Deep neural networks for choice analysis: Enhancing behavioral regularity with gradient regularization

Siqi Feng[a,b], Rui Yao[c], Stephane Hess[d], Ricardo Daziano[b], Timothy Brathwaite[a], Joan Walker[a], and Shenhao Wang[*e,f]

[a]Department of Civil and Environmental Engineering, University of California, Berkeley
[b]School of Civil and Environmental Engineering, Cornell University
[c]Department of Technology, Management and Economics, Technical University of Denmark
[d]Institute for Transport Studies, University of Leeds
[e]Department of Urban and Regional Planning, University of Florida
[f]Department of Urban Studies and Planning, Massachusetts Institute of Technology

Abstract

Deep neural networks (DNNs) have been increasingly applied to travel demand modeling because of their automatic feature learning, high predictive performance, and economic interpretability. However, DNNs frequently present behaviorally irregular patterns, significantly limiting their practical potentials and theoretical validity in behavioral modeling. This study proposes strong and weak behavioral regularities as novel metrics to evaluate the monotonicity of the demand functions (a.k.a. "law of demand"), and further designs a constrained optimization framework with six gradient regularization methods to enhance such regularities for DNNs. The empirical benefits of the approach are illustrated by applying these gradient regularizers to the Chicago travel survey data, which enables us to examine the trade-off between prediction quality and behavioral regularity for large versus small sample scenarios and in-domain versus out-of-domain generalizations. The results demonstrate that, unlike models with strong behavioural foundations such as the multinomial logit, the benchmark DNNs cannot guarantee behavioral regularity. However, after applying gradient regularization, we significantly increase DNNs' behavioral regularity by around 30 percentage points while retaining their relatively high predictive power. In the small sample scenario, gradient regularization is more effective than in the large sample scenario, simultaneously improving behavioral regularity by nearly 50 percentage points and prediction accuracy by around three percentage points. Comparing to the in-domain generalizability, gradient regularization can more effectively enhance DNNs' out-of-domain generalizability because it improves behavioral regularity by nearly 40 percentage points, indicating the criticality of behavioral regularization for enhancing model transferability and application in forecasting. Future studies could use behavioral regularity as a metric along with log-likelihood and prediction accuracy in evaluating travel demand models, and investigate other methods to further enhance behavioral regularity when adopting complex machine learning models.

Keywords: travel demand, deep learning, behavioral regularization

*Corresponding author. E-mail: shenhaowang@ufl.edu.

1 Introduction

Deep neural networks (DNNs) have revolutionized fields such as computer vision and natural language processing, that in turn support technologies such as self-driving cars and large language models (van Dis et al., 2023; LeCun et al., 2015). DNNs have also been applied to economics (Zheng et al., 2023), including interpreting and predicting individual choice behavior (Wang et al., 2020a,b). It is in this latter area that DNNs offer a contrast with the conventionally used discrete choice models (DCMs), which are typically based on random utility maximization (Ben-Akiva and Lerman, 1985) and the assumption that travelers choose the alternative with the highest random utility in the choice set. One drawback of this traditional modeling paradigm is the time-consuming trial-and-error process for an "optimal" specification of the model, in particular the utility function (van Cranenburgh et al., 2022), which represents economic preferences. Additionally, the decisions made in this process are often subjective. By contrast, DNNs are capable of automated feature learning, i.e., the specification of a DNN-based choice model is automatically learned from the input data, which avoids the aforementioned specification search process, and reduces the level of subjectivity. The high prediction accuracy of DNNs is a result of their complex model structure, which helps capture intricate behavioral relationships and can provide new insights beyond those of conventional DCMs.

Traditional choice modelers often see DNNs as "black-box" models, although DNNs actually contain complete economic information for choice analysis (Wang et al., 2020b). However, existing DNNs often exhibit behaviorally irregular patterns because the demand functions in DNNs are not guaranteed to decrease monotonically with generalized costs. The "law of demand" in economics indicates an inverse relationship between generalized costs and the aggregate demand. While DCMs such as random utility models (RUMs), do not impose specific directionalities a priori, the specification search conducted by an analyst will not accept models that lead to counter-intuitive results. With DNNs, the analyst has less control, and non-monotonic patterns have been detected empirically in DNNs' predictions, even with model ensembles (Wang et al., 2020a,b; Xia et al., 2023). This fact might be a drawback of the nonlinear structure of DNNs, making them flexible to fit data but difficult to restrict the gradient's direction with limited data samples. The issue often deteriorates in case of out of sample application, i.e., applying a trained DNN to a testing set with unseen distributions (Quiñonero-Candela et al., 2008). In fact, the out-of-domain generalizability of DNNs has attracted rising interests in several computer science fields, including domain adaption (Wang and Deng, 2018) and transfer learning (Pan and Yang, 2009).

To improve the monotonicity of DNNs, we propose to regularize the loss function in training, which has been shown in computer science to enhance the robustness of DNNs (Lyu et al., 2015; Ross and Doshi-Velez, 2018). However, this approach has rarely been considered in previous DNN-based choice models, or used for enhancing behavioral regularity (Wang et al., 2020b; Zheng et al., 2021). In this paper, we address the issue of behavioral irregularity by first defining metrics based on monotonicity of the demand functions, and further designing and implementing a constrained optimization framework that regularizes the input gradient in order to explicitly constrain the gradient's direction. We then design experiments to examine the performance of behaviorally regularized DNNs in terms of behavioral regularity and predictive performance, differentiating the in-domain versus out-of-domain generalization. We also consider factors such as sample size, which is valuable for practical modeling because large samples are costly for travel surveys. The multinomial logit (MNL) model is chosen as a benchmark model due to its concise expression and perfect regularity, while our regularization framework is general. The results show that by using appropriate regularization, DNNs can achieve high regularity without sacrificing their prediction quality, which makes them competitive in real-world applications.

The rest of this paper is organized as follows. Section 2 briefly reviews the literature about the behavioral irregularity issue of DNNs with possible solutions. Section 3 introduces the theory, formulates the problem, and develops a solution framework based on gradient regularization. Section 4 sets up the mode choice experiment, while Section 5 illustrates and analyzes the empirical results. Finally, Section 6 concludes the study and looks ahead to future research. To facilitate future research, we uploaded this work to the following

GitHub repository: https://github.com/siqi-feng/DNN-behavioral-regularity.

2 Literature review

The economic choice behavior of humans generally follows the "law of demand" in economics, which states the inverse relationship between price and quantity demanded. This law leads to a monotonic change in market demand due to the change in consumers' purchasing power, including price and income changes (Chiappori, 1985; Härdle et al., 1991; Hildenbrand, 1983; Quah, 2000). The transportation field has also observed the negative influence of travel costs on travel demand (McFadden, 1974; Souche, 2010; Yao and Morikawa, 2005), based on which demand management policies such as road pricing (May, 1992; Yang and Bell, 1997) were developed. Although such market rationality is widely recognized, individual choice behavior might be irrational (Becker, 1962; Knez et al., 1985). Lichtenstein and Slovic (1971) studied preference reversal in decision making, which is a typical counterexample of individual rationality. Studies in bounded rationality theory (Simon, 1957; Di and Liu, 2016; Watling et al., 2018) and prospect theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992) also relax the strict monotonicity assumption in modeling demand.

The aforementioned law is generally followed by the design of random utility models. In the MNL model, for example, an increase in the travel cost of an alternative would be expected to decrease its deterministic utility, thus decreasing its choice probability by design. If initial model estimation leads to counter-intuitively signed coefficients, such as positive cost coefficients, then this is easily spotted by an analyst and serves as an invitation to refine the model specification or deal with data issues. Once all signs are as expected, the monotonic relationship is guaranteed. By contrast, this is not the case in DNNs because of the complex nonlinear model structure, especially when the number of hidden layers increases. For example, Xia et al. (2023) observed non-monotonic demand predictions with increasing generalized costs in a mode choice experiment with DNNs, which suggests the need to investigate the monotonicity of DNNs to improve their behavioral regularity. Although shallow neural networks might reduce the risk of non-monotonic behavior of DNNs (Alwosheel et al., 2019; Zhao et al., 2020), this might come at the cost of reduced modeling flexibility and universal approximation power. Alternatively, Han et al. (2022) and Sifringer et al. (2020) proposed to use DNNs only for learning latent representation in the utility function, while resorting to the DCM framework to ensure model monotonicity. On the other hand, depending on model design, these hybrid DNN models might still produce irregular predictions (Wang et al., 2021; Wong and Farooq, 2021). Moreover, hybrid DNN models are a compromise for regularity since they again require the subjective process of model specification. To fully utilize the capability of DNNs, previous studies have attempted to migrate the non-monotonic issue through model ensemble. However, irregular patterns might still be observed after averaging over multiple trainings (Wang et al., 2020a,b; Xia et al., 2023). One promising direction is to directly integrate domain-specific knowledge into the design and training of DNNs, such as incorporating monotonicity constraints in model training (Haj-Yahia et al., 2023). However, there is no consensus on how to measure or improve the model regularity of DNNs within the choice modeling field. This paper contributes to the development of a behavioral regularity measure and a novel regularization framework.

As discussed in computer science applications, the regularity of DNNs can be improved by employing either hard or soft constraints. The first category enforces monotonicity by model construction, e.g., constraining the positiveness of weights in hidden layers (Daniels and Velikova, 2010; Dugas et al., 2009; Sill, 1997), restricting the derivative to be positive (Neumann et al., 2013; Wehenkel and Louppe, 2019), down-weighting samples that violate monotonicity (Archer and Wang, 1993), and incorporating deep lattice network for learning monotonic functions (You et al., 2017). The second category achieves monotonicity by regularization, i.e., by augmenting a regularization term in the loss function to jointly improve model monotonicity. Regularization is firmly rooted in constrained optimization, including the Lagrangian method as an example (Boyd and Vandenberghe, 2004). It has been widely applied as a local method to penalize constraint violations. For example, Sill and Abu-Mostafa (1996) penalized squared deviations in monotonic-

ity for virtual pairs of input variables, while Gupta et al. (2019) proposed a pointwise loss that embeds prior knowledge about monotonicity. Moreover, gradient regularization has also been used to enhance model robustness against adversarial examples (Lyu et al., 2015; Ross and Doshi-Velez, 2018), e.g., penalizing the squared L_2 norm of the gradient of the loss with respect to (w.r.t.) inputs (Drucker and Le Cun, 1991; Ororbia II et al., 2017), penalizing the squared Frobenius norm of the Jacobian matrix of probabilities (Sokolić et al., 2017) and utilities (Jakubovitz and Giryes, 2018) w.r.t. inputs. Note that regularizing the gradient norms might be less effective to improve monotonicity than regularizing the gradient's direction like Haj-Yahia et al. (2023). Inspired by the aforementioned regularization methods, we will design our own approaches in the demand modeling context.

3 Methodology

3.1 DNNs for choice analysis

The discrete choice problem is cast as a supervised classification problem in DNN-based choice analysis. Assuming there are in total D explanatory variables $(x_1, \dots, x_D)$ for all alternatives, the attribute vector of individual n can be written as $\mathbf{x}_n = [x_{n1}, \dots, x_{nD}]^\top$. Then, a DNN model predicts the probability of n choosing i out of J alternatives, i.e., $P_{ni} : \mathbb{R}^D \to (0, 1)$ and $\sum_{i=1}^{J} P_{ni} = 1$. The observed choice vector $\mathbf{y}_n \in \{0, 1\}^J$ of n is used for DNN training, where $y_{ni} = 1$ if alternative i is chosen, and $y_{ni} = 0$ otherwise. Similar to conventional RUMs, DNN models aim to find specifications with high predictive power and behavioral regularity.

However, contrasting to the manual model specifications of RUMs, DNNs automatically learn model specifications with their unique representation learning capability. Specifically, utility vector $\mathbf{V}_n = [V_{n1}, \dots, V_{nJ}]^\top$ is specified through a series of transformations, termed as hidden layers $(f_1, \dots, f_H)$. Each layer f_h contains a learnable parameter matrix W_h, a bias vector $\mathbf{b}_h$, and an activation function $\varphi(\cdot)$ (e.g., the rectified linear unit, ReLU) to transform $\mathbf{x}_n$. Specifically, each layer transformation can be written as

$$f_h(\mathbf{x}_n) = \varphi(W_h \mathbf{x}_n + \mathbf{b}_h) \tag{1}$$

and the utility vector $\mathbf{V}_n$ is computed in a composite form:

$$\mathbf{V}_n = (f_H \circ f_{H-1} \circ \cdots \circ f_2 \circ f_1)(\mathbf{x}_n) \tag{2}$$

Finally, a softmax classification layer (i.e., the logistic function) outputs the choice probability of i as

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_{j=1}^{J} e^{V_{nj}}} \tag{3}$$

The DNN structure generalizes the classical linear MNL model. If a DNN is specified with a single hidden layer and identity activation function, the utility function in Eq. (2) would collapse to

$$\mathbf{V}_n = f(\mathbf{x}_n) = W \mathbf{x}_n + \mathbf{b} \tag{4}$$

where W can be interpreted as coefficients and $\mathbf{b}$ as alternative-specific constants. Although closely related to the MNL model, DNNs allow flexible model specification through multi-layer nonlinear transformations. We illustrate in Fig. 1 a feedforward DNN structure with four hidden layers and one classification layer for a choice modeling problem with D attributes for J alternatives.

3.2 Behavioral regularity metrics

In this study, we propose a novel metric to evaluate behavioral regularity, which measures the monotonicity of the aggregate choice probability functions. The proposed metric essentially measures the monotonicity

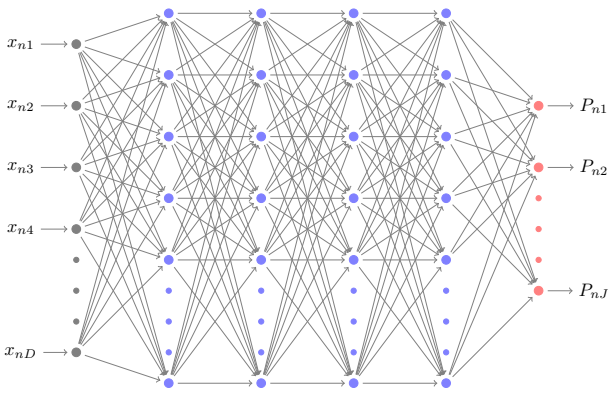


Figure 1: A feedforward DNN structure with four hidden layers and one classification layer.

consistency between the model and prior knowledge on the "correct" signs of parameter estimates, commonly used in the subjective process for selecting the "optimal" specification of RUMs. We define the behavioral regularity metric of alternative i w.r.t. a cost variable x_d as

$$B_{id} = \int \mathbb{1}\left\{\frac{\partial P_i(\mathbf{x})}{\partial x_d} < \varepsilon\right\} \rho(z) dz \tag{5}$$

where $P_i(\mathbf{x})$ maps a representative individual's attributes $\mathbf{x}$ (including both individual-specific sociodemographic factors and alternative-specific cost variables) to the individual's probability of choosing i, z is a sociodemographic factor of a population group with density $\rho(z)$, and $\mathbb{1}\{\cdot\}$ is an indicator function that equals 1 if $\partial P_i(\mathbf{x})/\partial x_d < \varepsilon$, and 0 otherwise. Parameter ε represents the modeler's prior assumptions on the monotonicity of P_i w.r.t. x_d:

(1) $\varepsilon = 0$, termed as *strong* regularity, requires P_i to be strictly decreasing w.r.t. x_d. The formulation assumes that all individuals in population group z respond negatively to x_d.

(2) $\varepsilon > 0$, termed as *weak* regularity, relaxes the strict monotonicity assumption and allows P_i to be non-increasing w.r.t. x_d. The formulation assumes that some individuals do not respond (with zero derivative) or even respond positively (with positive derivative) to x_d, implying that the behavioral regularity metric becomes weak.

As an illustration, a classical linear MNL model with a negative parameter w.r.t. x_d yields $B_{id} = 1$, which implies that all individual behaviors are consistent with the demand monotonicity assumption and well-captured by the model.

The population-based behavioral regularity measure in Eq. (5) can be approximated by the mean behavioral regularity across individuals:

$$B_{id} \approx \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left\{\frac{\Delta P_{ni}}{\Delta x_{nd}} < \varepsilon\right\} \tag{6}$$

where N is the sample size, and the partial derivative is computed with finite differences. The proposed empirical regularity measure in Eq. (6) approaches the exact metric in Eq. (5) when the sample size increases. Our behavioral regularity metrics can be extended to incorporate taste heterogeneity across population groups and even individuals within each group by distinguishing ε w.r.t. different groups, that is, ε can be further specified as ε_z to reflect the group-specific thresholds. Meanwhile, our behavioral regularity metrics B_{id} only require the aggregate regularity rather than individual one, which is inspired by classical economics discussions that market rationality is a fundamental law while individual behaviors might present more diverse and irrational patterns (Becker, 1962).

3.3 Achieving behavioral regularity by constrained optimization

3.3.1 Unconstrained likelihood maximization

DNN-based choice models can be estimated using the likelihood maximization framework. Given a set of hyperparameters, an unconstrained DNN learns parameters W through minimizing the cross-entropy L:

$$\min_W L(W) = \min_W \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{J} -y_{ni} \log P_i(\mathbf{x}_n; W) \tag{7}$$

The unconstrained formulation in Eq. (7) is sufficient for the estimation of conventional DCMs, since they often satisfy convexity conditions under linear utility specification. In a linear MNL model, for example, choice probability P_{ni} increases monotonically with utility V_{ni} according to Eq. (3). The linear specification in Eq. (4) induces monotonicity of utility w.r.t. cost variables. Therefore, if individuals indeed perceive higher utility with lower costs, the optimization would result in negative parameter estimates, and the behavioral monotonicity is clearly satisfied by evoking the chain rule. The multi-layer nonlinear transformations in neural networks allow for approximation of arbitrary functions, but these complex transformations might lead to non-monotonic choice probability functions, especially when the network is deep. In this case, unconstrained likelihood maximization can no longer guarantee a behaviorally regular model.

3.3.2 Constrained likelihood maximization

To address the irregularity issue of DNNs, we introduce a set of behavioral regularity constraints into the optimization problem, which yields

$$\min_W \quad L(W) \tag{8}$$

$$\text{s.t.} \quad R(\mathbf{x}_n; W) \leq 0, \quad n = 1, \ldots, N \tag{9}$$

where R constrains the attributes $\mathbf{x}_n$ of an individual n, defined as $R : \mathbb{R}^{J \times D} \to \mathbb{R}$, where dimension J is the number of alternatives and D is the number of attributes. Hence the behavioral regularity constraints are imposed at the individual level to achieve the aggregate behavioral regularity in B_{id}. The specific behavioral regularity constraints will be designed in the next subsection.

Training DNNs with constraints is challenging. We tackle this problem by treating the *hard* constraints in Eq. (9) as *soft* constraints, motivated by the Lagrangian relaxation method. Given a hyperparameter λ, we consider the following optimization problem:

$$\min_W L(W) + \lambda \sum_{n=1}^{N} R(\mathbf{x}_n; W) \tag{10}$$

where λ controls the strength of the behavioral regularity constraint and can be interpreted as a Lagrangian multiplier for constrained optimization. We note that the relaxation formulation in Eq. (10) is similar to the regularization methods (e.g., L_1 and L_2 norms) that are commonly applied in machine learning for model sparsity, while our motivation is to improve the behavioral regularity of the DNN choice models.

Compared to the hard constraint formulation, soft regularization can flexibly accommodate the various degrees of validity in our behavioral regularity assumptions. Similar to the motivation for the weak regularity metric, our approach allows each individual n to somewhat violate the preset constraint $R(\mathbf{x}_n)$, thus accommodating the potentially irregular behavior of certain individuals. As a result, hyperparameter λ provides insight into the consistency between behavioral regularity assumptions and the actual behavior of studied individuals. If a larger λ is required to achieve higher predictive performance, it might imply that the actual behavior is inconsistent with prior assumptions, thus providing extra insight into the validity of behavioral regularity constraints.

3.4 Gradient regularization

We design gradient regularizers to improve the behavioral regularity of DNN-based choice models. Specifically, we constrain the demand feedback on generalized costs by the gradient's direction (i.e., signs of the parameter estimates) and magnitude.

For individual n, the Jacobian matrix (gradient) of demand vector $\mathbf{P} = [P_1, \dots, P_J]^\top$ w.r.t. cost variables $\{x_1, \dots, x_D\}$ can be written as

$$\nabla \mathbf{P}(\mathbf{x}_n) = \begin{bmatrix} \frac{\partial P_1}{\partial x_1}(\mathbf{x}_n) & \cdots & \frac{\partial P_1}{\partial x_D}(\mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ \frac{\partial P_J}{\partial x_1}(\mathbf{x}_n) & \cdots & \frac{\partial P_J}{\partial x_D}(\mathbf{x}_n) \end{bmatrix} \tag{11}$$

which includes three types of partial derivatives:

(1) Direct derivatives: e.g., the probability of driving w.r.t. the driving cost.
(2) Cross derivatives: e.g., the probability of driving w.r.t. the train travel time.
(3) Sociodemographic derivatives: e.g., the probability of driving w.r.t. the age of traveler.

The constrained likelihood maximization framework in Eq. (10) allows us to impose different gradient constraints to the three types of partial derivatives in the Jacobian matrix. Since behavioral regularity is reflected by the gradient's direction, we introduce a mask matrix for individual n:

$$\Psi(\mathbf{x}_n) = \begin{bmatrix} \mathbb{1}\left\{\frac{\partial P_1}{\partial x_1}(\mathbf{x}_n) \notin \mathbb{S}_{11}\right\} & \cdots & \mathbb{1}\left\{\frac{\partial P_1}{\partial x_D}(\mathbf{x}_n) \notin \mathbb{S}_{1D}\right\} \\ \vdots & \ddots & \vdots \\ \mathbb{1}\left\{\frac{\partial P_J}{\partial x_1}(\mathbf{x}_n) \notin \mathbb{S}_{J1}\right\} & \cdots & \mathbb{1}\left\{\frac{\partial P_J}{\partial x_D}(\mathbf{x}_n) \notin \mathbb{S}_{JD}\right\} \end{bmatrix} \tag{12}$$

where $\mathbb{1}\{\cdot\}$ is an indicator function that equals 1 if $\partial P_i(\mathbf{x}_n)/\partial x_d \notin \mathbb{S}_{id}$, and 0 otherwise; and set $\mathbb{S}_{id}$ defines the expected sign of the partial derivative. Combing the mask matrix $\Psi(\mathbf{x}_n)$ and the Jacobian matrix $\nabla \mathbf{P}(\mathbf{x}_n)$, we define the *sum-based* gradient regularization using the Frobenius inner product[1]:

$$R_\sigma(\mathbf{x}_n) = \langle \Psi(\mathbf{x}_n), \nabla \mathbf{P}(\mathbf{x}_n) \rangle_F \tag{13}$$

This sum-based gradient regularization flexibly accommodates different prior assumptions on the signs of the derivatives. Set $\mathbb{S}_{id}$ can take negative values ($\mathbb{S}_{id} = \mathbb{R}^-$), positive values ($\mathbb{S}_{id} = \mathbb{R}^+$), or any real values ($\mathbb{S}_{id} = \mathbb{R}$), depending on the prior assumption on attribute x_d's effect on demand $P_i(\mathbf{x}_n)$. For example, by imposing $\mathbb{S} = \mathbb{R}^-$ on the direct derivatives, they are expected to be negative and penalized if non-negative. On the other hand, when there is no prior assumption regarding a derivative, we allow all possible signs by taking $\mathbb{S} = \mathbb{R}$. Despite such flexibility, we only impose negative constraints on the direct derivatives throughout our empirical experiments, which is the least controversial among all possibilities.

Alternative to the sum-based approach, we could also regularize the gradient's magnitude, implying that demands are not expected to change drastically with small cost perturbations. Using the same notations, we define the *norm-based* regularization as

$$R_\nu(\mathbf{x}_n) = \|\nabla \mathbf{P}(\mathbf{x}_n)\|_F^2 = \langle \nabla \mathbf{P}(\mathbf{x}_n), \nabla \mathbf{P}(\mathbf{x}_n) \rangle_F \tag{14}$$

This norm-based regularization is relatively common in the computer science literature (Drucker and Le Cun, 1991; Jakubovitz and Giryes, 2018), thus serving as a benchmark for our empirical experiments. Although smoothness is a relatively common assumption from a pure mathematical perspective, it is not founded on strong behavioral regularity beliefs due to possible threshold effects of pricing.

The regularization terms proposed above are termed as probability gradient regularizers (PGRs) because they exploit the analytical relationship between demand monotonicity and probability gradients $\nabla \mathbf{P}(\mathbf{x}_n)$.

[1] For real matrices, we have $\langle A, B \rangle_F = \sum_{i,j} A_{ij} B_{ij}$.

Due to the computational chain among utilities, choice probabilities, and log-likelihoods, it is also possible to replace the probability gradients by utility and log-likelihood gradients. The two alternative regularizers are defined as:

(1) Utility gradient regularizers (UGRs): According to the softmax function in Eq. (3), choice probability $P_i(\mathbf{x}_n)$ increases monotonically with utility $V_i(\mathbf{x}_n)$. Consequently, demand monotonicity can be retained by regularizing the utility monotonicity w.r.t. generalized costs and evoking the chain rule. Therefore, we construct UGRs by replacing $\mathbf{P}(\mathbf{x}_n)$ with $\mathbf{V}(\mathbf{x}_n)$ in derivation.

(2) Log-likelihood gradient regularizers (LGRs): We define the individual- and alternative-specific log-likelihood as $l_i(\mathbf{x}_n) = -y_{ni} \log P_i(\mathbf{x}_n)$. Since logarithmic transformation is monotonic, demand monotonicity can be retained by regularizing the log-likelihood monotonicity w.r.t. generalized costs. Thus we construct LGRs by replacing $\mathbf{P}(\mathbf{x}_n)$ with $\mathbf{l}(\mathbf{x}_n)$ in derivation.

In brief, by combining sum- and norm-based regularization with probability, utility, and log-likelihood gradients, we have designed six gradient regularizers. They are hereafter referred to as the sum-PGR, sum-UGR, sum-LGR, norm-PGR, norm-UGR, and norm-LGR, all of which will be tested thoroughly in our empirical experiments.

4 Setup of experiments

4.1 Datasets

Our experiments use My Daily Travel survey data[2], which were collected in the Chicago metropolitan area in 2018–2019. After preliminary cleaning, the full dataset retains 26,974 trips made using four travel modes: driving (automobile), walking, train, and bicycle. Around 70% of the trips use automobile, while the proportion of bicycle trips is negligible. Hence the walking and bicycle modes were merged into a single active mode to create a more balanced dataset. Based on the spatial information of each trip in terms of origin and destination, we compiled level of service data by utilizing Google Directions API to collect the travel time of each mode, where active times were calculated by averaging walking and bicycle times. Train costs were provided by the dataset, while driving costs were computed by summing the money paid to toll plazas en route and parking lots. The K-nearest neighbors algorithm was applied to impute the missing data, especially for driving and train costs. This study uses nine major individual- and alternative-specific input variables, among which seven are numerical and two are categorical. The former includes the travel time of each mode, driving and train costs, age of traveler, and number of household vehicles, while the latter includes indicators of sexuality and high income household (annual income $\geq$ \$75,000). Table A.1 summarizes the basic statistics of the full dataset, where many variables have right-skewed distributions.

The original dataset was then reprocessed to create three datasets to examine the effects of large versus small sample sizes, and in-domain versus out-of-domain generalizations. The first dataset, named as *10K-Random*, incorporates 10,000 trips with 70% randomly sampled for training and 30% for testing. This dataset is used as the benchmark because it has a sufficient sample size and its random sampling scheme is common to examine the in-domain generalizability. The second dataset, named as *1K-Random*, includes 1,000 trips with the same training-testing split as the 10K-Random dataset. By comparing the results between these two datasets, we could evaluate how predictive performance and behavioral regularity vary with sample sizes. The third dataset, named as *10K-Sorted*, has the same sample size as the 10K-Random dataset. However, unlike the random split, we sorted the 10,000 trips by driving cost, with the lower 70% used as the training set and the upper 30% as the testing set. As shown in Table A.2, the distributions of variables are quite different between the training and testing sets, with significantly higher means and standard deviations in the testing set. In this way, the training-testing split scheme facilitates the testing carried out on more expensive trips. This out-of-domain generalizability is not only of theoretical interests, but also highly relevant in practice because it investigates model transferability, i.e., how the models perform

[2]See https://www.cmap.illinois.gov/data/transportation/travel-survey.

in a target context distinct from their source context. In fact, the sampling scheme of the 10K-Sorted dataset resembles a common transportation policy setting, in which policy makers utilize data collected from the status quo to discuss how travelers would respond to price changes.

4.2 Experimental design

We specify a hyperparameter space of DNNs to search for the optimal architecture and hyperparameter configuration. The hyperparameters were randomly selected from Table 1 for DNN training and finalized based on the model performance in the corresponding testing set. The final optimal DNN architecture has four hidden layers and 100 neurons per layer. Since Table 1 also incorporates regularization strength λ, the random hyperparameter search helps examine the balance between prediction quality and behavioral regularity. To fully demonstrate the effects of λ, we took λ values from 10^{-5} to 10^{3}. It is expected that the DNNs with extremely small λ's would approximate the benchmark DNN, while those with extremely large λ's would have lower prediction quality. The DNNs were trained with standardized input variables by `PyTorch` of Python, under default settings of the Adam algorithm (e.g., the learning rate is 10^{-3}). This algorithm, derived from adaptive moment estimation, utilizes stochastic gradient descent to train deep learning models (Kingma and Ba, 2014). The batch sizes of the training sets were set to 1/10 of the sample sizes. For simplicity, only direct derivatives are regularized during DNN training. To mitigate model randomness, we analyze the ensemble performance by averaging the results of 10 trainings with random seeds 0–9. Finally, to evaluate the behavioral regularity of DNNs, we set parameter ε to -10^{-4} for strong regularity and 10^{-4} for weak regularity.

Table 1: Hyperparameter space.

Depth	$\{2, 3, 4, 5, 6\}$
Width	$\{50, 100, 150\}$
Regularization strength λ	$\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1, 10, 10^{2}, 10^{3}\}$

For comparative analysis, we also estimated a linear MNL model by `PyLogit` of Python with unstandardized input variables. The MNL is widely applied in choice analysis due to its convenience in estimation and interpretation (Ben-Akiva and Lerman, 1985). The utility specification is shown below:

$$V_{n1} = w_{\text{time},1}\text{time}_{nc} + w_{\text{cost},1}\text{cost}_{n1} \tag{15}$$

$$V_{n2} = w_{0,2} + w_{\text{sex},2}\text{sex}_n + w_{\text{inc},2}\text{inc}_n + w_{\text{age},2}\text{age}_n + w_{\text{veh},2}\text{veh}_n + w_{\text{time},2}\text{time}_{n2} + w_{\text{cost},2}\text{cost}_{n2} \tag{16}$$

$$V_{n3} = w_{0,3} + w_{\text{sex},3}\text{sex}_n + w_{\text{inc},3}\text{inc}_n + w_{\text{age},3}\text{age}_n + w_{\text{veh},3}\text{veh}_n + w_{\text{time},3}\text{time}_{n3} \tag{17}$$

where subscripts 1–3 represent automobile, train, and the active mode, respectively, and automobile is set as the reference alternative for individual-specific variables. In our experiments, the MNL provides important insight as a benchmark model, but it is *not* necessary to judge the performance of DNNs based on their prediction similarity to the MNL model.

5 Results

In this section, we present the results of our empirical work in two stages. Section 5.1 compares the DNNs with gradient regularization to the benchmark DNN and MNL models regarding their prediction quality and behavioral regularity metrics, where we do this separately for the large sample (10K-Random), small sample (1K-Random), and out-of-domain generalization (10K-Sorted) scenarios. Section 5.2 investigates how the regularization strength influences the trade-off between prediction quality and behavioral regularity under the same three settings. Section 5.1 presents the optimum DNNs with gradient regularizers, while Section 5.2 provides the underlying reasoning why gradient regularization can enhance prediction quality and behavioral regularity.

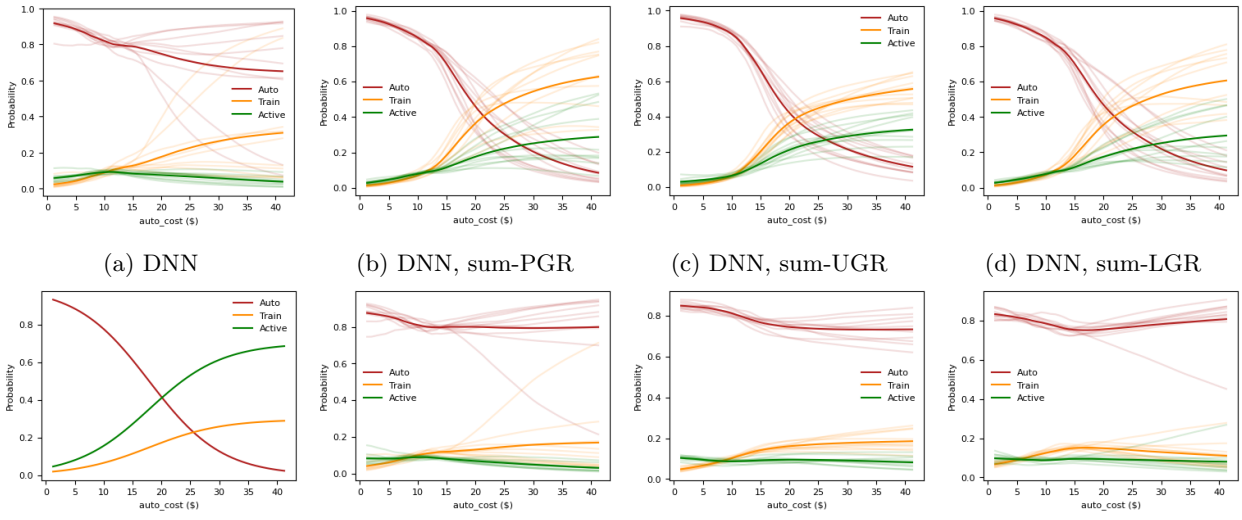
5.1 Enhancing model performance with gradient regularization

5.1.1 Large sample scenario

Using the 10K-Random dataset, we investigate the regularized DNNs by four metrics in the testing set, namely accuracy (hit rate), log-likelihood, and strong and weak regularities. The former two describe the models' prediction quality, while the latter two describe their behavioral regularity. Table 2 summarizes the performance of six DNN models with adequate strengths of gradient regularization, alongside the MNL and DNN benchmark models. To provide visual intuition, we plot the choice probability functions of the three travel modes for the eight models in Fig. 2, where light and dark curves represent the results of training replications and ensembles, respectively. Fig. 2 uses an "average individual" as the market representative, and varies the driving cost while keeping all other variables constant. By examining the large sample and in-domain scenarios (10K-Random dataset), we have three major empirical findings.

Table 2: Performance of DNNs and the MNL in the testing set (10K-Random dataset).

Panel 1: Sum-based gradient regularization					
	MNL	DNN	DNN, PGR	DNN, UGR	DNN, LGR
Accuracy	72.0%	72.9%	73.0%	72.9%	73.0%
Log-likelihood	−2058.4	−1903.2	−1906.0	−1921.6	−1904.9
Strong regularity	1	0.6760	0.9772	0.9491	0.9758
Weak regularity	1	0.7114	0.9973	0.9729	0.9970
Panel 2: Norm-based gradient regularization					
	MNL	DNN	DNN, PGR	DNN, UGR	DNN, LGR
Accuracy	72.0%	72.9%	73.1%	73.0%	73.0%
Log-likelihood	−2058.4	−1903.2	−1911.8	−1945.5	−1981.4
Strong regularity	1	0.6760	0.5181	0.6004	0.5079
Weak regularity	1	0.7114	0.5548	0.6476	0.5347



(a) DNN (b) DNN, sum-PGR (c) DNN, sum-UGR (d) DNN, sum-LGR

(e) MNL (f) DNN, norm-PGR (g) DNN, norm-UGR (h) DNN, norm-LGR

Figure 2: Individual demands as functions of driving cost (10K-Random dataset).

Firstly, without gradient regularization, the benchmark DNN slightly outperforms the MNL in prediction quality but significantly underperforms the MNL in behavioral regularity. The benchmark DNN improves the MNL's accuracy by 0.9 percentage point and the MNL's log-likelihood by 7.5% of its absolute value. It appears that log-likelihood is a more sensitive indicator of predictive performance than accuracy, potentially caused by mode imbalance in the dataset: automobiles account for more than 60% of the trips, impeding further improvements in prediction accuracy. However, the benchmark DNN presents significant behavioral irregularity, as suggested by only 67.6% strong regularity and 71.1% weak regularity. This finding is also illustrated intuitively in Fig. 2a. Although the average market share (dark lines) of automobile presents an overall declining trend, many local DNN models (light lines) are counter-intuitive: the market share of automobile could drop too abruptly or even increase when the driving cost increases. This combination of DNNs' high predictive performance and low behavioral regularity aligns with the findings from many previous studies (Wang et al., 2020a,b; Wong and Farooq, 2021; Xia et al., 2023).

Secondly, sum-based gradient regularizers can significantly improve the behavioral regularity of DNNs without sacrificing their prediction quality. With sum-based regularization, DNNs' demand functions become highly monotonic, as indicated by strong and weak regularity metrics both approaching around 95–99%. Specifically, the sum-PGR, which directly regularizes the demand function, improves the strong and weak regularities of the benchmark DNN respectively by 30.1 and 28.6 percentage points; while the sum-UGR and sum-LGR, which exploit demand monotonicity through the chain rule, lead to comparable performance in prediction quality and behavioral regularity. This finding is further elaborated by Fig. 2b–d, where the regularized DNNs have individual demand functions more consistent with the MNL: automobile is less favored due to increased costs, while train and the active mode see higher demand. The regularized demand functions are more monotonic not only in the ensemble models, but also in the training replications (light curves in the figure). Comparing three sum-based methods, they differ by only 0.1 percentage points in prediction accuracy and less than 3.0 percentage points in behavioral regularities.

On the other hand, norm-based gradient regularizers fail to enhance behavioral regularity or prediction quality. For example, the norm-UGR leads to a 0.1 percentage point increase in accuracy, but a 7.6 percentage point decrease in strong regularity and a 6.4 percentage point decrease in weak regularity. These results are shared by all three norm-based approaches, potentially because they tend to flatten and smoothen the demand curves. As shown in Fig. 2f–h, individual demand curves remain nearly constant at around 80% regardless of the large variation in driving cost from the minimum \$1.24 to the maximum \$41.25. With relatively strong norm-based regularization, the demand curves become nearly flat and could not reflect the decision mechanism: travelers might not respond to cost changes at certain points, but are highly unlikely to be insensitive to all cost changes. Since smaller λ's lead to better performance, we might conclude that the dataset or model itself is somewhat behaviorally regular and thus does not change abruptly due to cost perturbations. In brief, although regularizing the gradient norm is a common practice in computer science (Drucker and Le Cun, 1991; Jakubovitz and Giryes, 2018; Sokolić et al., 2017), it is not founded on prior beliefs in behavioral regularity and thus not effective for our purposes.

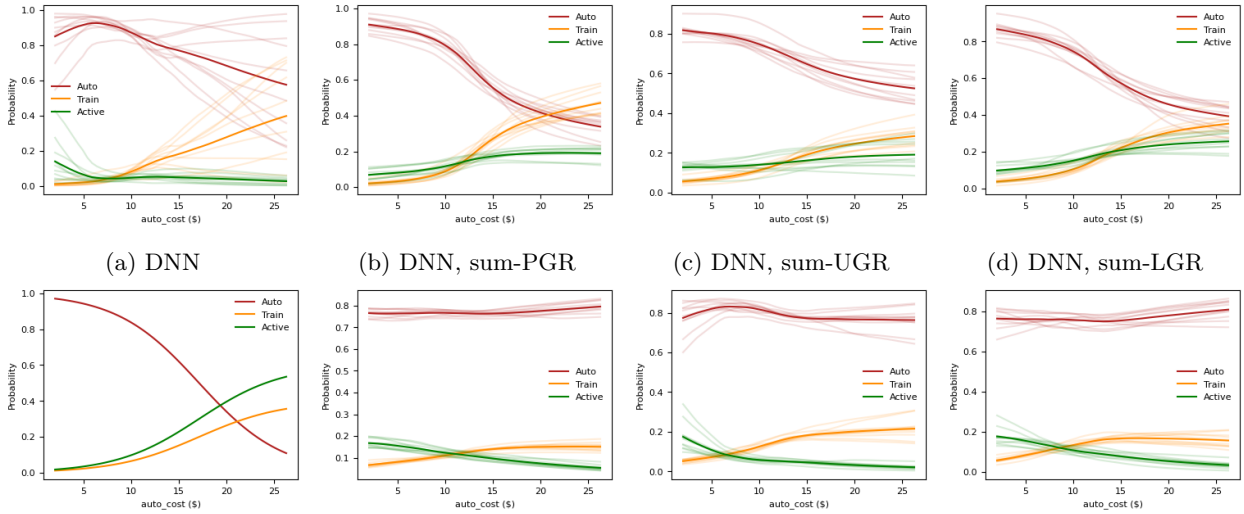
5.1.2 Small sample scenario

The same analysis is applied to a smaller sample of 1K randomly selected trips, i.e., the 1K-Random dataset. This small-sample analysis helps demonstrate the benefit of gradient regularization in choice modeling practice, where large-scale data collection is difficult due to resource limitations or privacy concerns. For example, Ben-Elia et al. (2013) collected stated-preference data from only 36 participants, each with 20 repetitions in different scenarios, leading to a sample size of only 720 observations. The Borlänge GPS dataset used by Fosgerau et al. (2013) and Mai et al. (2015) consists of only 1,832 trips. To examine the small sample scenario, Table 3 illustrates the model performance and Fig. 3 elaborates on the individual demand functions, using the same format as Table 2 and Fig. 2.

We find that gradient regularization is even more effective than in the large sample scenario with details as follows. Firstly, without regularization, the benchmark DNN is even less behaviorally regular in

Table 3: Performance of DNNs and the MNL in the testing set (1K-Random dataset).

Panel 1: Sum-based gradient regularization					
	MNL	DNN	DNN, PGR	DNN, UGR	DNN, LGR
Accuracy	73.7%	71.4%	74.8%	73.9%	74.0%
Log-likelihood	−211.5	−221.6	−199.3	−202.9	−203.9
Strong regularity	1	0.4850	0.9763	0.9437	0.9830
Weak regularity	1	0.5503	0.9990	0.9517	0.9993
Panel 2: Norm-based gradient regularization					
	MNL	DNN	DNN, PGR	DNN, UGR	DNN, LGR
Accuracy	73.7%	71.4%	74.3%	73.2%	74.1%
Log-likelihood	−211.5	−221.6	−200.5	−204.8	−201.0
Strong regularity	1	0.4850	0.4820	0.4053	0.4340
Weak regularity	1	0.5503	0.5693	0.4583	0.4807



(a) DNN (b) DNN, sum-PGR (c) DNN, sum-UGR (d) DNN, sum-LGR

(e) MNL (f) DNN, norm-PGR (g) DNN, norm-UGR (h) DNN, norm-LGR

Figure 3: Individual demands as functions of driving cost (1K-Random dataset).

the small sample scenario, as indicated by a 19.1 percentage point drop in strong regularity and a 16.1 percentage point drop in weak regularity. Meanwhile, it is also inferior to MNL in predictive power, for both accuracy and log-likelihood. As shown in Fig. 3a, its individual demand curves are mostly non-monotonic and inconsistent across trainings. Secondly, sum-based gradient regularizers succeed in enhancing all four metrics. In other words, the regularized DNNs outperform the benchmark DNN in both prediction quality and behavioral regularity. For example, the sum-PGR improves its accuracy by 3.4 percentage points and its strong regularity by 49.1 percentage points, which are more significant than in the large sample scenario. Finally, norm-based gradient regularization can improve prediction quality, but not behavioral regularity. The regularized DNNs have similar or higher accuracy and log-likelihood than the benchmark DNN, along with similar or lower behavioral regularities. The individual demand curves in Fig. 3f–h are nearly flat and less divergent across trainings.

5.1.3 *Out-of-domain generalization*

The large and small sample scenarios above assist in examining only in-domain generalization due to the random split of training and testing data. Although random split is the most common practice, we are also interested in the out-of-domain generalizability of DNNs. Out-of-domain generalization is highly relevant to transportation engineering, system design, and urban planning; for example, engineers often cite the ridership performance of other cities' metro systems when evaluating whether a city should build a similar system. Our data split scheme in the 10K-Sorted dataset can be interpreted as using the individual choice behavior before road pricing or parking price increases (as in the training set) to extrapolate the behavior after that (as in the testing set). Here we emulate such policy setting by testing the DNNs' out-of-domain generalizability using prediction quality and behavioral regularity metrics.

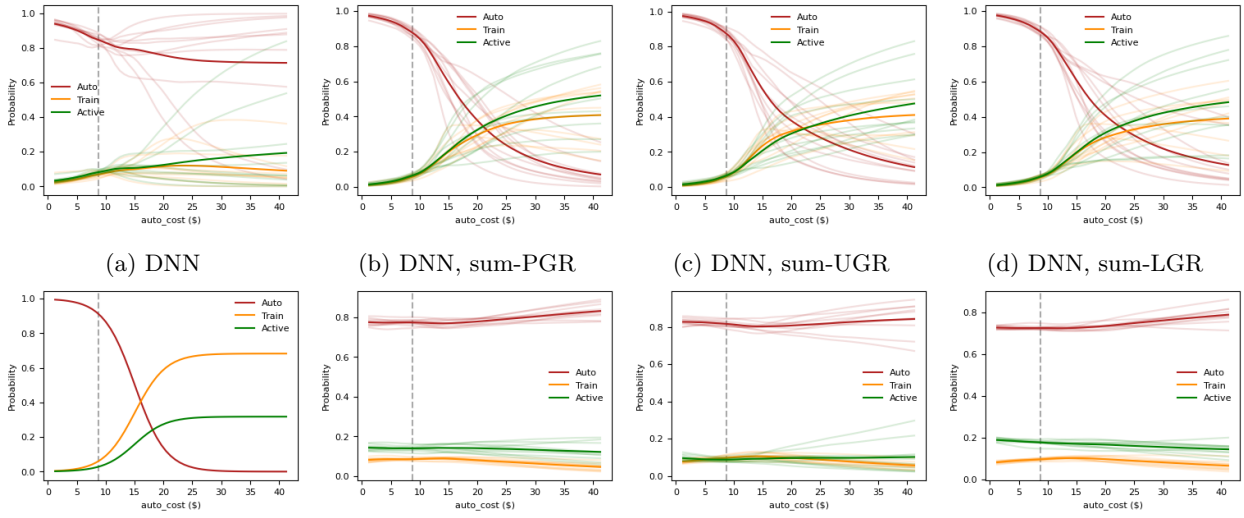
The results suggest that gradient regularization could drastically improve the out-of-domain generalizability of DNNs, even more effectively than improving their in-domain generalizability. Table 4 summarizes the model performance of DNNs and the MNL and Fig. 4 visualizes the individual demand functions varying with the driving cost. Firstly, we can easily find the substantial decrease in the MNL's prediction quality: its accuracy drops by 2.4 percentage points compared with Table 2. By contrast, the DNNs exhibit great flexibility with high accuracy and log-likelihood. The benchmark DNN has adequate prediction quality but performs unstably in the testing set, as shown to the right of the data split threshold (dashed gray line). Secondly, with sum-based gradient regularization, the benchmark DNN's log-likelihood slightly increases. For example, the sum-PGR raises the log-likelihood by 2.7% of its absolute value, in contrast to the 0.1% drop in Table 2. The sum-PGR also increases the strong and weak regularities of the benchmark DNN by 38.3 and 33.3 percentage points, which are higher than the improvements in Table 2. This is also illustrated in Fig. 4. Finally, regularizing the gradient norm can improve the benchmark DNN's prediction quality, such as rising the log-likelihood by 4.9% of its absolute value, at the cost of reducing behavioral regularities.

Table 4: Performance of DNNs and the MNL in the testing set (10K-Sorted dataset).

Panel 1: Sum-based gradient regularization					
	MNL	DNN	DNN, PGR	DNN, UGR	DNN, LGR
Accuracy	69.6%	78.4%	78.1%	77.8%	78.2%
Log-likelihood	−3268.0	−1933.7	−1881.2	−1813.7	−1903.5
Strong regularity	1	0.5307	0.9141	0.9296	0.9098
Weak regularity	1	0.6645	0.9978	0.9959	0.9964
Panel 2: Norm-based gradient regularization					
	MNL	DNN	DNN, PGR	DNN, UGR	DNN, LGR
Accuracy	69.6%	78.4%	78.6%	78.4%	78.5%
Log-likelihood	−3268.0	−1933.7	−1846.8	−1839.8	−1891.6
Strong regularity	1	0.5307	0.3057	0.4172	0.2815
Weak regularity	1	0.6645	0.3734	0.5273	0.3412

5.2 Trade-off between prediction quality and behavioral regularity

Section 5.1 demonstrates how sum-based gradient regularization significantly enhances the benchmark DNN's behavioral regularity and prediction quality under three settings. This subsection will explain why it happens by investigating the trade-off between prediction quality and behavioral regularity in two scenarios of in-domain generalization (large and small samples) and out-of-domain generalization. Although the optimization problem in Eq. (10) demonstrates a clear substitution effect between prediction quality and behavioral



(a) DNN (b) DNN, sum-PGR (c) DNN, sum-UGR (d) DNN, sum-LGR

(e) MNL (f) DNN, norm-PGR (g) DNN, norm-UGR (h) DNN, norm-LGR

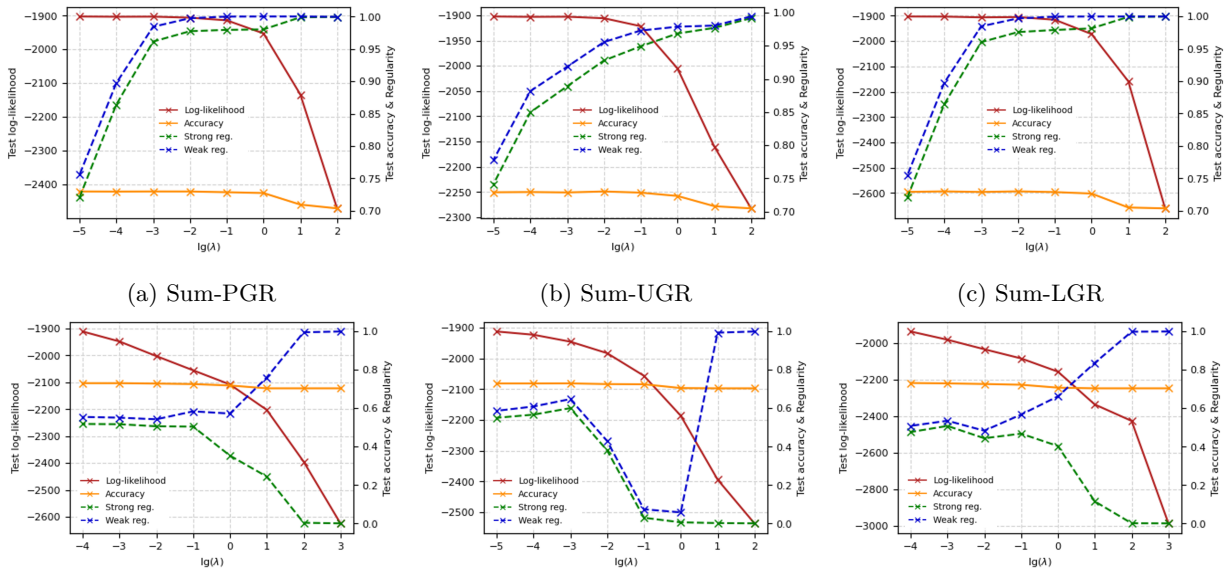
Figure 4: Individual demands as functions of driving cost (10K-Sorted dataset).

regularity in the training set, it remains an open question whether this effect persists in the testing set. As shown below, we find the same substitution effect in the large sample scenario (10K-Random dataset), indicating that prediction quality decreases and behavioral regularity increases with stronger sum-based gradient regularization. But interestingly, complementary effects between prediction quality and behavioral regularity exists in the small sample scenario (1K-Random dataset) and out-of-domain generalization (10K-Sorted dataset): stronger sum-based gradient regularization can enhance both prediction quality and behavioral regularity, detailed as follows.

5.2.1 Large sample scenario: substitution effects

In a relatively large sample, higher behavioral regularity is associated with lower prediction quality. The effects of regularization strength λ on the four metrics are visualized in Fig. 5, where the horizontal axis uses logarithmic coordinates $\lg(\lambda)$ to better show the results when λ is small. As shown in Fig. 5, the DNNs' prediction quality declines with increasing λ, especially as measured by log-likelihood, since accuracy is not very sensitive to the regularization strength. The decline in prediction quality is particularly noticeable after an elbow point, such as $\lambda = 1$ for the sum-PGR and sum-LGR. In other words, although regularization would generally reduce prediction quality, there exists a range of λ that almost preserves prediction quality while significantly enhancing behavioral regularity, such as $\lambda \in [10^{-5}, 0.1]$ for the sum-UGR, consistent with our findings in Section 5.1.

Fig. 5 also presents two important differences between log-likelihood and accuracy, as well as between sum- and norm-based gradient regularizers. Firstly, log-likelihood is much more sensitive than accuracy in measuring prediction quality, which is theoretically valid and empirically expected for imbalanced data. Therefore, we recommend that future studies use at least both metrics to reflect the probabilistic nature of DCMs. Secondly, our behavioral regularity metrics can demonstrate the flattening effects of norm-based regularization on individual demand curves, because weak regularity approaches 1 and strong regularity approaches 0 for very large λ's, as shown in Fig. 5. The results suggest that strong regularity might be more appropriate than the weak one, at least for describing the global declining trend of demand curves, although weak behavioral regularity metric could still be important for describing local insensitivity to cost changes.



(a) Sum-PGR (b) Sum-UGR (c) Sum-LGR

(d) Norm-PGR (e) Norm-UGR (f) Norm-LGR

Figure 5: Effects of regularization strength (10K-Random dataset).

5.2.2 Small sample scenario: complementary effects

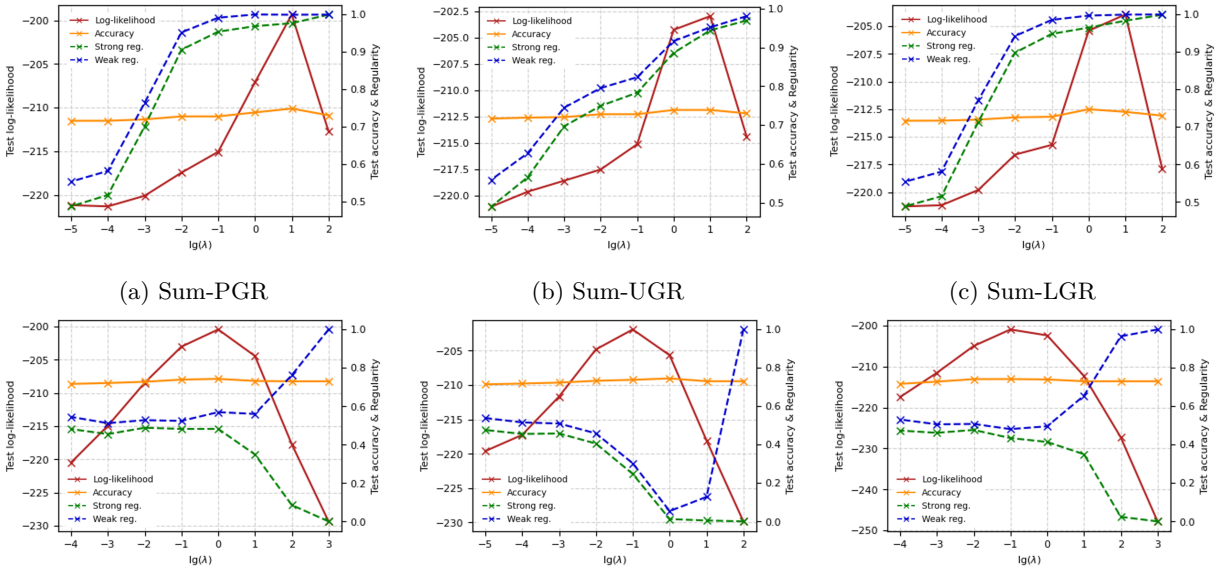
Interestingly, in a relatively small sample, stronger sum-based gradient regularization can simultaneously improve prediction quality and behavioral regularity, thereby advancing the Pareto frontier of all four performance metrics. From Fig. 6, which visualizes the effects of λ for the 1K-Random dataset, we can find a range for each regularizer such that prediction quality and behavioral regularity increase together. When λ increases from 10^{-5} to 10, for example, the sum-UGR improves the DNN's log-likelihood by 8.2% of its absolute value and strong regularity by 45.4 percentage points. This phenomenon suggests the presence of Pareto efficiency in DNN training, despite the substitution effects observed in Fig. 5. The flattening effects of norm-based regularizers are still obvious as λ increases, but there exists a range where prediction quality improves significantly. We can see that, for example, the log-likelihood of the DNN with norm-UGR improves by 8.1% of its absolute value when λ increases from 10^{-5} to 0.1.

On the other hand, when λ exceeds a certain critical point, we still observe substitution effects between prediction quality and behavioral regularity. For example, when λ increases from 10 to 100, the sum-UGR reduces the DNN's log-likelihood by 5.7% of its absolute value, but still improves the DNN's strong regularity by 2.6 percentage points. This also demonstrates the flexibility of soft constraints by identifying the optimum λ, which reflects the alignment of the data with our behavioral assumptions.

5.2.3 Out-of-domain generalization: complementary effects

To explore the out-of-domain generalizability of DNNs with gradient regularizers, we visualize the effects of λ for the 10K-Sorted dataset in Fig. 7. Interestingly, the trend of each metric combines the characteristics of the first two scenarios, i.e., for norm-based regularization there are substitution effects between prediction quality and behavioral regularity, while for sum-based regularization there are complementary effects. When λ increases from 10^{-5} to 0.1, for example, the sum-UGR improves the DNN's log-likelihood by 5.7% of its absolute value and strong regularity by 31.3 percentage points.

Given the relatively large sample size of the 10K-Sorted dataset, it is not surprising to see substitution effects as in Fig. 5. By contrast, the presence of complementary effects and Pareto efficiency may indicate the effectiveness of our gradient regularizers. It is noteworthy that complementary effects also appear locally

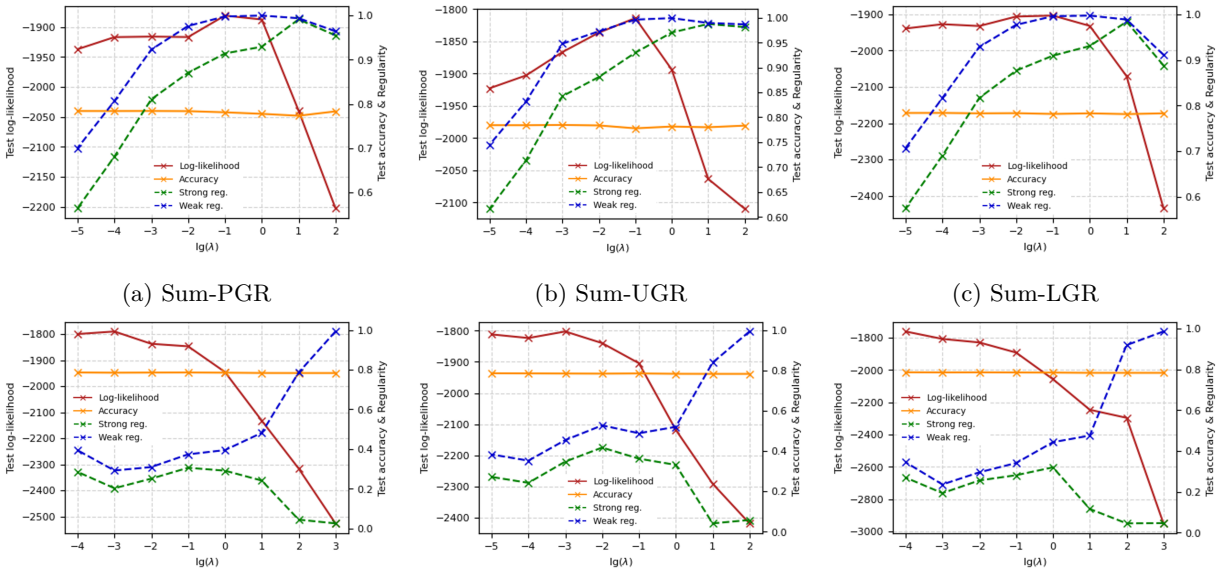


(a) Sum-PGR (b) Sum-UGR (c) Sum-LGR

(d) Norm-PGR (e) Norm-UGR (f) Norm-LGR

Figure 6: Effects of regularization strength (1K-Random dataset).

for out-of-domain generalization. When sum-based regularization is excessive, we still observe substitution effects or even simultaneous decrease in prediction quality and behavioral regularity. Taking the sum-UGR as an example, the former corresponds to $\lambda \in [0.1, 1]$ and the latter corresponds to $\lambda \in [1, 100]$.



(a) Sum-PGR (b) Sum-UGR (c) Sum-LGR

(d) Norm-PGR (e) Norm-UGR (f) Norm-LGR

Figure 7: Effects of regularization strength (10K-Sorted dataset).

6 Conclusions

DNNs often suffer from behaviorally irregular patterns that greatly limit their practical use and theoretical appeal in travel behavior analysis, especially in applications for forecasting. However, there is no consensus on how to measure or improve the model regularity of DNNs within the choice modeling field. This paper makes contributions by developing the behavioral regularity metrics and a gradient regularization framework. Specifically, we propose the "law of demand" in economics as a novel measure of the behavioral regularity of DNNs w.r.t. generalized costs. Using a constrained optimization framework, we design six gradient regularizers to enhance the strong and weak behavioral regularities of DNNs. Empirically, these gradient regularizers are applied to the Chicago travel survey data, through which we examine the trade-off between prediction quality and behavioral regularity in the small vs. large sample scenarios, and in-domain vs. out-of-domain generalization.

We find that sum-based gradient regularization can significantly improve the overall behavioral regularity of DNNs without sacrificing their prediction quality in the large sample scenario. By contrast, norm-based gradient regularization fails to enhance behavioral regularity or prediction quality. We observe different relationships between prediction quality and behavioral regularity. There exists a substitution effect between prediction quality and behavioral regularity in the large sample scenario, but a complementary effect in the small sample scenario. Utilizing the 10K-Sorted dataset, we further find that gradient regularization could be more effective for out-of-domain generalization than in-domain generalization, which is critical for analysis in new situations.

Our results demonstrate how and why the gradient regularizers enhance prediction quality and behavioral regularity, particularly for the small sample scenario and out-of-domain generalization. The findings could be further understood by an analogy between gradient regularization and informative Bayesian prior. The latter can assist in enhancing model performance particularly for the small sample scenario and out-of-domain generalization, because a correct prior belief is more critical than observed data points when the sample size is limited. Therefore, our gradient regularization is similar to the Bayesian prior, although it can be more generally designed and applied because it directly controls the target gradient rather than indirectly restricting the parameter space.

This study pioneers in proposing a new behavioral metric and designing a practical regularization framework for DNNs. To address the irregularity issue observed previously (Wang et al., 2020a,b; Wong and Farooq, 2021; Xia et al., 2023), we incorporate domain-knowledge into DNNs by regularizing the gradient's direction rather than its magnitude as in the computer science literature (Drucker and Le Cun, 1991; Jakubovitz and Giryes, 2018; Sokolić et al., 2017). Only direct partial derivatives are regularized in this paper, thus sidestepping potential validity debates on the behavioral regularity assumptions. Although we only studied some of the most straightforward regularizers, our general framework allows for more complicated regularization. Future studies could use behavioral regularity as a metric along with log-likelihood and accuracy to evaluate travel demand models, and investigate other methods to further enhance behavioral regularity when adopting complex machine learning models.

Contributions of authors

Shenhao Wang conceived of the presented idea; Shenhao Wang and Rui Yao developed the theory and reviewed previous studies; Shenhao Wang, Rui Yao, and Siqi Feng designed the experiments and discussed the results; Siqi Feng conducted the experiments and drafted the manuscript; Shenhao Wang and Rui Yao revised the manuscript; Stephane Hess, Ricardo Daziano, Timothy Brathwaite, and Joan Walker provided comments, Shenhao Wang supervised this work. All authors contributed to the final manuscript.

Acknowledgement

Stephane Hess acknowledges support from the European Research Council through the advanced grant 101020940-SYNERGY.

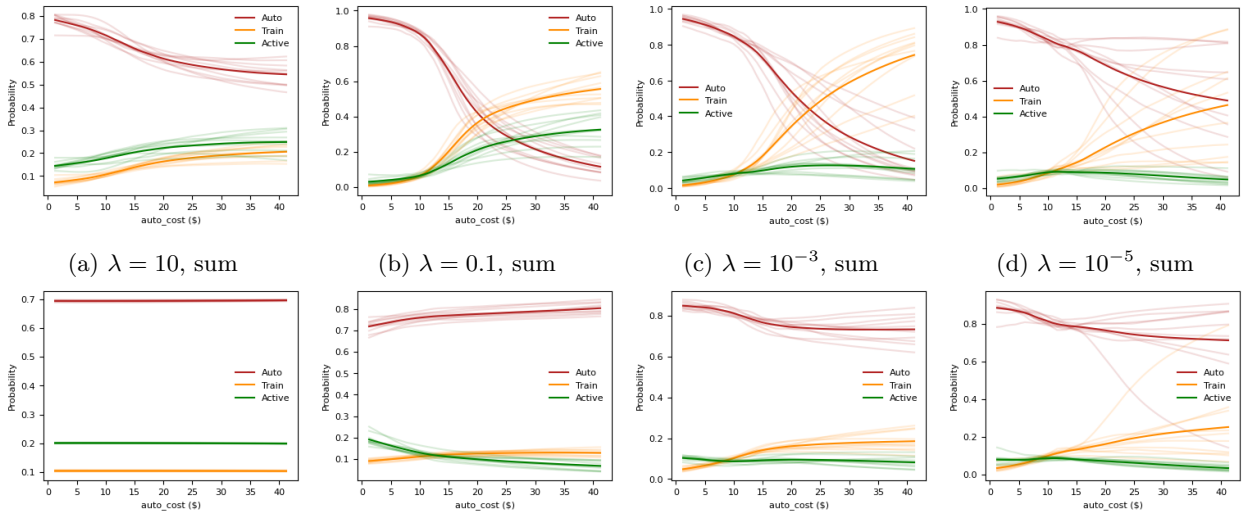
A Summary statistics of datasets

The key statistics of the full dataset as well as the 10K-Sorted dataset are summarized in Tables A.1 and A.2, respectively.

Table A.1: Summary statistics of the full dataset.

Numerical variables	mean	std.	min.	25%	50%	75%	max.
age (year)	39.06	13.30	6.00	29.00	37.00	47.00	84.00
vehicle	1.44	1.00	0.00	1.00	1.00	2.00	8.00
auto_time	13.92	10.37	0.17	6.77	10.60	17.85	86.20
auto_cost ($)	9.22	3.36	1.19	7.09	7.19	10.51	48.24
train_time (min)	61.05	84.76	3.17	23.08	39.29	62.95	961.07
train_cost ($)	2.56	0.40	0.00	2.38	2.42	2.52	10.00
active_time (min)	74.89	94.38	1.98	20.62	39.83	85.70	960.21
Categorical variables							
sexuality	12,241 (1: male); 14,733 (0: female)						
income	16,145 (1: high income); 10,829 (0: low income)						

B Individual demands as functions of regularization strength



(a) $\lambda = 10$, sum (b) $\lambda = 0.1$, sum (c) $\lambda = 10^{-3}$, sum (d) $\lambda = 10^{-5}$, sum

(e) $\lambda = 10$, norm (f) $\lambda = 0.1$, norm (g) $\lambda = 10^{-3}$, norm (h) $\lambda = 10^{-5}$, norm

Figure B.1: Individual demands as functions of driving cost (10K-Random dataset).

Table A.2: Summary statistics of the 10K-Sorted dataset.

Training set							
Numerical variables	mean	std.	min.	25%	50%	75%	max.
age (year)	38.37	13.36	13.00	28.00	36.00	46.00	84.00
vehicle	1.33	0.99	0.00	1.00	1.00	2.00	8.00
auto_time	8.81	4.14	0.17	5.70	8.20	11.22	25.73
auto_cost ($)	7.26	0.69	1.24	7.06	7.09	7.25	8.80
train_time (min)	32.47	18.20	3.17	19.25	28.92	42.70	166.33
train_cost ($)	2.39	0.16	0.00	2.37	2.42	2.44	10.00
active_time (min)	30.53	17.80	1.98	15.94	27.59	42.58	208.94
Categorical variables							
sexuality	3,100 (1: male); 3,900 (0: female)						
income	4,048 (1: high income); 2,952 (0: low income)						
Testing set							
Numerical variables	mean	std.	min.	25%	50%	75%	max.
age (year)	40.48	13.33	6.00	30.00	39.00	50.00	84.00
vehicle	1.62	0.96	0.00	1.00	2.00	2.00	8.00
auto_time	25.46	10.72	4.12	17.65	23.09	31.14	86.07
auto_cost ($)	13.71	2.62	8.80	12.47	14.49	15.63	41.25
train_time (min)	123.98	126.05	10.63	59.58	84.44	129.30	955.32
train_cost ($)	2.94	0.51	0.00	2.47	2.94	3.08	10.00
active_time (min)	174.59	115.01	9.49	92.67	134.29	220.89	875.94
Categorical variables							
sexuality	1,408 (1: male); 1,592 (0: female)						
income	1,865 (1: high income); 1,135 (0: low income)						

References

Alwosheel, A., Van Cranenburgh, S., Chorus, C.G., 2019. 'Computer says no' is not enough: Using prototypical examples to diagnose artificial neural networks for discrete choice analysis. Journal of choice modelling 33, 100186.

Archer, N.P., Wang, S., 1993. Application of the back propagation neural network algorithm with monotonicity constraints for two-group classification problems. Decision Sciences 24, 60–75.

Becker, G.S., 1962. Irrational behavior and economic theory. Journal of Political Economy 70, 1–13.

Ben-Akiva, M.E., Lerman, S.R., 1985. Discrete choice analysis: Theory and application to travel demand. volume 9. MIT Press.

Ben-Elia, E., Di Pace, R., Bifulco, G.N., Shiftan, Y., 2013. The impact of travel information's accuracy on route-choice. Transportation Research Part C: Emerging Technologies 26, 146–159.

Boyd, S.P., Vandenberghe, L., 2004. Convex optimization. Cambridge university press.

Chiappori, P.A., 1985. Distribution of income and the "law of demand". Econometrica 53, 109–127.

van Cranenburgh, S., Wang, S., Vij, A., Pereira, F., Walker, J., 2022. Choice modelling in the age of machine learning - Discussion paper. Journal of Choice Modelling 42, 100340.

Daniels, H., Velikova, M., 2010. Monotone and partially monotone neural networks. IEEE Transactions on Neural Networks 21, 906–917.

Di, X., Liu, H.X., 2016. Boundedly rational route choice behavior: A review of models and methodologies. Transportation Research Part B: Methodological 85, 142–179.

van Dis, E.A., Bollen, J., Zuidema, W., van Rooij, R., Bockting, C.L., 2023. Chatgpt: five priorities for research. Nature 614, 224–226.

Drucker, H., Le Cun, Y., 1991. Double backpropagation increasing generalization performance, in: IJCNN-91-Seattle International Joint Conference on Neural Networks, IEEE. pp. 145–150.

Dugas, C., Bengio, Y., Bélisle, F., Nadeau, C., Garcia, R., 2009. Incorporating functional knowledge in neural networks. Journal of Machine Learning Research 10, 1239–1262.

Fosgerau, M., Frejinger, E., Karlstrom, A., 2013. A link based network route choice model with unrestricted choice set. Transportation Research Part B: Methodological 56, 70–80.

Gupta, A., Shukla, N., Marla, L., Kolbeinsson, A., Yellepeddi, K., 2019. How to incorporate monotonicity in deep networks while preserving flexibility? arXiv preprint arXiv:1909.10662 .

Haj-Yahia, S., Mansour, O., Toledo, T., 2023. Incorporating domain knowledge in deep neural networks for discrete choice models. arXiv preprint arXiv:2306.00016.

Han, Y., Pereira, F.C., Ben-Akiva, M., Zegras, C., 2022. A neural-embedded discrete choice model: Learning taste representation with strengthened interpretability. Transportation Research Part B: Methodological 163, 166–186.

Härdle, W., Hildenbrand, W., Jerison, M., 1991. Empirical evidence on the law of demand. Econometrica 59, 1525–1549.

Hildenbrand, W., 1983. On the "law of demand". Econometrica 51, 997–1019.

Jakubovitz, D., Giryes, R., 2018. Improving dnn robustness to adversarial attacks using jacobian regularization, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 514–529.

Kahneman, D., Tversky, A., 1979. Prospect theory: An analysis of decision under risk. Econometrica 47, 263–292.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Knez, P., Smith, V.L., Williams, A.W., 1985. Individual rationality, market rationality, and value estimation. The American Economic Review 75, 397–402.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature 521, 436–444.

Lichtenstein, S., Slovic, P., 1971. Reversals of preference between bids and choices in gambling decisions. Journal of experimental psychology 89, 46–55.

Lyu, C., Huang, K., Liang, H.N., 2015. A unified gradient regularization family for adversarial examples, in: 2015 IEEE International Conference on Data Mining, IEEE. pp. 301–309.

Mai, T., Fosgerau, M., Frejinger, E., 2015. A nested recursive logit model for route choice analysis. Transportation Research Part B: Methodological 75, 100–112.

May, A., 1992. Road pricing: An international perspective. Transportation 19, 313–333.

McFadden, D., 1974. The measurement of urban travel demand. Journal of Public Economics 3, 303–328.

Neumann, K., Rolf, M., Steil, J.J., 2013. Reliable integration of continuous constraints into extreme learning machines. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 21, 35–50.

Ororbia II, A.G., Kifer, D., Giles, C.L., 2017. Unifying adversarial training algorithms with data gradient regularization. Neural computation 29, 867–887.

Pan, S.J., Yang, Q., 2009. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22, 1345–1359.

Quah, J.K.H., 2000. The monotonicity of individual and market demand. Econometrica 68, 911–930.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D., 2008. Dataset shift in machine learning. MIT Press.

Ross, A., Doshi-Velez, F., 2018. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1660–1669.

Sifringer, B., Lurkin, V., Alahi, A., 2020. Enhancing discrete choice models with representation learning. Transportation Research Part B: Methodological 140, 236–261.

Sill, J., 1997. Monotonic networks. Advances in Neural Information Processing Systems 10, 661–667.

Sill, J., Abu-Mostafa, Y., 1996. Monotonicity hints. Advances in neural information processing systems 9, 634–640.

Simon, H.A., 1957. Models of man; Social and rational. Wiley.

Sokolić, J., Giryes, R., Sapiro, G., Rodrigues, M.R., 2017. Robust large margin deep neural networks. IEEE Transactions on Signal Processing 65, 4265–4280.

Souche, S., 2010. Measuring the structural determinants of urban travel demand. Transport Policy 17, 127–134.

Tversky, A., Kahneman, D., 1992. Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and uncertainty 5, 297–323.

Wang, M., Deng, W., 2018. Deep visual domain adaptation: A survey. Neurocomputing 312, 135–153.

Wang, S., Mo, B., Zhao, J., 2020a. Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions. Transportation Research Part C: Emerging Technologies 112, 234–251.

Wang, S., Mo, B., Zhao, J., 2021. Theory-based residual neural networks: A synergy of discrete choice models and deep neural networks. Transportation research part B: methodological 146, 333–358.

Wang, S., Wang, Q., Zhao, J., 2020b. Deep neural networks for choice analysis: Extracting complete economic information for interpretation. Transportation Research Part C: Emerging Technologies 118, 102701.

Watling, D.P., Rasmussen, T.K., Prato, C.G., Nielsen, O.A., 2018. Stochastic user equilibrium with a bounded choice model. Transportation Research Part B: Methodological 114, 254–280.

Wehenkel, A., Louppe, G., 2019. Unconstrained monotonic neural networks. Advances in Neural Information Processing Systems 32, 1545–1555.

Wong, M., Farooq, B., 2021. Reslogit: A residual neural network logit model for data-driven choice modelling. Transportation Research Part C: Emerging Technologies 126, 103050.

Xia, Y., Chen, H., Zimmermann, R., 2023. A random effect bayesian neural network (re-bnn) for travel mode choice analysis across multiple regions. Travel Behaviour and Society 30, 118–134.

Yang, H., Bell, M.G., 1997. Traffic restraint, road pricing and network equilibrium. Transportation Research Part B: Methodological 31, 303–314.

Yao, E., Morikawa, T., 2005. A study of on integrated intercity travel demand model. Transportation Research Part A: Policy and Practice 39, 367–381.

You, S., Ding, D., Canini, K., Pfeifer, J., Gupta, M., 2017. Deep lattice networks and partial monotonic functions. Advances in neural information processing systems 30.

Zhao, X., Yan, X., Yu, A., Van Hentenryck, P., 2020. Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. Travel Behaviour and Society 20, 22–35.

Zheng, Y., Wang, S., Zhao, J., 2021. Equality of opportunity in travel behavior prediction with deep neural networks and discrete choice models. Transportation Research Part C: Emerging Technologies 132, 103410.

Zheng, Y., Xu, Z., Xiao, A., 2023. Deep learning in economics: a systematic and critical review. Artificial Intelligence Review , 1–43.