Deep Neural Networks and Discrete Choice Models (Part 1)

Shenhao Wang
191022

Outline

1	2	3	4
Choice Modeling	Predictive ML for Choice Analysis	DNN for Economic Interpretation in Choice Analysis	Interpretability in ML

Part 1. Discrete Choice Analysis

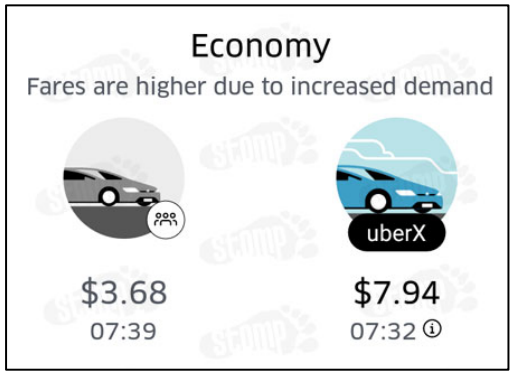
(Compress 1.202 to 40 minutes)

Discrete choice (travel mode choice)

Discrete choice examples in urban transportation

Travel mode choice 

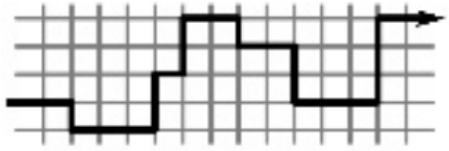
Uber single passenger vs. ride-sharing



Residential & job location choice

Auto ownership 

Trip purpose

Route choice 

etc.

Discrete choice beyond urban transportation

Choice: iPhone vs. Samsung
(economics & marketing)

Choice: Trump vs. Hillary
(political science)

Why is discrete choice analysis (individual decision-making) important?

1. It is everywhere (e.g. choice of class, breakfast, etc.)
2. It is the basic building block of many fields in social science.
3. It can be aggregated to determine more important results (e.g. survival of firms; election results)

Aggregate vs. disaggregate demand analysis

Aggregate demand analysis:

- Urban grids
- Census blocks, census tracts, cities, etc.

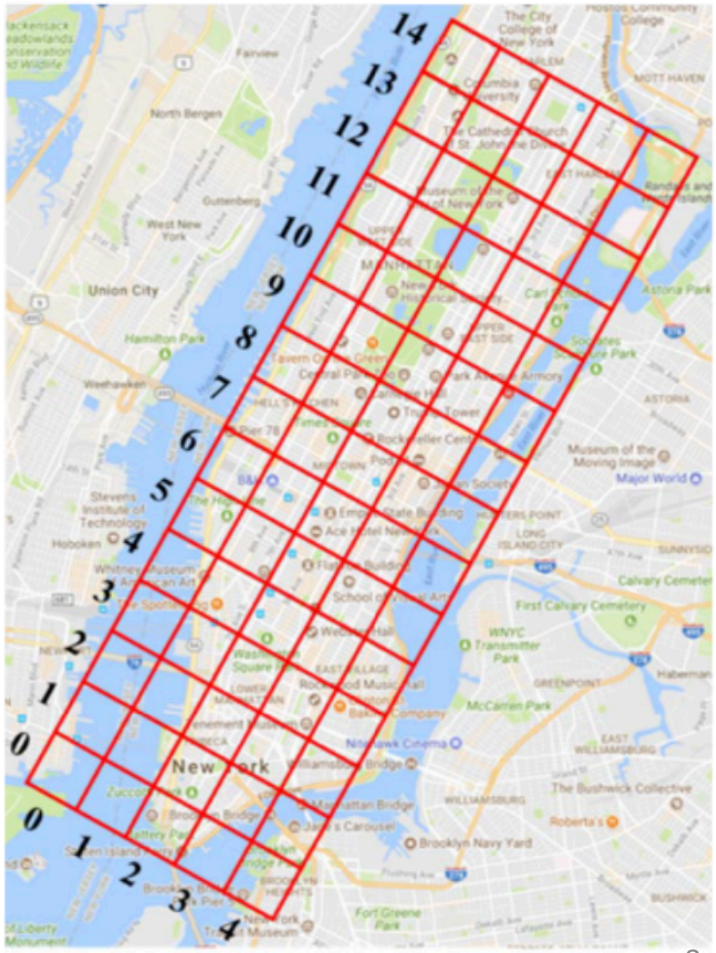
Disaggregate demand analysis (discrete choice):

- Individuals/households

Relationship

- From disaggregate to aggregate: easy
- From aggregate to disaggregate: difficult

Taxi Demand Prediction



Classical analytical framework for discrete choice

Decision-Maker

- Individuals (person/household)
- Socio-economic variables (e.g. age, gender, income)

Alternatives

- Decision-maker i selects one and only one alternative from a choice set $C = \{1, 2, \dots, K\}$
- Collectively exhaustive & mutually exclusive

Attributes of alternatives

- e.g. price, quality

Decision rule

- Utility maximization, etc.

Classical choice models: random utility maximization (RUM)

An example. Choose between bus and auto:

If $U(bus) > U(auto)$, choose bus.

If $U(auto) > U(bus)$, choose auto.

$U(bus) = V(bus) + \epsilon_{bus} = \beta_1 x_{bus\_ivt} + \epsilon_{bus}$

$U(auto) = V(auto) + \epsilon_{auto} = \beta_1 x_{auto\_ivt} + \epsilon_{auto}$

Intuition: compare in-vehicle travel time of taking a bus and driving.

A more generic form:

$$U(k) = \beta_k^T \phi_x(x_k) + w_k^T \phi_z(z) + \epsilon_k; \forall k$$

x_k: alternative-specific variables: price & quality of alternatives.

z: individual-specific variables: income, age, etc.

ϕ_x, ϕ_z: feature transformation (e.g. linear, quadratic, etc.)

Classical choice models: random utility maximization (RUM)

$$
\begin{aligned}
P_i(bus) &= P(U_{bus} \geq U_{auto}) \\
&= P(V_{bus} + \epsilon_{bus} \geq V_{auto} + \epsilon_{auto}) \\
&= P(V_{bus} - V_{auto} \geq \epsilon_{auto} - \epsilon_{bus}) \\
&= P(V_{bus} - V_{auto} \geq \epsilon) \\
&= F_{\epsilon}(V_{bus} - V_{auto})
\end{aligned}
$$

$F_{\epsilon}(V_{bus} - V_{auto})$ is the cumulative distribution function (CDF) of ϵ.

Assume the extreme value distribution of ϵ_{bus} and ϵ_{auto}: $\epsilon_{bus} \sim EV(0, \mu)$, $\epsilon_{auto} \sim EV(0, \mu)$

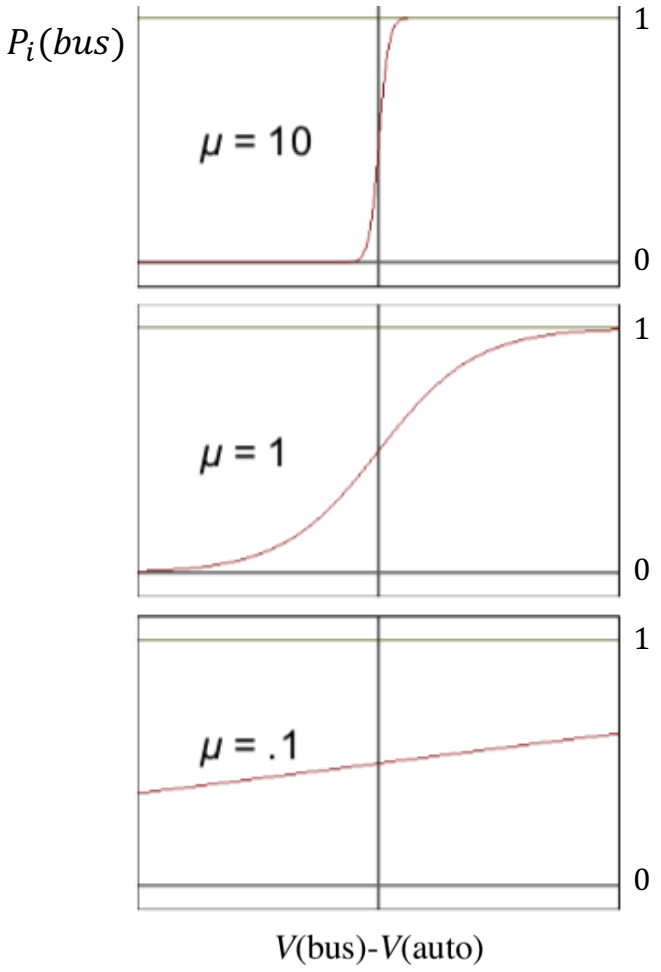
$$
P_i(bus) = \frac{e^{\mu V_{bus}}}{e^{\mu V_{bus}} + e^{\mu V_{auto}}}
$$

Intuition of decision rules in discrete choice models

$$P_i(bus) = \frac{e^{\mu V_{bus}}}{e^{\mu V_{bus}} + e^{\mu V_{auto}}}$$

With different values of μ, the shape of $P_i(bus)$ is different

- $\mu \to +\infty$; deterministic decision making
- $\mu \to 0$; pure random decision making
- "temperature" in ML ("randomness/entropy"; $T = \frac{1}{\mu}$)



Choice probability function in multinomial logit model (MNL)

When the choice set has more than two alternatives, then

$$P_i(k) = \frac{e^{\mu V_k}}{\sum_{j \in C} e^{\mu V_j}}$$

e.g. $C = \{auto, bus, walk\}$

$$P_i(auto) = \frac{e^{\mu V_{auto}}}{e^{\mu V_{auto}} + e^{\mu V_{bus}} + e^{\mu V_{walk}}}$$

Q: Did you see this formula somewhere in the previous lectures?

A: Yes. It is the softmax activation function in DNN! Note the different stories between MNL and DNN.

MNL training: maximum likelihood estimation (MLE)

Maximum likelihood estimation

- Asymptotically efficient

Identification

- Only utility difference matters
- Scale factor does not matter

Optimization

- Convex MLE in MNL

Independence of Irrelevant Alternatives (IIA) in MNL Model

Scenario 1: choose between auto and bus

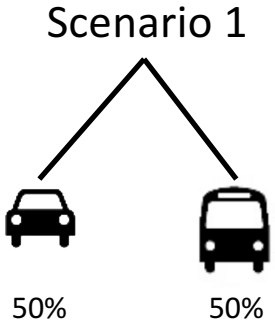
$$P_i(auto) = \frac{e^{\mu V_{auto}}}{e^{\mu V_{bus}} + e^{\mu V_{auto}}};\ P_i(bus) = \frac{e^{\mu V_{bus}}}{e^{\mu V_{bus}} + e^{\mu V_{auto}}};\ \frac{P_i\left(auto \middle| \{auto, bus\}\right)}{P_i\left(bus \middle| \{auto, bus\}\right)} = \frac{e^{\mu V_{auto}}}{e^{\mu V_{bus}}}$$

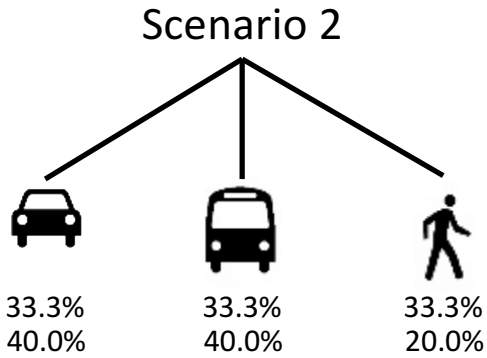
Scenario 2: choose between auto, bus, and walk

$$P_i(auto) = \frac{e^{\mu V_{auto}}}{e^{\mu V_{auto}} + e^{\mu V_{bus}} + e^{\mu V_{walk}}};\ P_i(bus) = \frac{e^{\mu V_{bus}}}{e^{\mu V_{auto}} + e^{\mu V_{bus}} + e^{\mu V_{walk}}};\ \frac{P_i\left(auto \middle| \{auto, bus, walk\}\right)}{P_i\left(bus \middle| \{auto, bus, walk\}\right)} = \frac{e^{\mu V_{auto}}}{e^{\mu V_{bus}}}$$

Independence of Irrelevant Alternatives (IIA) constraint:

$$\frac{P_i(auto|\{auto, bus\})}{P_i(bus|\{auto, bus\})} = \frac{P_i(auto|\{auto, bus, walk\})}{P_i(bus|\{auto, bus, walk\})}$$



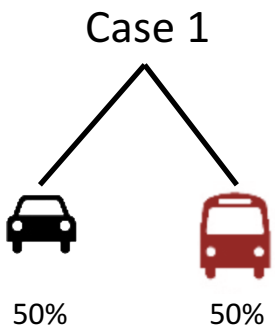


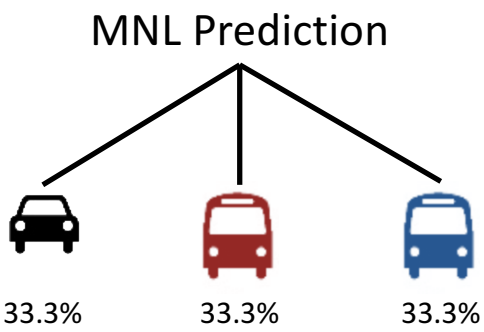
Evaluate Independence of Irrelevant Alternatives (IIA) in MNL Model

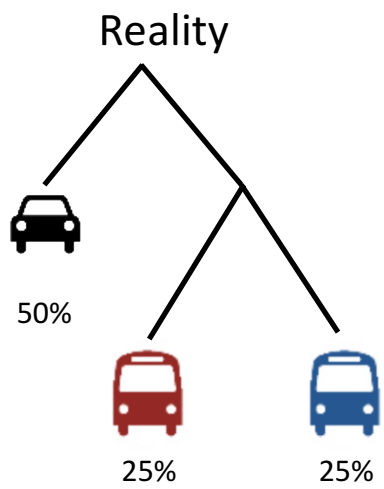
Negative

Against reality: blue and red buses

Motivation for nested logit model







Positive

Model/analytical simplicity

It is one constraint, and will be used later…

What information can we get from MNL?

Economic Information	**Formula in MNL**
Choice prediction	$\text{argmax}_k \, P_i(k)$
Choice probability	$P_i(k)$
Market share	$\sum_i P_i(k)$
Substitution pattern between two alternatives	$P_i(k_1)/P_i(k_2)$
Social welfare	$\sum_i \frac{1}{\alpha_i}[\log \sum_k e^{V_{ik}}] + C$
Probability derivative	$\partial P_i(k)/\partial x_{ij}$
Elasticity	$\frac{\partial P_i(k)/P_i(k)}{\partial x_{ij}/x_{ij}}$
Marginal rate of substitution	$-\frac{\partial P_i(k)/\partial x_{ij_1}}{\partial P_i(k)/\partial x_{ij_2}}$
Value of time (x_{ij_1}: time; x_{ij_2}: price)	$-\frac{\partial P_i(k)/\partial x_{ij_1}}{\partial P_i(k)/\partial x_{ij_2}}$

A full list of economic information from MNL

One view: DNN is a black-box!

Q: To what extent can we extract from DNN models the long list of economic information as in the MNL model?

Part 2. Predicting Individual Choices with ML

Treat choice analysis as a prediction task

$$f: x \rightarrow y; y \in \{1, 2, \ldots K\}$$

Use any predictive ML classifier for choice analysis

Deep neural networks (any architecture, hyper parameters, etc.)
Discriminant analysis
Bayesian model (Naïve Bayesian, etc.)
Support vector machine (any kernel)
K nearest neighbors (KNN)
Decision trees
Random forests
etc.

Research frontier: simply focus on prediction accuracy

Author (Year)	Task	Sample Size	Models	Best Model
Nijkamp et al. (1996) [42]	Travel Mode	1,396	DNN, MNL	DNN
Rao et al. (1998) [48]	Travel Mode	4,335	DNN, MNL	DNN
Hensher and Ton (2000) [28]	Travel Mode	801	DNN, NL	DNN/NL
Xie et al. (2003) [61]	Travel Mode	34,680	DT, DNN, MNL	DNN
Cantarella et al. (2005) [11]	Travel Mode	1,067	DNN, MNL	DNN
Celikoglu (2006) [12]	Travel Mode	N.A.	DNN, RBFNN, GRNN, MNL	RBFNN
Pulugurta et al. (2013) [47]	Travel Mode	5,822	RBM, MNL	RBM
Tang et al. (2015) [53]	Travel Mode	14,000	DT, MNL	DT
Omrani (2015) [43]	Travel Mode	9,500	DNN, RBFNN, MNL, SVM	DNN
Sekhar and Madhu (2016) [50]	Travel Mode	5,000	RF, DT, MNL	RF
Hagenauer and Helbich (2017) [25]	Travel Mode	230,608	MNL, DNN, NB, SVM, CTs, BOOSTING, BAGGING, RF	RF
Tang et al. (2018)	Travel Mode	14,000	DNN	DNN
Wang and Ross (2018) [59]	Travel Mode	51,910	BOOSTING, MNL	BOOSTING
Cheng et al. (2019) [13]	Travel Mode	7,276	RF, SVM, BOOSTING, MNL	RF
Pirra and Dianna (2019) [45]	Travel Mode	39,167	SVM	SVM

Table 1: ML classifiers in past studies; (abbreviations are the same as introduced in Section 1)

Critiques

Lack an empirical benchmark study for predictive ML in choice analysis

Lack the discussions about interpretability, robustness, ML fairness, ML transparency, etc.

Working paper: large-scale analysis as an empirical benchmark study

Predicting Travel Mode Choice with 86 Machine Learning Classifiers: An Empirical Benchmark Study

Shenhao Wang
Baichuan Mo
Jinhua Zhao

Massachusetts Institute of Technology

Abstract

Researchers are applying a large number of machine learning (ML) classifiers to predict travel behavior, but the results are data-specific and the selection of ML classifiers is author-specific. To obtain generalizable results, this paper provides an empirical benchmark by using 86 classifiers from 14 model families to predict the travel mode choice based on the National Household Travel Survey (NHTS) 2017 dataset. The 86 ML classifiers from 14 model families incorporate all the important ML classifiers discussed in previous studies. The large number of observations (about 800,000) in the NHTS2017 dataset enables us to analyze the effect of different sample sizes as a meta-dimension on prediction accuracy. We found that **ensemble models**, including boosting, bagging, and random forests, perform the best among all the classifiers, and that **deep neural networks** (DNNs) perform the best among all the non-ensemble models. Classical **discrete choice models** (DCMs) only predict at the medium or relatively low range

Statement About Prediction: DNNs Outperform classical MNL in Simulations and the Empirical Experiments in the Past Studies (inconclusive)

Note: It depends on model complexity, sample size and completeness of domain knowledge

References: Nijkamp et al., 1996; Rao et al., 1998; Xie et al., 2003; Omrani, 2015; etc.

"A model that consistently predicts accurately must have captured something"

Prediction → Interpretation

Part 3. DNNs for Choice Analysis: Extracting Complete Economic Information for Interpretation

Working paper (under review in TR-Part C)

Shenhao Wang, Qingyi Wang, Jinhua Zhao

Oct 2019

Research purpose: demonstrate how to interpret DNNs for economic information

Strength

1. Completeness

2. Automatic Learning

Challenges

1. Sensitivity to Hyperparameters (statistics)

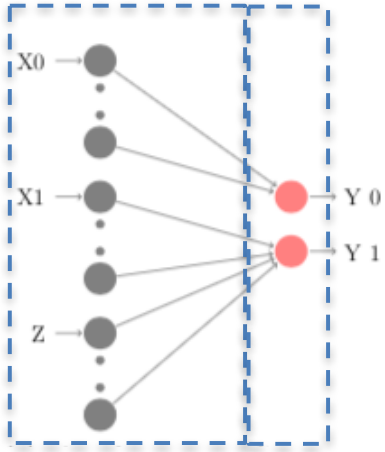
2. Model Non-identification (Optimization)

3. Local Irregularity: exploding gradients & non-monotonicity (Robustness)

Comparing MNL and DNN: MNL and DNN are similar

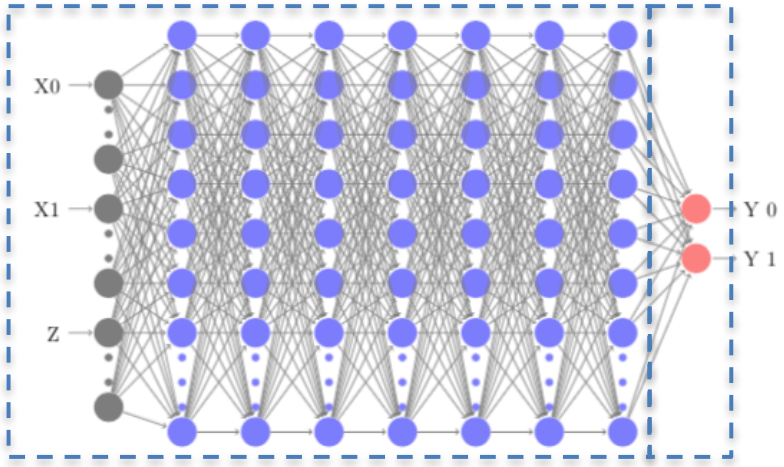
MNL

$$P_i(k) = \frac{e^{\mu V_k}}{\sum_{j \in C} e^{\mu V_j}}$$



DNN

$$P_i(k) = \frac{e^{\Phi_k(x)}}{\sum_{j \in C} e^{\Phi_j(x)}}$$

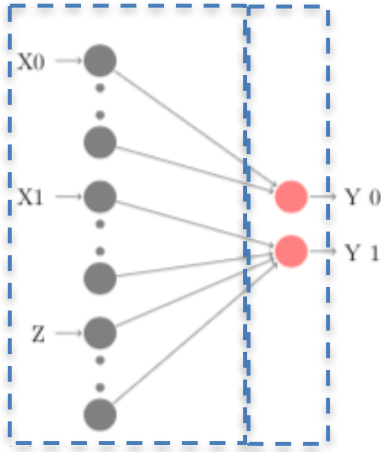


1. Utility specification & comparison exist in both MNL and DNN
2. McFadden (1974) proof about the RUM and softmax activation function.

Comparing MNL and DNN: MNL is a special case of DNNs

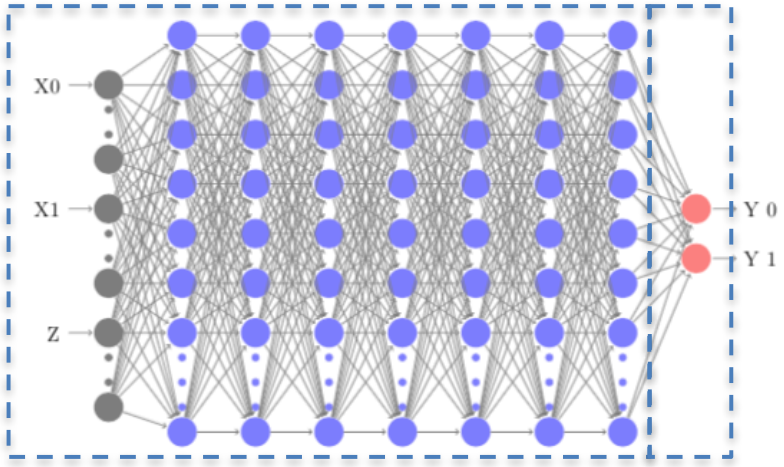
MNL

$$P_i(k) = \frac{e^{\mu V_k}}{\sum_{j \in C} e^{\mu V_j}}$$



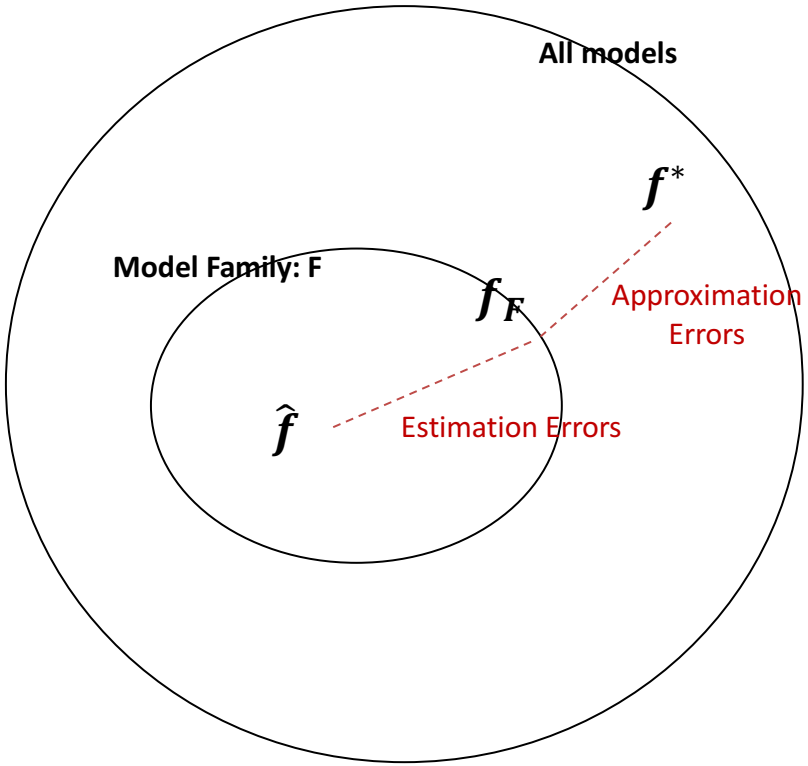
DNN

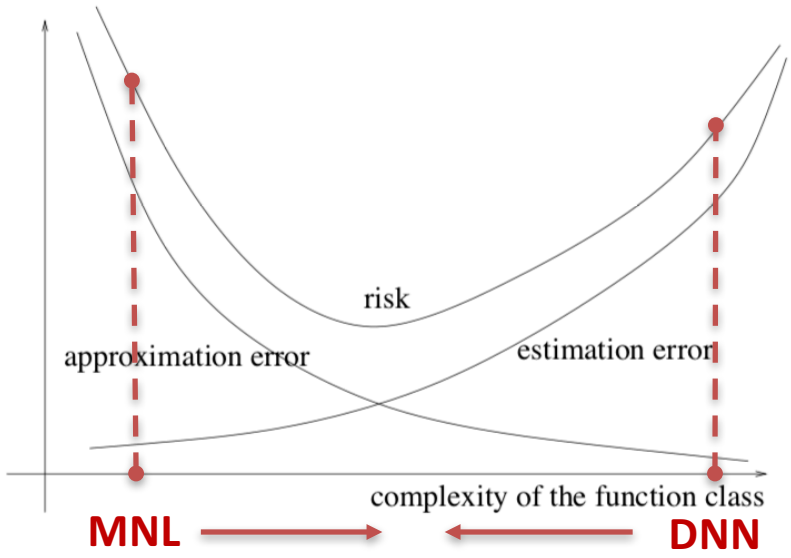
$$P_i(k) = \frac{e^{\Phi_k(x)}}{\sum_{j \in C} e^{\Phi_j(x)}}$$



1. Visual Intuition: SNN vs. DNN
2. DNN: universal approximator theorem

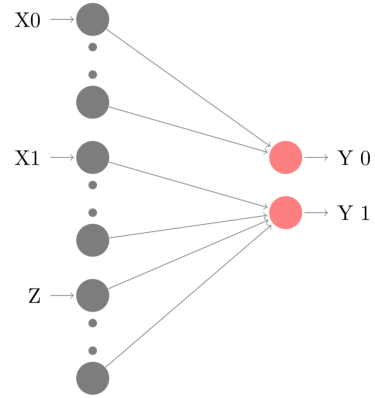
Statistical learning theory:
Prediction errors = approximation + estimation errors



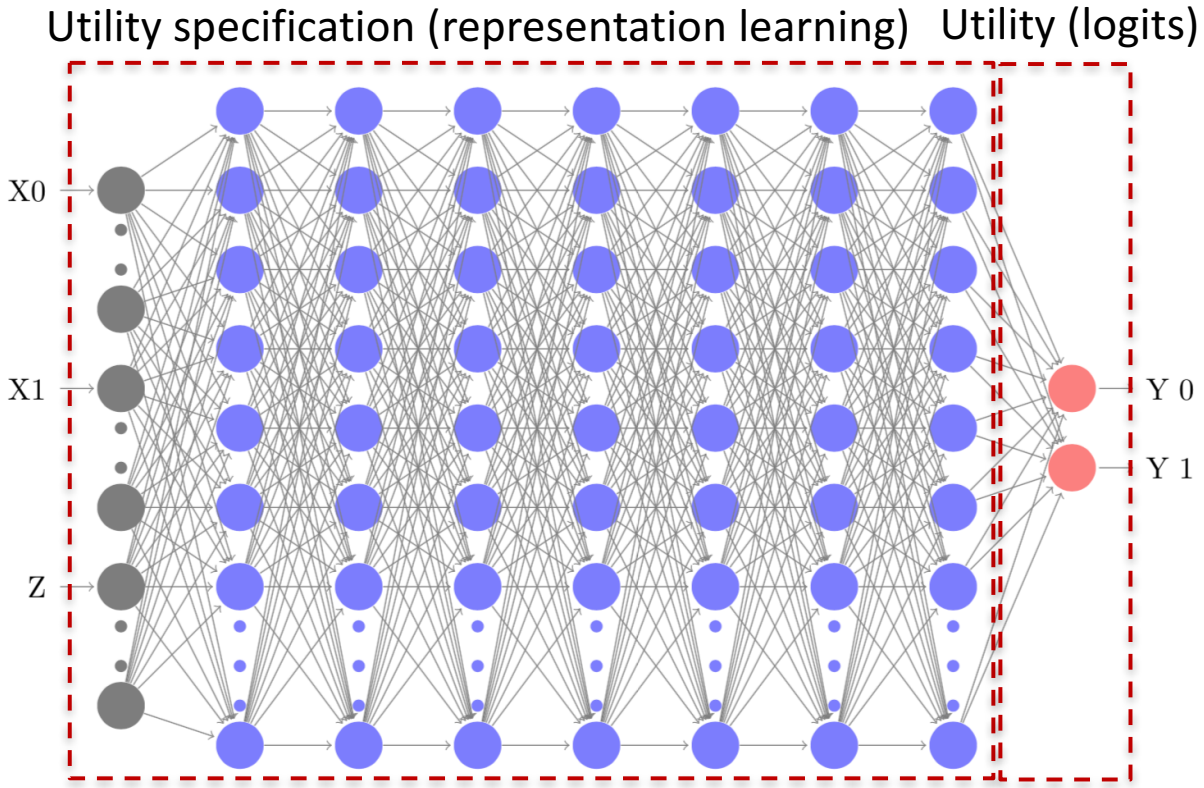


1. How to improve MNL? Increase the model complexity
2. How to improve DNN? Reduce the model complexity

Extract economic information from DNN in choice analysis



MNL with linear specification



DNN for Binary Choice

Choice probability functions [Soft labels]: **s(x)**

Utility functions [Logits]: **V(x)**

(Notations are changed to those in the working paper)

Derive economic information from DNNs as complete as DCMs

By using

1. Functions: **s(x) or V(x)**
2. Gradients of the Functions

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

A full list of economic information

Experiment setup

Comparing three model groups:

- HP-DNNs (100 models from the hyperparameter searching)
- DNNs (100 repeated trainings of DNNs with the best hyperparameter)
- MNLs

Predicting travel mode choice (walking, public transit, ridesharing, AV, and driving)

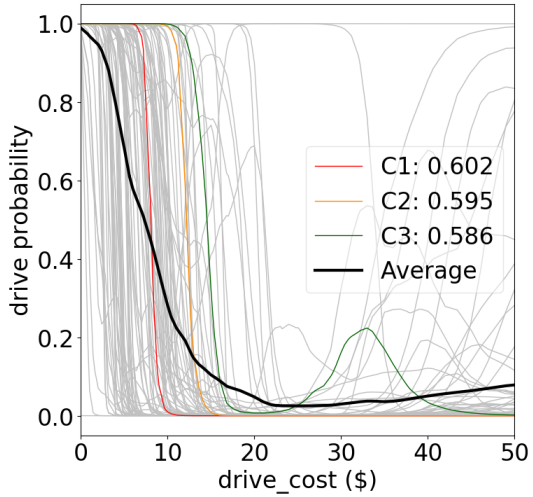
Sample size: 8,418

Discussing each piece of economic information

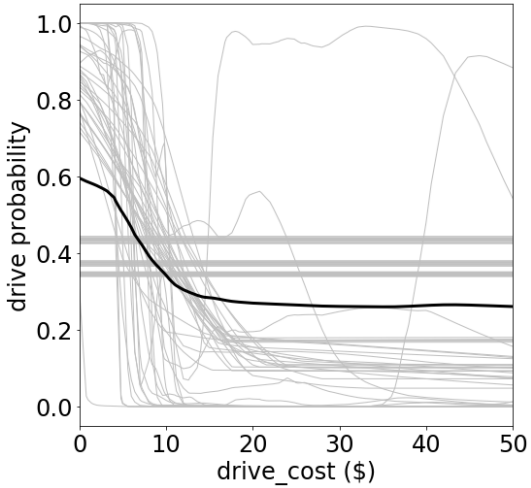
Choice Probability Functions
[Soft Labels]

$$s_k\left(x_j; x_{\setminus j}\right);\ k = Driving$$

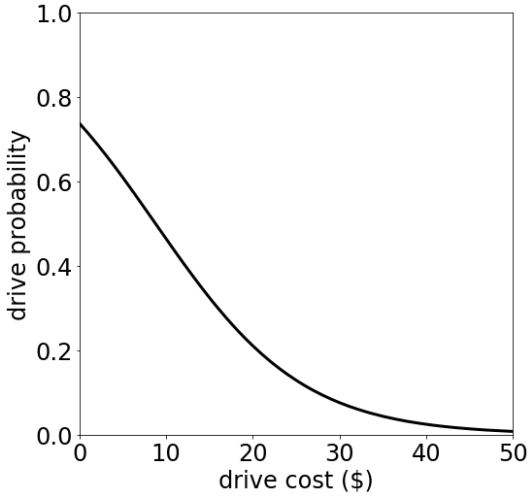
DNNs



HP-DNNs

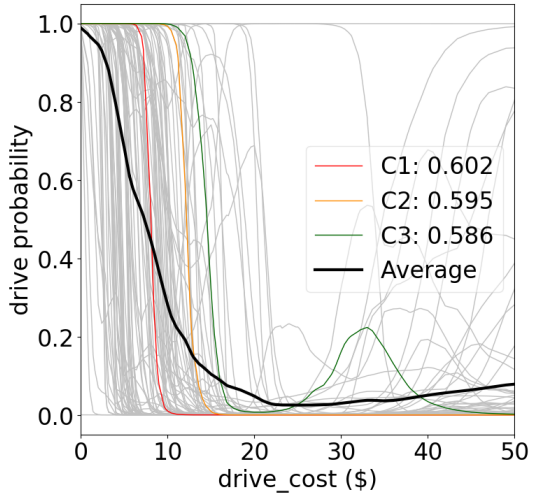


MNL



Side remark: risks caused by the local irregularity (the red curve)

DNNs



DNN application example:

- Public schools hire/fire teachers based on DNN prediction.

Consequence:

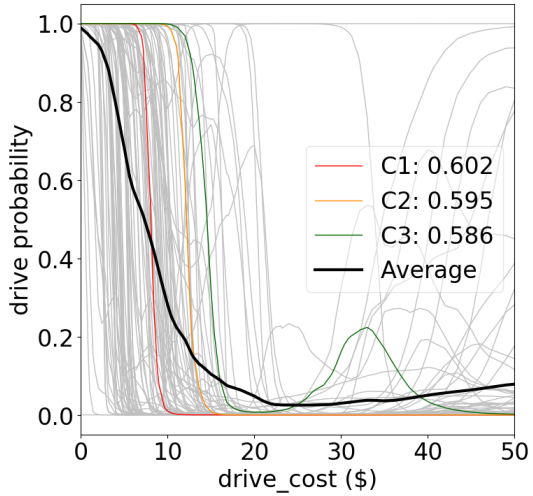
- Teachers are fired due to ϵ perturbation. E.g. different colors of shirts.

Important research frontier

- Deployable ML

Side Q1: Did you see this local irregularity (the red curve) in previous lectures?

DNNs



Lecture 4 Justin Dauwels

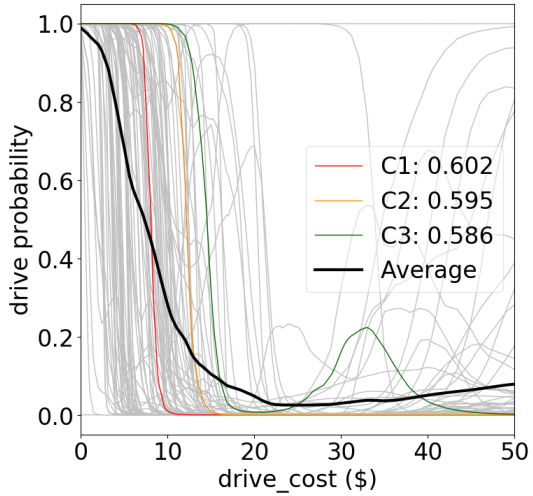
"panda"
57.7% confidence

"gibbon"
99.3% confidence

Gradients and robustness perspectives are two sides of the same coin.

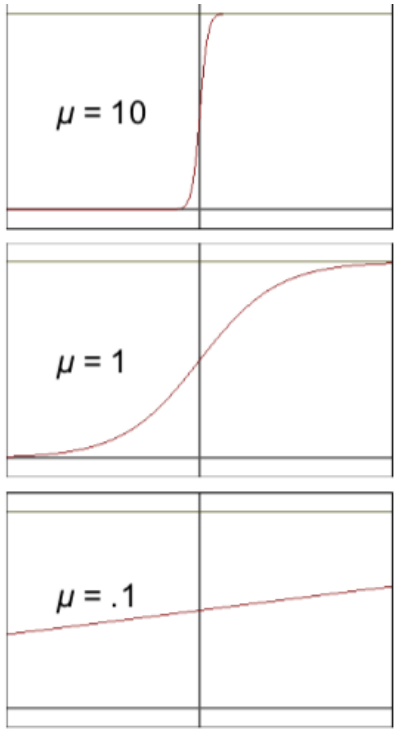
Side Q2: Did you see this local irregularity (the red curve) in the classical choice models in today's lecture?

DNNs



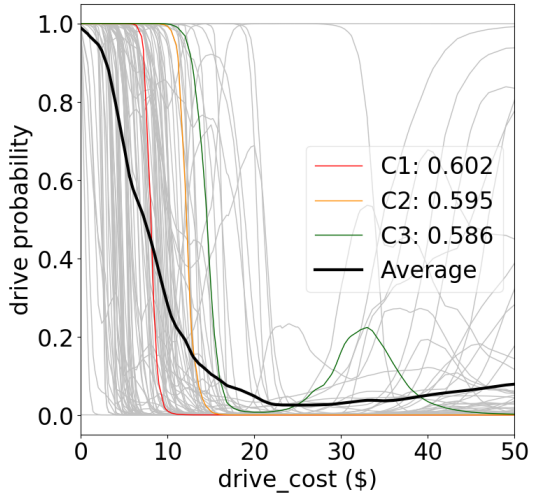
MNL

$$P_i(bus) = \frac{e^{\mu V_{bus}}}{e^{\mu V_{bus}} + e^{\mu V_{auto}}}$$



Side Q3: Do you think whether the "local irregularity" (the red curve) is reasonable/intuitive/correct?

DNNs



Yes, with three perspectives:

MNL perspective:

- MNL with different temperatures

Decision-making perspective:

- Deterministic decision rule

ML perspective

- Decision tree

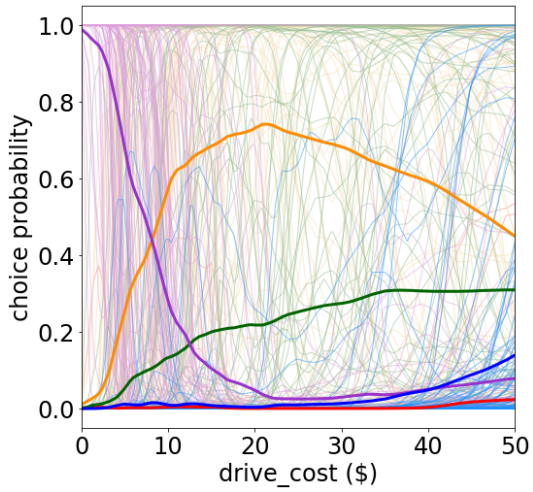
How to formally answer this question?

- Need to answer it with simulations

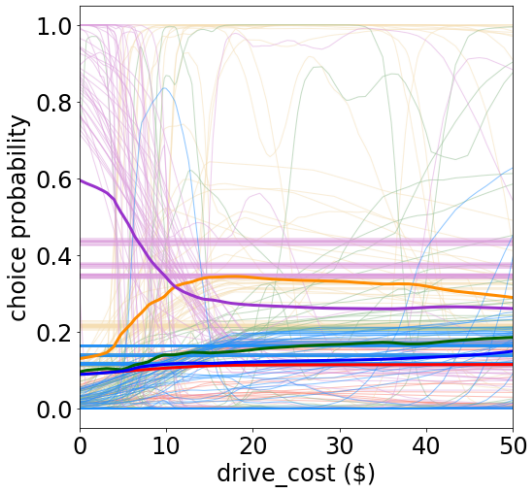
Substitution Patterns of Five Alternatives

$$s_k\left(x_j; x_{\setminus j}\right), k \in \{1, \ldots 5\}$$

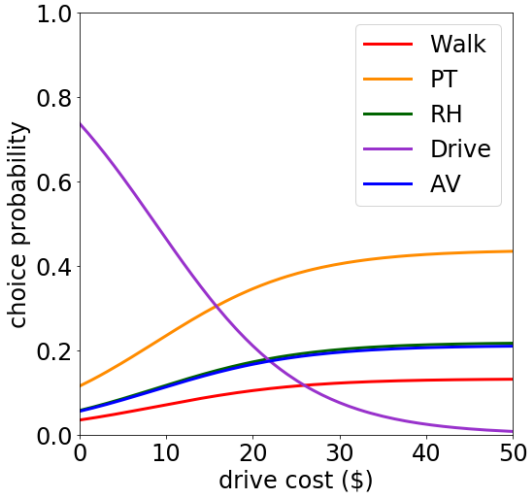
DNNs



HP-DNNs



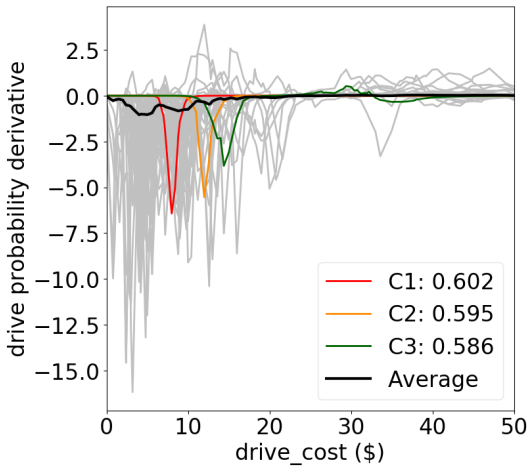
MNL



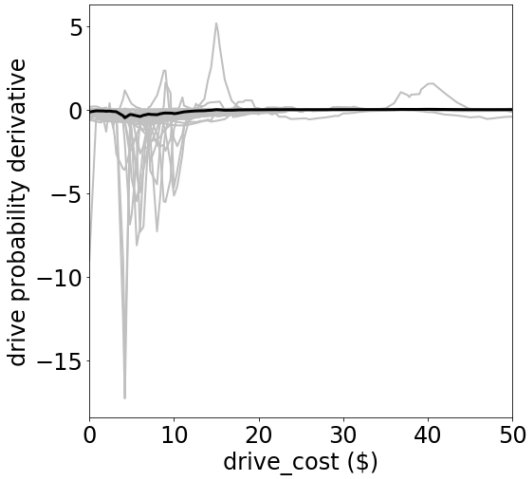
Gradients of Choice Probability Functions

$\partial s_k(x)/\partial x_j$

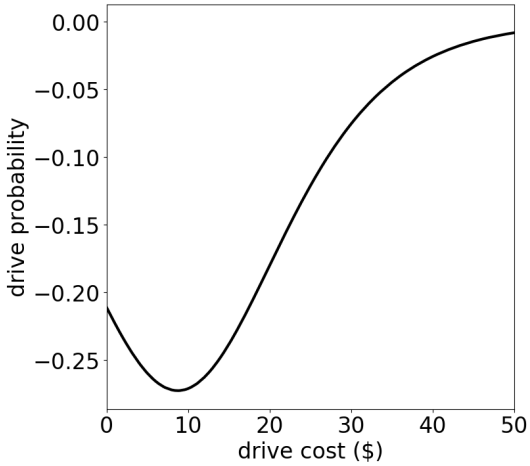
DNNs



HP-DNNs



MNL



Elasticities

$$\frac{\partial s_k(x)/s_k(x)}{\partial x_j/x_j}$$

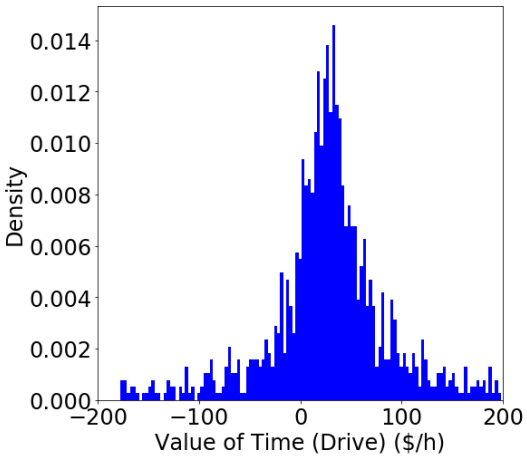
DNN

	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)

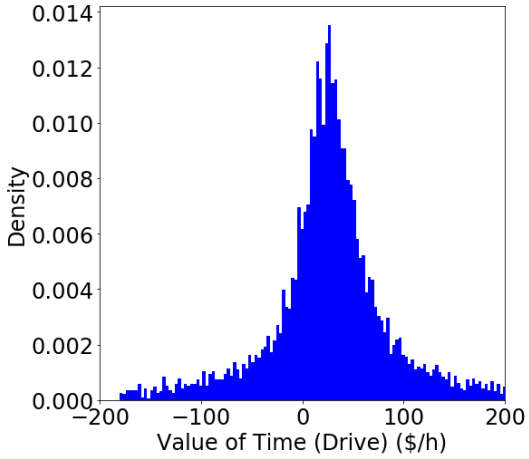
Heterogeneous VOT

$$\frac{\partial s_k(x)/\partial x_{j_1}}{\partial s_k(x)/\partial x_{j_2}}$$

DNN (Testing)



DNN (Training)



Other Economic Information

MRS

Market share

Social welfare

etc.

How to address the three problems?

Challenges	Theoretical Foundations	Solutions
High sensitivity to hyper-parameters	**Statistical challenge: balance approximation and estimation errors**	1. **Regularization methods**: model ensemble (Krizhevsky et al. 2012), data augmentation, dropouts (Hinton et al. 2012), early stopping; 2. **Architectural design**: AlexNet (Krizhevsky, et al. 2012), GoogleNet (Szegedy et al. 2015), and ResNet (He et al. 2016) 3. **Automatic hyperparameter tuning**: Bayesian neural networks (Snoek 2012, Snoek 2015), or reinforcement learning (Zoph 2016, Zoph 2017) 4. **Statistical learning theory** (Vapnik 2013, Bartlett 2002, Bartlett 2017, Neyshabur 2015, Golowich 2017)
Model non-identification	**Optimization challenge: find global optimum**	**One new perspective: it does not matter…**
Local irregularities	**Robustness challenge: stabilize local information**	1. **Robust training methods**: adversarial training (Kurakin et al. 2016), defensive knowledge distillation (Papernot et al. 2016), mini-max robust training (Madry et al. 2017) 2. **Monotonicity constraints**: (Gupta et al. 2016)

Contribution of this paper

1. First paper to demonstrate the relationship between MNL and DNN for choice analysis through utility interpretation

2. Derive complete economic information from DNN (not a black box!)

3. Reveal the potential risks involved in DNN-based choice models

4. Connect the risks to three theoretical challenges

5. Point out future directions of improving DNN for choice analysis

Generic open-ended questions

Q: Is the automatic learning only rhetoric? (MNL is also automatic learning.)

A: Yes. The key is the universal approximator theorem.

Q: There are many universal approximators. Why DNN?

A: it is subtle…

Q: If DNN performs so badly with ϵ perturbation, how can it have very high prediction accuracy?

A: …

Q: To improve DNN, we need to regularize the model to make it simpler. Is this argument against the spirit of going deeper?

A: …