradient. This phenomenon of exploding gradients is acknowledged in the robust DNN discussions, because exploding gradients render a system vulnerable [88, 87]. Many training results present a non-monotonic pattern. For example, C3 represents a counter-intuitive case where the probability of driving starts to increase dramatically as the driving costs are larger than \$25. However, it is worth noting that this local irregularity problem can be mitigated with model ensemble and larger sample size: the choice probability functions in Figure 3e are much more regular than those in Figure 3d.

5.2.2. Substitution Pattern of Alternatives

The substitution pattern of the alternatives is of both practical and theoretical importance in choice analysis. In practice, researchers need to understand how market shares vary with input variables; in theory, the substitution pattern constitutes the critical difference between multinomial logit, nested logit, and mixed logit models. Figure 4 visualizes how the choice probability functions vary as the driving cost increases. By visualizing the choice probabilities of all the alternatives, Figure 4 is an one-step extension of Figure 3.



(a) Opt-DNNs (8K-SGP) (b) Random-DNNs (8K-SGP) (c) MNL (8K-SGP)

(d) Opt-DNNs (8K-LD) (e) Opt-DNNs (80K-LD) (f) Random-DNNs (80K-LD) (g) MNL (80K-LD)

Fig. 4. Substitution patterns of all the alternatives with varying driving costs

The substitution pattern of the Opt-DNNs is more reasonable than that of the Random-DNNs and more flexible than that of the MNL models. As shown in Figure 4a, when the driving cost

is smaller than \$20, the substitution pattern of the Opt-DNNs aggregated over the 100 models illustrates that the five alternatives are substitute to each other, since the driving probability is decreasing while others are increasing. When the driving cost is larger than \$20, the substitution pattern between walking, ridesharing, driving, and using an AV still reveals the substitute nature. This reasonable substitution pattern is retained in the LD dataset (Figures 4d and 4e). In a choice modeling setting, the alternatives in a choice set are typically substitutes: people are expected to switch from driving to other travel modes, as the driving cost increases. Therefore, the aggregated substitution pattern has mostly reflected the correct relationship of the five alternatives.

However, the three theoretical challenges also permeate into the substitution patterns, particularly when the sample size is small. The large variation in Figures 4b and 4f illustrates the high sensitivity to hyperparameters; the large variation in Figures 4a, 4d, and 4e illustrates the model non-identification problem; and the individual curves in Figures 4a, 4d reveal the local irregularity. Note that larger sample, hyperparameter searching, and regularization methods can mitigate but not fully solve these problems. For example in Figure 4a, the average substitution pattern of the Opt-DNNs indicate that people are less likely to choose the public transit as the driving cost increases, when the driving cost is larger than \$20. As a comparison, the substitution pattern in the two MNL models, although perhaps exceedingly restrictive, reflects the travel mode alternatives being substitute goods.

5.2.3. Market Shares

Table 2 summarizes the market shares predicted by the three model groups in the 8K-SGP and 80K-LD datasets. We found that, while the choice probability functions of Opt-DNNs can be locally unreasonable, the aggregated market shares of Opt-DNNs are very close to the true market shares, and the market shares predicted by Opt-DNNs are as accurate as the MNL in the testing sets. Specifically, in both the training and testing sets of the 8K-SGP and 80K-LD data sets, the errors between the predicted market shares of Opt-DNNs and the true market shares are within the range of 1.0%. It appears that the three challenges of DNNs do not emerge in the calculation of market shares. The local irregularity could be cancelled out owing to the aggregation over the sample. The model non-identification appears less a problem as the market shares across the Opt-DNN trainings are very stable, as shown by the small standard deviations in the parenthesis. The high sensitivity to hyperparameters is addressed by the selection of the Opt-DNNs from the hyperparameter searching process, as the market shares of the Opt-DNNs are much more accurate than the Random-DNNs. It is also interesting to compare Opt-DNNs and MNL: while MNL is guaranteed to provide market shares exactly the same as the true ones in the training sets [3], the predicted market shares of Opt-DNNs are as accurate as MNL in the testing sets. Specifically, we computed the sum of the absolute errors between the predicted and the true market shares for Opt-DNNs and MNL in the testing sets. In the 8K-SGP, the absolute error of Opt-DNNs is 4.28%,

[3]The first order conditions in training MNL actually match the estimated and true market shares in the training sets.

slightly higher than 3.05% from MNL; in the 80K-LD, the absolute error of Opt-DNNs is 1.90%, slightly lower than 2.10% from MNL.

Table 2: Market shares of travel modes (training and testing); each entry represents the average value of the market share over 100 trainings, and the number in the parenthesis is the standard deviation.

Panel 1. Training sets				
	Opt-DNNs (8K-SGP)	Random-DNNs (8K-SGP)	MNL (8K-SGP)	True Market Share
Walk	10.2% (0.6%)	12.7% (2.4%)	10.6%	10.6%
Public Transit	23.1% (1.0%)	20.8% (2.8%)	23.0%	23.0%
Ride Hail	10.5% (0.6%)	12.7% (2.4%)	10.7%	10.7%
Drive	45.6% (0.8%)	41.0% (4.5%)	44.9%	44.9%
AV	10.6% (0.6%)	12.80% (2.4%)	10.8%	10.8%
	Opt-DNNs (80K-LD)	Random-DNNs (80K-LD)	MNL (80K-LD)	True Market Share
Walk	17.9% (1.8%)	19.7% (3.6%)	17.6%	17.6%
Public Transit	35.1% (2.5%)	32.7% (3.8%)	35.3%	35.3%
Cycle	2.9% (0.3%)	6.9% (4.6%)	3.0%	3.0%
Drive	44.1% (2.8 %)	40.7% (4.6%)	44.1%	44.1%
Panel 2. Testing sets				
	Opt-DNNs (8K-SGP)	Random-DNNs (8K-SGP)	MNL (8K-SGP)	True Market Share
Walk	9.1% (1.3%)	12.3% (2.7%)	10.34%	9.48%
Public Transit	23.4% (2.1%)	21.0% (3.2%)	23.1%	23.9%
Ride Hail	10.3% (1.2%)	12.7% (2.5%)	10.5%	10.8%
Drive	46.7% (1.8%)	41.2% (4.9%)	45.2%	44.5%
AV	10.5% (1.3%)	12.8% (2.5%)	10.8%	11.2%
	Opt-DNNs (80K-LD)	Random-DNNs (80K-LD)	MNL (80K-LD)	True Market Share
Walk	18.0% (1.8%)	19.8% (3.6%)	17.7%	17.3%
Public Transit	35.0% (2.5%)	32.6% (3.8%)	35.3%	34.7%
Cycle	2.8% (0.3%)	6.9% (4.6%)	2.9%	2.8%
Drive	44.2% (2.8%)	40.7% (4.6%)	44.1%	45.1%

5.2.4. Social Welfare

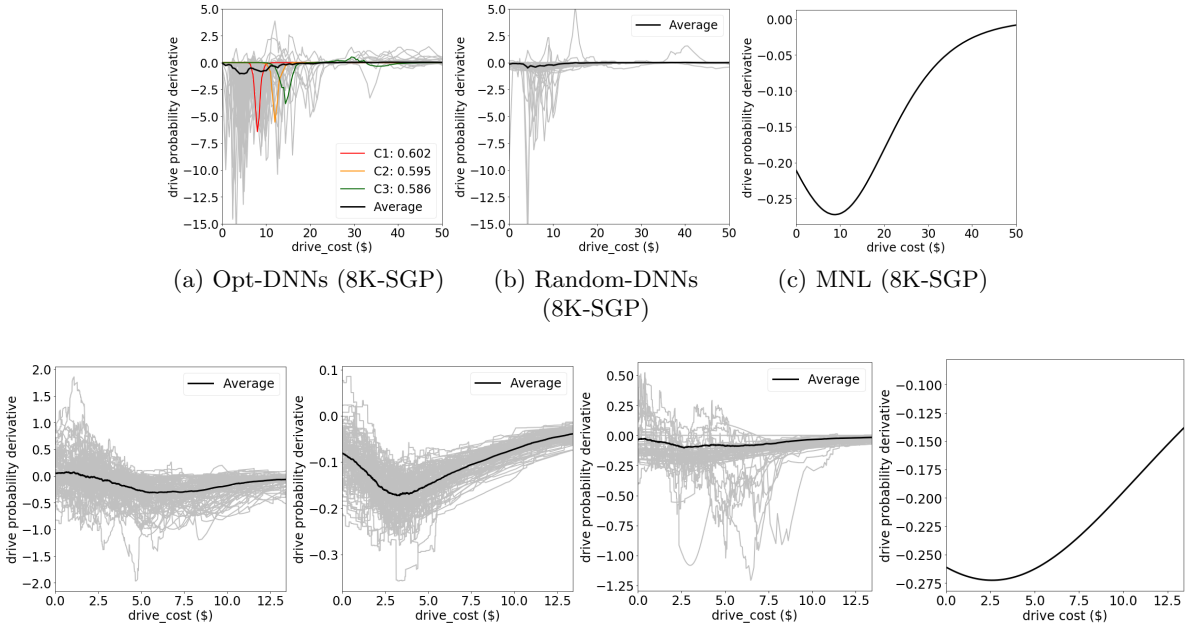
Since DNNs have an implicit utility interpretation, we can observe how social welfare changes as action variables change the values. To demonstrate this process, we simulate one dollar decrease of the driving cost, and calculate the average social welfare change in the Opt-DNNs in the 8K-SGP dataset. We found that the social welfare increases by about \$520 in the Opt-DNN models after averaging over all 100 trainings. Interestingly, the magnitude of this social welfare change (\$520) is very intuitive and consistent with the one computed from MNL models, which is \$491 dollars. In the process of computing the social welfare change, we used the α_i averaged across 100 trainings as the individual i's marginal value of utility, slightly different from the formula in Table 1. Specifically, the formula is $\sum_i \frac{1}{\bar{\alpha}_i}\left[\log(\sum_{k=1}^{K} e^{\hat{V}_{ik}^1}) - \log(\sum_{k=1}^{K} e^{\hat{V}_{ik}^0})\right]$, in which $\bar{\alpha}_i = \frac{1}{S}\sum_s \alpha_{i,s}$; $\alpha_{i,s}$ represents individual i's marginal value of utility for each DNN model s; V_{ik}^1 and V_{ik}^0 represent the utility values under two different scenarios. Without using $\bar{\alpha}_i$, individuals' marginal value of utility can take unreasonable values. The problem associated with the disaggregate gradient information

will be discussed in details in the following section.

5.3. Gradient-Based Interpretation

5.3.1. Gradients of Choice Probability Functions

The gradient of choice probability functions offers opportunities to extract more important economic information. Since researchers often seek to understand how to take actions to trigger behavioral changes, the most relevant information is the partial derivative of the choice probability function with respect to a targeting input variable. Figure 5 visualizes the corresponding derivatives of the choice probability functions. As shown below, both the strength and the challenges identified in the choice probability functions are retained in the probability derivatives.



(a) Opt-DNNs (8K-SGP) (b) Random-DNNs (8K-SGP) (c) MNL (8K-SGP)

(d) Opt-DNNs (8K-LD) (e) Opt-DNNs (80K-LD) (f) Random-DNNs (80K-LD) (g) MNL (80K-LD)

Fig. 5. Probability derivatives of choosing driving with varying driving costs

In Figure 5a, the majority of the Opt-DNNs, such as the three curves (C1, C2, and C3), take negative values and have inverse bell shapes. This inverse bell shaped curve is intuitive because people are not as sensitive to price changes when price is close to zero or infinity, but are more sensitive when price is close to a certain tipping point. The shapes revealed by Opt-DNNs are similar to the MNL models. The probability derivative of MNL models is $\partial s(x)/\partial x = s(x)(1 - s(x)) \times (\partial V(x)/\partial x)$, which is mostly negative and take a very regular truncated inverse bell shape, as shown in Figures 5c and 5g.

The sensitivity to hyperparameters, the model non-identification, and the local irregularity also exist here for probability derivatives. Random-DNNs reveal more unreasonable behavioral patterns than Opt-DNNs, as many of the input gradients are flat on zero, demonstrating the

importance of selecting correct hyperparameters. The variation of individual trainings in Figures 5a, 5d, and 5e demonstrates the challenge of model non-identification. With fixed training samples and hyperparameters, the DNN trainings can lead to different training results, thus creating difficulty for researchers to choose a final model for interpretation. The exploding gradients and the non-monotonicity issues, as the two indicators of local irregularity, are also clearly illustrated in the individual trainings in Figures 5a and 5d, although are less severe in Figure 5e. The absolute values of many probability derivatives are of large magnitude; for example in Figure 5a, at the peak of the C1 curve, \$1 increase of driving costs leads to about 6.5% change in choice probability[4], which is much larger than the MNL models. Similar to the previous discussions, large sample size, hyperparameter searching, model ensemble, regularization, and information aggregation can mitigate these challenges.

5.3.2. Elasticities

To compare across input variables, researchers often compute elasticities because the elasticities are standardized derivatives. Given that DNNs provide choice probability derivatives, it is straightforward to compute the elasticities from DNNs. For MNL models, the formula to compute elasticities are attached in Appendix IV. Tables 3 and 4 present the elasticities of travel mode choices with respect to input variables. In Panel 1, each entry represents the average elasticity of the respondents in the testing set based on one Opt-DNN model, and the value in the parenthesis is the standard deviation of individuals' elasticity values. Panel 2 is the average elasticity of the testing set from a MNL model with linear specification, and the value in the parenthesis represents the standard deviation.

The average elasticities in the Opt-DNN are reasonable in terms of both the signs and magnitudes. We highlight the elasticities that relate the travel modes to their own alternative-specific variables. These highlighted elasticities are all negative, which is very reasonable since higher travel cost and time should lead to lower probability of adopting the corresponding travel mode. In Table 3, the magnitudes in the DNN models are higher than the typical results from the MNL models, although the relative magnitudes of the elasticity coefficients in DNNs are similar to the MNL model. For example, DNN models indicate that 1% increase in public transit cost, walking time, waiting time, and in-vehicle travel time leads to the decrease of 4.3%, 1.7%, 2.5%, and 1.6% probabilities in using public transit, and the absolute magnitudes of these numbers are larger than but the relative magnitudes are similar to the MNL model, in which the corresponding probability decreases are 0.56%, 0.31%, 0.26%, and 0.48%. This difference in the absolute magnitude is already manifested in the previous discussion that the gradients of DNN models are larger than that of MNL models. In addition, the highlighted self-elasticities in the DNNs are overall of a larger magnitude than the cross-elasticity values, which is also reasonable.

Local irregularity is revealed here by the large standard deviations of the elasticities. For

[4]This 6.5% appears much smaller than the values in Figure 3. It is because of the difference between arc and point elasticities.

Table 3: Elasticities of five travel modes with respect to input variables (8K-SGP dataset)

Panel 1: DNN Model	Walk	Public Transit	Ride Hailing	Driving	AV
Walk time	**-5.308(6.9)**	0.399(5.9)	-0.119(7.1)	-0.030(4.6)	-1.360(6.8)
Public transit cost	-1.585(9.6)	**-4.336(9.6)**	-1.648(11.1)	1.081(5.9)	1.292(9.5)
Public transit walk time	0.123(6.9)	**-1.707(6.5)**	0.047(7.3)	0.621(4.7)	0.844(6.7)
public transit wait time	0.985(8.7)	**-2.520(8.9)**	-0.518(9.1)	0.092(5.8)	0.366(8.8)
Public transit in-vehicle time	0.057(9.0)	**-1.608(9.0)**	0.484(9.4)	0.778(5.8)	1.273(8.9)
Ride hail cost	-2.353(7.6)	0.005(6.9)	**-4.498(8.9)**	0.304(5.6)	-0.243(9.0)
Ride hail wait time	0.234(8.8)	1.471(8.3)	**-2.536(10.1)**	-0.253(5.7)	-0.228(8.8)
Ride hail in-vehicle time	0.299(7.8)	-0.224(7.4)	**-5.890(9.4)**	0.740(5.4)	0.739(7.6)
Drive cost	1.124(6.6)	2.545(5.9)	3.760(6.8)	**-1.886(5.0)**	2.273(6.9)
Drive walk time	2.033(5.3)	0.552(5.0)	2.503(5.6)	**-0.412(3.8)**	1.787(5.4)
Drive in-vehicle time	1.824(9.0)	4.163(8.2)	3.640(9.9)	**-3.199(7.4)**	3.268(9.1)
AV cost	-0.562(6.5)	-0.198(6.2)	0.819(6.9)	0.337(4.6)	**-4.289(7.6)**
AV wait time	-0.068(7.9)	-0.695(7.4)	2.400(8.4)	0.284(4.6)	**-1.591(7.8)**
AV in-vehicle time	-0.784(6.2)	0.221(5.6)	0.955(7.1)	0.079(4.3)	**-4.534(6.8)**
Age	-1.003(18.7)	2.502(18.4)	-4.385(20.0)	0.949(13.7)	-1.936(18.6)
Income	1.127(10.7)	0.727(10.5)	0.957(11.9)	-0.002(6.7)	2.539(10.8)
Panel 2: MNL Model	Walk	Public Transit	Ride Hailing	Driving	AV
Walk time	**-1.916(1.8)**	0.130(0.1)	0.130(0.1)	0.130(0.1)	0.130(0.1)
Public transit cost	0.137(0.1)	**-0.566(0.4)**	0.137(0.1)	0.137(0.1)	0.137(0.1)
Public transit access time	0.083(0.1)	**-0.318(0.3)**	0.083(0.1)	0.083(0.1)	0.083(0.1)
Public transit transfer time	0.072(0.1)	**-0.265(0.2)**	0.072(0.1)	0.072(0.1)	0.072(0.1)
Public transit in-vehicle time	0.126(0.1)	**-0.478(0.4)**	0.126(0.1)	0.126(0.1)	0.126(0.1)
Ride hail cost	0.028(0.0)	0.028(0.0)	**-0.248(0.2)**	0.028(0.0)	0.028(0.0)
Ride hail wait time	0.033(0.0)	0.033(0.0)	**-0.304(0.2)**	0.033(0.0)	0.033(0.0)
Ride hail in-vehicle time	0.076(0.1)	0.076(0.1)	**-0.716(0.5)**	0.076(0.1)	0.076(0.1)
Drive cost	0.292(0.2)	0.292(0.2)	0.292(0.2)	**-0.756(1.1)**	0.292(0.2)
Drive walk time	0.120(0.1)	0.120(0.1)	0.120(0.1)	**-0.211(0.3)**	0.120(0.1)
Drive in-vehicle time	0.291(0.2)	0.291(0.2)	0.291(0.2)	**-0.463(0.6)**	0.291(0.2)
AV cost	0.044(0.1)	0.044(0.1)	0.044(0.1)	0.044(0.1)	**-0.413(0.4)**
AV wait time	0.029(0.0)	0.029(0.0)	0.029(0.0)	0.029(0.0)	**-0.254(0.2)**
AV in-vehicle time	0.067(0.1)	0.067(0.1)	0.067(0.1)	0.067(0.1)	**-0.638(0.6)**
Age	-0.168(0.1)	0.546(0.2)	-0.695(0.2)	0.002(0.1)	-0.363(0.2)
Income	-0.102(0.1)	-0.177(0.1)	0.061(0.0)	0.056(0.0)	0.148(0.1)

example in Table 3, as the walking elasticity regarding walking time is -5.3 on average, its standard deviation is 6.9. This large standard deviation is caused by local irregularity, as individuals can have dramatically different elasticity values. The other two challenges, the high sensitivity and the model non-identification, are not presented in the process of computing the average elasticities, because the Opt-DNNs are trained by the same set of hyperparameters and the model non-identification is not seen in only one Opt-DNN.

The difference between the Opt-DNN and the MNL in terms of the coefficient magnitude seems less salient in the 80K-LD dataset, as shown in Table 4. Note that the coefficients of the DNN model on the main diagonal in Panel 1 are mainly negative, which are the same as the findings in Table 3. Interestingly, even the absolute magnitudes of the DNN models become similar to the MNL models. This result reflects the smoother choice probability functions in the large sample compared to the small sample cases, as discussed in Figure 3. It would be very difficult to provide a definitive answer to the question what size can be treated as "large" for DNN models. We will

Table 4: Elasticities of travel modes with respect to input variables (80K-LD dataset)

Panel 1: DNN Model	Walk	Public Transit	Cycle	Driving
Walk time	**-1.494(0.8)**	-0.437(0.9)	-0.926(1.0)	0.680(0.7)
Public transit cost	0.260(0.3)	**-0.199(0.3)**	0.118(0.3)	-0.002(0.2)
Public transit access time	0.310(0.5)	**-0.590(0.6)**	0.012(0.4)	0.239(0.3)
Public transit transfer time	-0.028(0.2)	**-0.118(0.3)**	-0.073(0.2)	0.069(0.2)
Public transit in-vehicle time	0.128(0.3)	**-0.324(0.6)**	0.026(0.4)	0.217(0.4)
Cycle time	-0.134(0.2)	0.376(0.4)	**0.113(0.3)**	-0.167(0.3)
Drive cost	0.070(0.2)	0.039(0.1)	0.025(0.1)	**-0.127(0.3)**
Drive in-vehicle time	0.280(0.5)	1.047(1.0)	0.561(0.8)	**-0.840(0.9)**
Age	-0.236(0.7)	-0.008(0.8)	-0.223(0.8)	0.204(0.8)
Panel 2: MNL Model	Walk	Public Transit	Cycle	Driving
Walk time	**-7.344(7.8)**	0.334(0.4)	0.334(0.4)	0.334(0.4)
Public transit cost	0.049(0.1)	**-0.066(0.1)**	0.049(0.1)	0.049(0.1)
Public transit access time	0.262(0.3)	**-0.459(0.4)**	0.262(0.3)	0.262(0.3)
Public transit transfer time	0.093(0.2)	**-0.110(0.2)**	0.093(0.2)	0.093(0.2)
Public transit in-vehicle time	0.222(0.3)	**-0.261(0.3)**	0.222(0.3)	0.222(0.3)
Cycle time	0.019(0.0)	0.019(0.0)	**-0.757(0.7)**	0.019(0.0)
Drive cost	0.070(0.1)	0.070(0.1)	0.070(0.1)	**-0.235(0.5)**
Drive in-vehicle time	0.811(0.9)	0.811(0.9)	0.811(0.9)	**-1.269(1.7)**
Age	-0.275(0.2)	0.163(0.1)	-0.304(0.2)	-0.004(0.1)

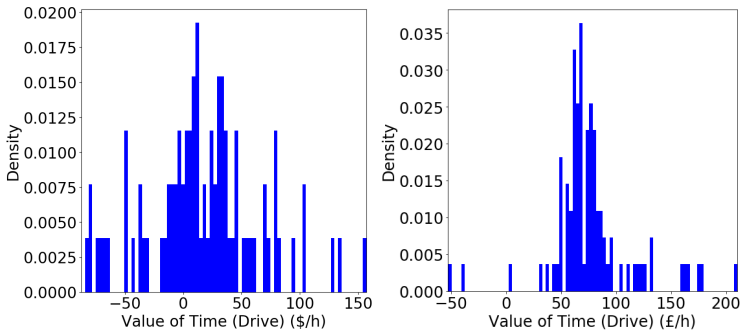
discuss this in the last section, but here we can conclude that the average elasticity coefficients from DNNs are largely intuitive.

5.3.3. Marginal Rates of Substitution: Heterogeneous Values of Time

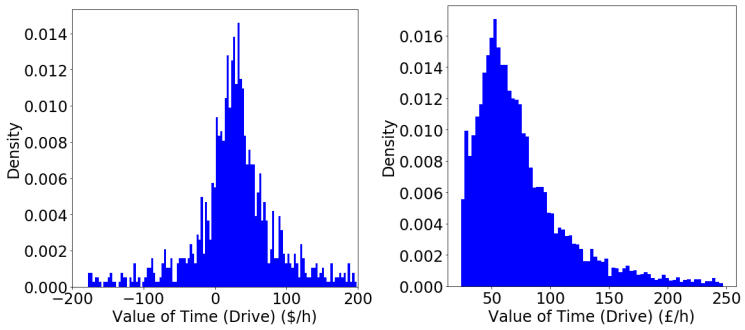
VOT, as one example of MRS, is one of the most important pieces of economic information obtained from choice models, since the monetary gain from time saving is the most prevalent benefit from the improvement of any transportation system. As VOT is computed as the ratio of two parameters in a MNL model, the ratio of two probability derivatives represents the VOT in DNNs. Figures 6a and 6b represent the distribution of the average VOT among all individuals over 100 trainings of Opt-DNNs; Figures 6c and 6d represent the distribution of the heterogeneous VOT of the individuals in one Opt-DNN model. The distributions of VOT in Figures 6c and 6d resemble the typical analysis about the heterogeneous VOT in DCMs.

The aggregate information such as the average values of the VOT distributions are quite reasonable, while certain regions of the VOT distributions can be somewhat counter-intuitive as they may have a very large dispersion and even some negative values. For example, the median VOT in the testing set of the 8K-SGP dataset is $\$27.8/h$, and the VOT distribution is highly concentrated around its mean value, resembling the shape of a Gaussian distribution. Similar patterns can be observed for the 80K-LD dataset in Figures 6d. This finding of reasonable aggregate information but irregular disaggregate information is consistent with our previous findings.

The median $\$27/h$ VOT in Figure 6a is consistent with previous studies, in which VOT has been found to be between $\$7.8/h$ and $\$30.3/h$ for various travel modes [49]. VOT has also been found to be between 21% and 254% of the hourly wage in a review paper [114]. By using the average hourly



(a) Opt-DNNs (8K-SGP) (b) Opt-DNNs (80K-LD)



(c) One Opt-DNN (8K-SGP) (d) One Opt-DNN (80K-LD)

Fig. 6. Heterogeneous values of time; the extremely large and small values (below 5% percentile and above 95% percentile) are cut off from this histogram.

wage ($27.16/h$) of the U.S workers in 2018, we would expect the VOT to be between $5.7/h and $70.0/h. Our VOT obtained from DNNs is about in the middle of this range. Intuitively, the VOT should be of the same magnitude as the hourly wages, and $27/h is very close to the average hourly wage. However, on the other hand, the VOT obtained from DNNs can be unreasonable for certain individuals. It is highly unlikely for VOT to be negative, while DNNs detect a sizeable portion of people whose VOT are negative. This counter-intuitive result is caused by the local irregularity of the probability derivatives. But on the other side, larger sample size can mitigate this problem even at the individual level. In the 80K-LD dataset, the range of VOT is mainly between $40.0/h and $80.0/h and the median value is $69.1/h. Both the range and the median value are reasonable based on the findings of the previous studies.

6. Conclusions and Discussions

This study aims to interpret DNN models in the context of choice analysis and extract economic information as complete as obtained from classical DCMs. The economic information includes a complete list of choice predictions, choice probabilities, market share, substitution patterns of alternatives, social welfare, probability derivatives, elasticity, marginal rates of substitution (MRS), and heterogeneous values of time (VOT). The process of interpreting DNN models is different from

classical DCMs because DNNs are a very flexible model family, capable of automatically learning more flexible behavioral patterns than the regular patterns pre-specified by domain experts in the classical DCMs. As a result, we found that most economic information extracted from DNNs is reasonable and more flexible than the MNL models, particularly when the sample size is large. However, the economic information automatically learnt by DNNs can be sometimes unreliable, caused by three challenges: high sensitivity to hyperparameters, model non-identification, and local irregularity. Owing to the high sensitivity to hyperparameters, the DNN models without appropriate regularizations or architectures cannot provide valuable economic information. Owing to the model non-identification, researchers cannot obtain an ultimate function estimate for economic interpretation. Owing to the local irregularity, DNN models reveal unreasonable local behavioral patterns when the sample size is small. These three problems can be partially addressed by using large sample, simple random hyperparameter searching, model ensemble, regularization, and information aggregation. Particularly, the economic information based on the ensemble method, such as the average choice probability function, average probability derivatives, market shares, average social welfare change, average elasticities, and the median VOT, are mostly consistent with our behavioral intuition and previous studies.

There should be little doubt that DNNs can provide a full set of economic information as DCMs; however, many questions remain. As shown in this study, the 80K-LD dataset can provide more reasonable behavioral patterns than the 8K-LD and 8K-SGP datasets, demonstrating the importance of a large sample size. One challenging question is the exact sample size that can be counted as "large" for DNNs. Unfortunately, it is very difficult to provide a definitive answer, since it always depends on other factors such as model complexity and input dimensions. In principle, as models become more complicated, researchers need larger sample size, as shown in the proof about the estimation error of DNNs in Appendix I. The exact model complexity of DNN models empirically depends on the specific DNN architectures and is also theoretically ambiguous owing to the difficulty of deriving a tight upper bound on the estimation errors. The bottom line, regarding the DNN-based choice models, is that the sample size traditionally counted as large, such as thousands of observations, seem inadequate to provide reliable economic information in DNN-based choice models. This is not a surprise since the model complexity of DNNs is much larger than classical DCMs. While our study suggests that about $O(10^4)$ sample size seems to be adequate for economic information, we would encourage future studies to further explore this question, particularly when more complicated models and inputs, such as images and natural languages, are involved in choice modeling settings.

We also discussed the irregular behavioral patterns of the utility functions in the 8K-LD and 8K-SGP datasets. However, we would like to emphasize that this "irregularity" does not necessarily have a negative connotation. For example, while certain patterns in DNNs can be treated as unreasonable from the perspective of classical MNL models, these patterns exist in other types of behavioral and machine learning literature. As to the exploding gradients, behavioral theory suggests that people can have a sharp threshold price in decision-making, and as a result, the

gradients around the threshold price can be very large. The decision tree model suggests a similar behavioral mechanism: the input gradients around the cutoff points can take very large values. As to the non-monotonicity pattern, people might become more likely to buy certain commodity when the price of the commodity increases, since people treat higher prices as the signal of better quality of the commodities, leading to positive elasticity values of choice probability functions regarding the price variables. In short, the patterns revealed in the 8K-SGP and 8K-LD datasets can be positively evaluated as a success for identifying certain flexible behavioral patterns that cannot be found in a restrictive MNL model, or as a failure that identifies unreasonable and unrealistic behavioral patterns caused by small sample and high model complexity. We do not take a stance here, but leave this question for future studies.

This study has demonstrated the importance of using hyperparameter searching, repeated trainings, regularization methods, and aggregation over models and population to improve the reliability of the economic information. However, it remains an open question what the most effective methods are in terms of making the economic information more reliable from the classical DCM perspective. Recent studies in the ML community have suggested potential methods to address the three challenges. As to the high sensitivity to hyperparameters, potential remedies include a large number of regularization methods, such as domain constraints, Bayesian priors, model ensemble [59], data augmentation [15], dropouts [47], early stopping, and sparse connectivity; new DNN architectures, such as AlexNet [59], GoogleNet [99], and ResNet [43]; or smarter ways to tune hyperparameters, such as construction of a continuous hyperparameter space, Gaussian process, Bayesian neural networks [96, 97], or reinforcement learning [121, 122, 5], much richer than a simple random searching in discrete grids [13, 14]. As to the non-identification challenge, the optimization algorithm has been refined significantly in the past years to the extent that it converges to the simple first order stochastic gradient descent with momentum [57] and specific initialization methods [35, 44]. As to the local irregularity, robust training methods and monotonicity constraints can be used. To formally measure local irregularity, researchers evaluated model performance on adversarial examples [37, 62, 61]. To defend against the adversarial attacks, researchers designed adversarial training methods by incorporating the adversarial attacks into the training process [62], defensive knowledge distillation [79], mini-max robust training [70], and even simple gradient regularization [87]. This study can only open up the discussion about why new methods are necessary for reliable information in DNN-based choice analysis, but a definitive answer needs significant future efforts.

This study interprets DNN-based choice models by focusing on only the economic information, but economic information is not the only valuable information researchers can obtain from DNNs. Our method of computing the input gradients is the same as the gradient-based methods, which are often referred to under different names such as sensitivity analysis, saliency, or attribution maps in computer vision [92, 56, 95, 90, 98, 87], or attention mechanism in natural language processing [113] and generic DNN interpretation literature. But besides gradient-based methods, researchers can also use case-based methods, such as activation maximization (AM), to identify meaningful individual observations to interpret DNN models [31, 92, 72, 56]. Researchers can also interpret

DNN models by mapping the neurons of the hidden layers to the input space [115] or visualizing the activation maps in the last layer [119]. With these approaches, it is possible to extract from DNNs more valuable information beyond the economic information. In short, future researchers should explore all these tools to improve the interpretability of DNN-based choice models and apply the method introduced in this study to a massive number of choice analysis settings. Given the power of DNNs and the infinite opportunities in choice modeling, we believe that the interaction between utility-based choice analysis and DNNs for economic interpretation will be a fertile research area in the future.

7. Acknowledgement

We thank Singapore-MIT Alliance for Research and Technology (SMART) for partially funding this research. We thank Mary Rose Fissinger for her careful proofreading and Jerry Hausman for inspiring Shenhao with an initial idea of this paper in a casual talk.

8. Contributions of Authors

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

References

[1] Agnar Aamodt and Enric Plaza. "Case-based reasoning: Foundational issues, methodological variations, and system approaches". In: *AI communications* 7.1 (1994), pp. 39–59.

[2] Anuradha M Annaswamy et al. "Transactive Control in Smart Cities". In: *Proceedings of the IEEE* 106.4 (2018), pp. 518–537.

[3] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

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

[5] Bowen Baker et al. "Designing neural network architectures using reinforcement learning". In: *arXiv preprint arXiv:1611.02167* (2016).

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

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

[8] Peter L Bartlett et al. "Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks". In: *arXiv preprint arXiv:1703.02930* (2017).

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

[10] Moshe Ben-Akiva, John L Bowman, and Dinesh Gopinath. "Travel demand model system for the information era". In: *Transportation* 23.3 (1996), pp. 241–266.

[11] Moshe Ben-Akiva et al. *Discrete Choice Analysis*. 2014.

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

[13] James S Bergstra et al. "Algorithms for hyper-parameter optimization". In: *Advances in neural information processing systems*. 2011, pp. 2546–2554.

[14] James Bergstra and Yoshua Bengio. "Random search for hyper-parameter optimization". In: *Journal of Machine Learning Research* 13.Feb (2012), pp. 281–305.

[15] Chris M Bishop. "Training with noise is equivalent to Tikhonov regularization". In: *Neural computation* 7.1 (1995), pp. 108–116.

[16] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[17] Stanislav S Borysov, Jeppe Rich, and Francisco C Pereira. "How to generate micro-agents? A deep generative modeling approach to population synthesis". In: *Transportation Research Part C: Emerging Technologies* 106 (2019), pp. 73–97. ISSN: 0968-090X.

[18] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. "Introduction to statistical learning theory". In: *Advanced lectures on machine learning*. Springer, 2004, pp. 169–207.

[19] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[20] Giulio Erberto Cantarella and Stefano de Luca. "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 (2005), pp. 121–155.

[21] Hilmi Berk Celikoglu. "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 (2006), pp. 640–658.

[22] Anna Choromanska et al. "The loss surfaces of multilayer networks". In: *Artificial Intelligence and Statistics*. 2015, pp. 192–204.

[23] Jonathan D Cohen et al. *Measuring time preferences*. Tech. rep. National Bureau of Economic Research, 2016.

[24] Sander van Cranenburgh and Ahmad Alwosheel. "An artificial neural network based approach to investigate travellers' decision rules". In: *Transportation Research Part C: Emerging Technologies* 98 (2019), pp. 152–166.

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

[26] Yann N Dauphin et al. "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization". In: *Advances in neural information processing systems*. 2014, pp. 2933–2941.

[27] Juan De Dios Ortuzar and Luis G Willumsen. *Modelling transport*. John Wiley and Sons, 2011.

[28] Loan NN Do et al. "An effective spatial-temporal attention based neural network for traffic flow prediction". In: *Transportation research part C: emerging technologies* 108 (2019), pp. 12–28. ISSN: 0968-090X.

[29] Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning". In: (2017).

[30] Yanjie Duan et al. "An efficient realization of deep learning for traffic data imputation". In: *Transportation research part C: emerging technologies* 72 (2016), pp. 168–181.

[31] Dumitru Erhan et al. "Visualizing higher-layer features of a deep network". In: *University of Montreal* 1341.3 (2009), p. 1.

[32] Stefan Falkner, Aaron Klein, and Frank Hutter. "BOHB: Robust and efficient hyperparameter optimization at scale". In: *arXiv preprint arXiv:1807.01774* (2018).

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

[34] Aurélien Géron. *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems.* " O'Reilly Media, Inc.", 2017.

[35] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 2010, pp. 249–256.

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

[37] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples". In: *arXiv preprint arXiv:1412.6572* (2015).

[38] Ian Goodfellow et al. *Deep learning*. Vol. 1. MIT press Cambridge, 2016.

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

[40] Lars Kai Hansen and Peter Salamon. "Neural network ensembles". In: *IEEE transactions on pattern analysis and machine intelligence* 12.10 (1990), pp. 993–1001. ISSN: 0162-8828.

[41] Siyu Hao, Der-Horng Lee, and De Zhao. "Sequence to sequence learning with attention mechanism for short-term passenger flow prediction in large-scale metro system". In: *Transportation Research Part C: Emerging Technologies* 107 (2019), pp. 287–300. ISSN: 0968-090X.

[42] David Haussler and Philip M Long. "A generalization of Sauer's lemma". In: *Journal of Combinatorial Theory, Series A* 71.2 (1995), pp. 219–240.

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

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

[45] John Paul Helveston et al. "Will subsidies drive electric vehicle adoption? Measuring consumer preferences in the US and China". In: *Transportation Research Part A: Policy and Practice* 73 (2015), pp. 96–112.

[46] Tim Hillel, Mohammed ZEB Elshafie, and Ying Jin. "Recreating passenger mode choice-sets for transport simulation: A case study of London, UK". In: *Proceedings of the Institution of Civil Engineers-Smart Infrastructure and Construction* 171.1 (2018), pp. 29–42. ISSN: 2397-8759.

[47] Geoffrey E Hinton et al. "Improving neural networks by preventing co-adaptation of feature detectors". In: *arXiv preprint arXiv:1207.0580* (2012).

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

[49] Chinh Q Ho et al. "Vehicle value of travel time savings: Evidence from a group-based modelling approach". In: *Transportation Research Part A: Policy and Practice* 88 (2016), pp. 134–150.

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

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

[52] Xiuling Huang, Jie Sun, and Jian Sun. "A car-following model considering asymmetric driving behavior based on long short-term memory neural networks". In: *Transportation Research Part C: Emerging Technologies* 95 (2018), pp. 346–362.

[53] Naomi Irvine et al. "Neural Network Ensembles for Sensor-Based Human Activity Recognition Within Smart Environments". In: *Sensors* 20.1 (2020), p. 216.

[54] Patiphan Kaewwichian, Ladda Tanwanichkul, and Jumrus Pitaksringkarn. "Car Ownership Demand Modeling Using Machine Learning: Decision Trees and Neural Networks." In: *International Journal of Geomate* 17.62 (2019), pp. 219–230.

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

[56] Been Kim and Finale Doshi-Velez. "Interpretable Machine Learning (ICML Tutorials)". In: *International Conference of Machine Learning*. Sydney, 2017.

[57] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

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

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

[60] Anders Krogh and Jesper Vedelsby. "Neural network ensembles, cross validation, and active learning". In: *Advances in neural information processing systems*. 1995, pp. 231–238.

[61] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world". In: *arXiv preprint arXiv:1607.02533* (2017).

[62] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. "Adversarial machine learning at scale". In: *arXiv preprint arXiv:1611.01236* (2016).

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

[64] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science Business Media, 2013.

[65] Seunghyeon Lee et al. "An advanced deep learning approach to real-time estimation of lane-based queue lengths at a signalized junction". In: *Transportation research part C: emerging technologies* 109 (2019), pp. 117–136. ISSN: 0968-090X.

[66] Lisha Li et al. "Hyperband: A novel bandit-based approach to hyperparameter optimization". In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 6765–6816.

[67] Zachary C Lipton. "The mythos of model interpretability". In: *arXiv preprint arXiv:1606.03490* (2016).

[68] Lijuan Liu and Rung-Ching Chen. "A novel passenger flow prediction model using deep learning methods". In: *Transportation Research Part C: Emerging Technologies* 84 (2017), pp. 74–91.

[69] Tao Ma, Constantinos Antoniou, and Tomer Toledo. "Hybrid machine learning algorithm and statistical time series model for network-wide traffic forecast". In: *Transportation Research Part C: Emerging Technologies* 111 (2020), pp. 352–372. ISSN: 0968-090X.

[70] Aleksander Madry et al. "Towards deep learning models resistant to adversarial attacks". In: *arXiv preprint arXiv:1706.06083* (2017).

[71] Daniel McFadden. "Conditional logit analysis of qualitative choice behavior". In: (1974).

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

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

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

[75] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. "Norm-based capacity control in neural networks". In: *Conference on Learning Theory*. 2015, pp. 1376–1401.

[76] Anh Nguyen, Jason Yosinski, and Jeff Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 427–436.

[77] Peter Nijkamp, Aura Reggiani, and Tommaso Tritapepe. "Modelling inter-urban transport flows in Italy: A comparison between neural network analysis and logit analysis". In: *Transportation Research Part C: Emerging Technologies* 4.6 (1996), pp. 323–338.

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

[79] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples". In: *arXiv preprint arXiv:1605.07277* (2016).

[80] Miguel Paredes et al. "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, 2017, pp. 780–785.

[81] Tomaso Poggio et al. "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 (2017), pp. 503–519.

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

[83] Sarada Pulugurta, Ashutosh Arun, and Madhu Errampalli. "Use of artificial intelligence for mode choice analysis and comparison with traditional multinomial logit model". In: *Procedia-Social and Behavioral Sciences* 104 (2013), pp. 583–592.

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

[85] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "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, 2016, pp. 1135–1144.

[86] David Rolnick and Max Tegmark. "The power of deeper networks for expressing natural functions". In: *arXiv preprint arXiv:1705.05502* (2017).

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

[88] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. "Right for the right reasons: Training differentiable models by constraining their explanations". In: *arXiv preprint arXiv:1703.03717* (2017).

[89] Ch Ravi Sekhar and E Madhu. "Mode Choice Analysis Using Random Forrest Decision Trees". In: *Transportation Research Procedia* 17 (2016), pp. 644–652.

[90] Ramprasaath R Selvaraju et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 618–626.

[91] Toru Seo et al. "Interactive online machine learning approach for activity-travel survey". In: *Transportation Research Part B: Methodological* (2017).

[92] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034* (2013).

[93] Kenneth A Small, Erik T Verhoef, and Robin Lindsey. "Travel Demand". In: *The economics of urban transportation*. Vol. 2. Routledge, 2007.

[94] Kenneth Small and Clifford Winston. "The demand for transportation: models and applications". In: *Essays in Transportation Economics and Policy*. 1998.

[95] Daniel Smilkov et al. "Smoothgrad: removing noise by adding noise". In: *arXiv preprint arXiv:1706.03825* (2017).

[96] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. "Practical bayesian optimization of machine learning algorithms". In: *Advances in neural information processing systems*. 2012, pp. 2951–2959.

[97] Jasper Snoek et al. "Scalable bayesian optimization using deep neural networks". In: *International Conference on Machine Learning*. 2015, pp. 2171–2180.

[98] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3319–3328.

[99] Christian Szegedy et al. "Going deeper with convolutions". In: Cvpr, 2015.

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

[101] Kenneth E Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

[102] Chih-Fong Tsai and Jhen-Wei Wu. "Using neural network ensembles for bankruptcy prediction and credit scoring". In: *Expert systems with applications* 34.4 (2008), pp. 2639–2649. ISSN: 0957-4174.

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

[104] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge University Press, 2018.

[105] Pascal Vincent et al. "Extracting and composing robust features with denoising autoencoders". In: *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1096–1103.

[106] Ulrike Von Luxburg and Bernhard Schölkopf. "Statistical learning theory: Models, concepts, and results". In: *Handbook of the History of Logic*. Vol. 10. Elsevier, 2011, pp. 651–706.

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

[108] Xin Wu et al. "Hierarchical travel demand estimation using multiple data sources: A forward and backward propagation algorithmic framework on a layered computational graph". In: *Transportation Research Part C: Emerging Technologies* 96 (2018), pp. 321–346. ISSN: 0968-090X.

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

[110] Guangnian Xiao, Zhicai Juan, and Chunqin Zhang. "Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization". In: *Transportation Research Part C: Emerging Technologies* 71 (2016), pp. 447–463.

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

[112] Shuguan Yang et al. "A deep learning approach to real-time parking occupancy prediction in transportation networks incorporating multiple spatio-temporal data sources". In: *Transportation Research Part C: Emerging Technologies* 107 (2019), pp. 248–265. ISSN: 0968-090X.

[113] Wenpeng Yin et al. "Abcnn: Attention-based convolutional neural network for modeling sentence pairs". In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 259–272.

[114] Luca Zamparini and Aura Reggiani. "The value of travel time in passenger and freight transport: an overview". In: *Policy analysis of transport networks*. Routledge, 2016, pp. 161–178.

[115] Matthew D Zeiler and Rob Fergus. "Visualizing and understanding convolutional networks". In: *European conference on computer vision*. Springer, 2014, pp. 818–833.

[116] Chiyuan Zhang et al. "Understanding deep learning requires rethinking generalization". In: *arXiv preprint arXiv:1611.03530* (2016).

[117] Junbo Zhang et al. "Predicting citywide crowd flows using deep spatio-temporal residual networks". In: *Artificial Intelligence* 259 (2018), pp. 147–166. ISSN: 0004-3702.

[118] Zhenhua Zhang et al. "A deep learning approach for detecting traffic accidents from social media data". In: *Transportation research part C: emerging technologies* 86 (2018), pp. 580–596.

[119] Bolei Zhou et al. "Learning deep features for discriminative localization". In: *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, 2016, pp. 2921–2929.

[120] Bolei Zhou et al. "Object detectors emerge in deep scene cnns". In: *arXiv preprint arXiv:1412.6856* (2014).

[121] Barret Zoph and Quoc V Le. "Neural architecture search with reinforcement learning". In: *arXiv preprint arXiv:1611.01578* (2016).

[122] Barret Zoph et al. "Learning transferable architectures for scalable image recognition". In: *arXiv preprint arXiv:1707.07012* 2.6 (2017).

Appendix I: Large Estimation Error of DNNs

Definition 1. *Excess error of $\hat{f}$ is defined as*

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

$L(\hat{f})$ is the population error of the estimator; $L(f^*)$ is the population error of the true model; $L = \mathbb{E}_{x,y}[l(y, f(x))]$ and $l(y, f(x))$ is the loss function. Excess error measures to what extent the error of the estimator deviates from the true model, averaged over random sampling S. Note that the excess error can be decomposed as following.

$$\mathbb{E}_S[L(\hat{f}) - L(f^*)] = \mathbb{E}_S[L(\hat{f}) - L(f_F) + L(f_F) - L(f^*)] \tag{8}$$

in which $\mathbb{E}_S[L(\hat{f}) - L(f_F)]$ represents the estimation error and $\mathbb{E}_S[L(f_F) - L(f^*)]$ represents the approximation error. When the model family F is large enough, the approximation error becomes very small. For the simplicity of our discussion, we assume the approximation error of DNNs equals to zero. As a result, the following discussions use the terms of excess error and estimation error in an interchangable way.

Proposition 1. *The estimation error of $\hat{f}$ can be bounded by VC dimension*

$$\mathbb{E}_S[L_{0/1}(\hat{f}) - L_{0/1}(f^*)] \lesssim O(\frac{v}{N}) \tag{9}$$

in which v is the VC dimension of function class $\mathcal{F}$; N is the sample size; $L_{0/1}$ is the binary prediction error.

Proof. When no misspecification error exists, estimation error can be further decomposed as three terms

$$\begin{aligned}
\mathbb{E}_S[L(\hat{f}) - L(f^*)] &= \mathbb{E}_S[L(\hat{f}) - \hat{L}(\hat{f}) + \hat{L}(\hat{f}) - \hat{L}(f^*) + \hat{L}(f^*) - L(f^*)] && (10)\\
&\leq \mathbb{E}_S[L(\hat{f}) - \hat{L}(\hat{f})] && (11)\\
&\leq \mathbb{E}_S \sup_{f \in \mathcal{F}} [L(f) - \hat{L}(f)] && (12)
\end{aligned}$$

in which $\hat{L}(f) := \frac{1}{N}\sum_i l(y_i, f^*(x_i))$; the first inequality holds because $\mathbb{E}_S[\hat{L}(\hat{f}) - \hat{L}(f^*)] \leq 0$ based on the definition of $\hat{f} := \operatorname{argmin} \hat{L}(f)$ and $\mathbb{E}_S[\hat{L}(f^*) - L(f^*)] = 0$ based on the law of large numbers; the second inequality holds due to the sup operator.

Equation 12 can be bounded.

$$\mathbb{E}_S \sup_{f \in \mathcal{F}} [L(f) - \hat{L}(f)] \leq 2\mathbb{E}_{S,\epsilon} \sup_f \frac{1}{N} \sum_i l(f(x_i), y_i)\epsilon_i \tag{13}$$

This proof relies on the technique called symmetrization, as shown in the proof of Theorem 4.10 in [107]. Note that for prediction error, the loss function $l(f(x_i), y_i) = \mathbb{1}\{f(x_i) \neq y_i\} = y_i + (1 -$

$2y_i)f(x_i)$, as $y_i \in \{0,1\}$ and $f(x_i) \in \{0,1\}$. By applying contraction inequality to Equation 13,

$$2\mathbb{E}_{S,\epsilon} \sup_f \frac{1}{N} \sum_i l(f(x_i), y_i)\epsilon_i = 2\mathbb{E}_{S,\epsilon} \sup_f \frac{1}{N} \sum_i (y_i + (1-2y_i)f(x_i)) \times \epsilon_i \tag{14}$$

$$\leq 2\mathbb{E}_{S,\epsilon} \sup_f \frac{1}{N} \sum_i f(x_i)\epsilon_i \tag{15}$$

$$= 2\mathbb{E}_S \hat{\mathcal{R}}_N(\mathcal{F}\,|_S) \tag{16}$$

in which the second line uses the contraction inequality [64] and the third uses the definition of Rademacher complexity. Basically the question about the upper bound of estimation error is turned to the question about the complexity of function class of DNN $\mathcal{F}$. There are many ways to derive an upper bound on Rademacher complexity [7]. To obtain the v/N result, Dudley integral and chaining techniques are useful. Let $Z_f := \frac{1}{\sqrt{N}} \sum_i \epsilon_i f(x_i)$ and $Z_g := \frac{1}{\sqrt{N}} \sum_i \epsilon_i g(x_i)$, in which $f, g \in \mathcal{F}$. Based on Theorem 5.22 Dudley's entropy integral bound in [107],

$$\mathbb{E}_S \left[\sup_{f\in\mathcal{F}} Z_f\right] \leq \mathbb{E}_S \left[\sup_{f,g\in\mathcal{F}} Z_f - Z_g\right] \tag{17}$$

$$\leq 2\mathbb{E}_S\Big[\sup_{f',g'\in\mathcal{F};\rho_x(f',g')\leq\delta} Z_{f'} - Z_{g'}\Big] + 32\int_{\delta/4}^{D} \sqrt{\log N_x(u;\mathcal{F})}du \tag{18}$$

in which $\rho_x^2(f',g') = \frac{1}{N}\sum_{i=1}^N (f'(x_i) - g'(x_i))^2$; f' and g' are the components around the δ distance of one element in the δ cover of function class $\mathcal{F}$; D is the diameter of the function class $\mathcal{F}$ projected to dataset S, defined as $D := \sup_{f,g\in\mathcal{F}} \rho_x(f,g) \leq 1$; δ is any positive value in $[0, D]$. Equation 18 holds for any δ. The first term in Equation 18 measures the local complexity of DNN and the second term measures the error caused by discretization of the function space. The two terms could be bounded separately. For the first term,

$$\mathbb{E}_S\Big[\sup_{f',g'\in\mathcal{F};\rho_x(f',g')\leq\delta} Z_{f'} - Z_{g'}\Big] = \mathbb{E}_S\Big[\sup_{\rho_x(f',g')\leq\delta} \frac{1}{\sqrt{N}} \sum_i \epsilon_i(f'(x_i) - g'(x_i))\Big] \tag{19}$$

$$= \delta\mathbb{E}_S||\epsilon||_2 \tag{20}$$

$$\leq \delta\sqrt{\mathbb{E}\sum_i \epsilon_i^2} \tag{21}$$

$$\leq \delta\sqrt{N} \tag{22}$$

in which the second line uses the dual norm; the third line uses the fact that ϵ_i is a 1 sub-Gaussian random variable. For the second term in Equation 18, we need to use the Haussler fact [42] that

$$N_x(u;\mathcal{F}) \leq Cv(16e)^v(\frac{1}{u})^v$$

It implies

$$32\int_{\delta/4}^{D}\sqrt{\log N_x(u;\mathcal{F})}du \leq 32\int_{\delta/4}^{D}\sqrt{\log\left[Cv(16e)^v(\frac{1}{u})^v\right]}du \tag{23}$$

$$= 32\int_{\delta/4}^{D}\sqrt{\log C + \log v + v\log 16e + v\log\frac{1}{u}}\ du \tag{24}$$

$$\leq c_0\sqrt{v}\int_{\delta/4}^{D}\sqrt{\log\frac{1}{u}}\ du \tag{25}$$

$$\leq c_0\sqrt{v}\int_{0}^{D}\sqrt{\log\frac{1}{u}}\ du \tag{26}$$

$$\leq c_0'\sqrt{v} \tag{27}$$

By plugging in the upper bounds on the two terms back to Equation 18 and dividing both side by $\sqrt{N}$, it implies

$$\mathbb{E}_S\sup_{f\in\mathcal{F}}\frac{1}{N}\sum_i\epsilon_i f(x_i) \leq \inf_{\delta}\left[\delta + c_0'\sqrt{\frac{v}{N}}\right] \tag{28}$$

$$= c_0'\sqrt{\frac{v}{N}} \tag{29}$$

Therefore, the estimation error can be bounded:

$$\mathbb{E}_S[L(\hat{f}) - L(f^*)] \lesssim O(\sqrt{\frac{v}{N}}) \tag{30}$$

Remarks. Intuitively, v/N describes the tradeoff between model complexity and sample size. In a typical MNL model, v is of the same scale as the number of parameters and the input dimension d; on the contrary, DNN is a much more complex nonlinear model with much larger v. As proved by Bartlett (2017) [8], DNN with W denoting the number of weights and L denoting the depth has VC dimension $O(WL\log(W))$. For instance, when a dataset has 25 input variables, the VC dimension of a simple DNN with 8 layers and 100 neurons as its width is about $320,000$, as opposed to $v = 25$ as the VC dimension of MNL. Theorefore, the theoretical upper bound of DNN on its estimation error is much larger than MNL model.

Statistical learning theory is a very broad field that can be used to prove the upper bound on the estimation error [104, 107]. Proposition 1 is limited to the binary discrete output, although its extension to multiple classes and continuous output is also possible. The theoretically optimum upper bound on DNN's estimation error is still an ongoing research field. Statisticians have been exploring different methods to bound DNN, and the methods based on empirical process theory and the contraction inequaility could provide the tightest upper bound so far [36, 75, 7, 64]. The proof of tighter bounds based on contraction inequality also relies on the connection between different loss functions, the techniques of margin analysis and surrogate losses [6]. These proofs are beyond the scope of this study.

Appendix II: Hyperparameter Space

Table 5: Hyperparameter space

Depth	$[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$
Width	$[25, 50, 100, 150, 200]$
L_1 penalty constants	$[0.1, 10^{-2}, 10^{-3}, 10^{-5}, 10^{-10}, 10^{-20}]$
L_2 penalty constants	$[0.1, 10^{-2}, 10^{-3}, 10^{-5}, 10^{-10}, 10^{-20}]$
Dropout rates	$[0.01, 10^{-5}]$

Appendix III: Summary Statistics of Two Datasets

The key statistics of our samples are summarized in Tables 6 and 7.

Table 6: Summary statistics of the SGP dataset

Panel 1. Continuous Variables							
	mean	std	min	25%	50%	75%	max
Walk_walktime (min)	60.504	54.875	2.0	28.00	40.0	75.00	630.0
Bus_cost (S$)	2.069	1.266	0.0	1.12	1.8	2.52	7.0
Bus_walktime (min)	11.964	10.782	0.0	4.20	8.0	15.00	84.0
Bus_waittime (min)	7.732	5.033	0.0	4.00	7.0	10.00	42.0
Bus_ivt (min)	25.064	18.911	0.0	10.00	21.0	31.20	168.0
Ridesharing_cost (S$)	14.485	11.636	0.0	7.00	12.0	17.60	140.0
Ridesharing_waittime (min)	7.108	4.803	0.0	4.00	5.6	9.00	42.0
Ridesharing_ivt (min)	18.283	13.389	0.8	9.80	15.4	23.20	147.0
AV_cost (S$)	16.076	14.598	0.0	7.70	12.1	18.70	180.0
AV_waittime (min)	7.249	5.675	0.0	3.00	6.0	8.00	48.0
AV_ivt (min)	20.115	16.989	0.6	9.00	16.2	25.20	189.0
Drive_cost (S$)	10.494	10.568	0.0	3.20	7.0	16.00	70.0
Drive_walktime (min)	3.968	4.176	0.0	1.40	2.8	4.80	42.0
Drive_ivt (min)	17.430	14.101	0.8	8.00	14.4	22.40	168.0
Age (year)	41.349	12.478	18.0	31.00	41.0	50.00	82.0
Income (K S$)	9.827	5.013	0.0	7.00	9.0	13.50	20.0
Education	3.063	2.698	0.0	0.0	4.0	5.00	7.0
Panel 2. Discrete Variables (Counts)							
Gender	5,190 (1: Male); 3,228 (0: Female)						
Employment	5,064 (1: Employed); 3,354 (0: Unemployed)						

Appendix IV: Formula to Compute Elasticities

This appendix shows the formula used to compute elasticities. There are four types of formula, depending on the type of variable and utility functions. Recall that the utility functions are:

Table 7: Summary statistics of the LD dataset

Panel 1. Continuous Variables							
	mean	std	min	25%	50%	75%	max
Age (Year)	39.462	19.227	5.000	25.000	38.000	52.000	99.000
Distance (Meters)	4605	4782	77	1309	2814	6175	40941
Duration_walking (h)	1.129	1.118	0.025	0.351	0.723	1.514	9.278
Duration_cycling (h)	0.362	0.352	0.006	0.117	0.232	0.485	3.052
Duration_PT_access (h)	0.160	0.092	0.000	0.092	0.144	0.211	1.189
Duration_PT_in_vehicle (h)	0.262	0.230	0.000	0.083	0.192	0.383	2.147
Duration_pt_transfer_total (h)	0.044	0.078	0.000	0.000	0.000	0.083	0.865
Duration_driving (h)	0.282	0.252	0.000	0.108	0.192	0.369	2.061
Cost_transit (£)	1.563	1.535	0.000	0.000	1.500	2.400	13.490
Cost_driving_total (£)	1.903	3.485	0.000	0.290	0.570	1.290	17.160
Panel 2. Discrete Variables (Counts)							
Gender	42,690 (0: Female); 38,396 (1: Male)						
Driving License	31,051 (0: No); 50,035 (1: Yes)						
Car Ownership	23707 (0 Car); 35,229 (1 Car); 22,150 (2 Cars)						
Number of Transfers in PT	56,668 (0); 19,765 (1); 4,428 (2); 495 (3); 30 (4)						

$$V_{ik} = \beta_{0,k} + \beta_{x,k}^T x_{ik} + \beta_{z,k}^T z_i, \quad as \quad k \neq ref \tag{31}$$

$$V_{ik} = \beta_{x,k}^T x_{ik}, \quad as \quad k = ref \tag{32}$$

For simplicity, the following derivation omits the subscript i, uses x_k as a scalar to represent an alternative-specific variable, and uses z as a scalar to represent an individual-specific variable.

1. Self-elasticity of choice probability s_k with respect to an alternative-specific variable x_k. This formula can be found in Train (2009) [101].

$$\frac{\partial s_k / s_k}{\partial x_k / x_k} = \frac{\partial s_k}{\partial x_k} \times \frac{x_k}{s_k} \tag{33}$$

$$= \frac{\partial V_k}{\partial x_k} s_k(1 - s_k) \times \frac{x_k}{s_k} \tag{34}$$

$$= \beta_{x,k} x_k (1 - s_k) \tag{35}$$

2. Cross-elasticity of choice probability s_k with respect to an alternative-specific variable $x_{k'}$. This formula can also be found in Train (2009) [101]. Interestingly, the formula does not depend on the alternative k, so conditioning on the same k', the cross-elasticities are the

same.

$$\frac{\partial s_k/s_k}{\partial x_{k'}/x_{k'}} = \frac{\partial s_k}{\partial x_{k'}} \times \frac{x_{k'}}{s_k} \tag{36}$$

$$= -\beta_{x,k'} x_{k'} s_{k'} \tag{37}$$

3. Elasticity of choice probability s_k $(k = ref)$ with respect to individual-specific variable z_i. This formula cannot be found in standard textbooks [9, 101], so we show slightly more detailed derivations.

$$\frac{\partial s_k/s_k}{\partial z/z} = [\frac{\partial e^{V_k}}{\partial z} \times \frac{1}{\sum_j e^{V_j}} - e^{V_k}(\sum_j e^{V_j})^{-2} \frac{\partial \sum_j e^{V_j}}{\partial z}] \times \frac{z}{s_k} \tag{38}$$

$$= -\frac{e^{V_k}}{(\sum_j e^{V_j})^2} \times \sum_{j \neq k} e^{V_j} \beta_{z,j} \tag{39}$$

$$= -z \sum_{j \neq k} s_j \beta_{z,j} \tag{40}$$

4. Elasticity of choice probability s'_k $(k' \neq ref)$ with respect to individual-specific variable z_i. Here we use k to denote the reference alternative.

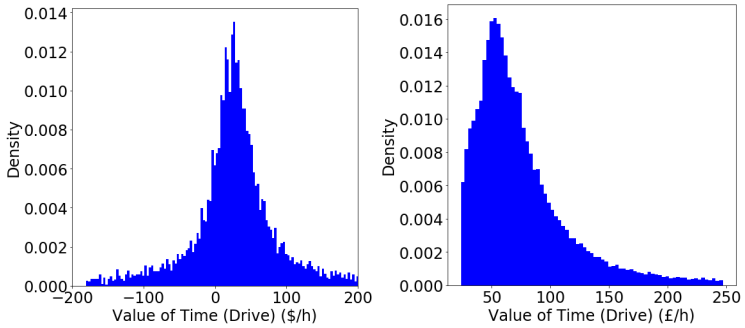
$$\frac{\partial s_{k'}/s_{k'}}{\partial z/z} = [\frac{\partial e^{V_{k'}}}{\partial z} \times \frac{1}{\sum_j e^{V_j}} - e^{V_{k'}}(\sum_j e^{V_j})^{-2} \frac{\partial \sum_j e^{V_j}}{\partial z}] \times \frac{z}{s_{k'}} \tag{41}$$

$$= [\frac{e^{V_{k'}}}{\sum_j e^{V_j}} \beta_{z,k'} - \frac{e^{V_{k'}}}{(\sum_j e^{V_j})^2} \times \sum_{j \neq k} \frac{\partial e^{V_j}}{\partial x}] \times \frac{z}{s_{k'}} \tag{42}$$

$$= z\beta_{z,k'} - z \sum_{j \neq k} s_j \beta_{z,j} \tag{43}$$

Appendix V: Further Details of Experimental Results

Figure 7 shows the distributions of the VOT in the training sets of the 8K-SGP and 80K-LD datasets. The distributions in the training sets are very similar to those in the testing sets, suggesting that our previous findings are robust.



(a) One Opt-DNN (8K-SGP; Training)

(b) One Opt-DNN (80K-LD; Training)

Fig. 7. Heterogeneous values of time in the training sets; the extremely large and small values are cut-off from this histogram.