Unravelling the role of deep learning optimization in behavioural models

Melvin Wong

Laboratory of Innovations in Transportation
Ryerson University
melvin.wong@ryerson.ca

EPFL Seminar
October 21, 2019
Outline

1. Introduction

2. Topic I: Deep learning optimization
   - Multiple Discrete-Continuous Models
   - Econometric analysis

3. Topic II: Modelling behaviour heterogeneity through DNNs
   - Residual Logit model
   - Empirical Examples

4. Summary
   - Recent papers

Wong, M.  DL Optimization in Choice Models
My Research Topics

- Incorporating modern machine learning methods into conventional econometric models
  - e.g. deep learning and neural networks
- Explaining ‘deep learning’
  - model interpretability
  - econometric analysis
- Data modelling
  - Optimization of complex behavioural models
  - ‘Big Data’ analysis
  - High performance computation
  - Open-source tools for implementing deep learning algos in choice models
Trends

Increasing appeal of artificial neural network (ANN) based models among travel behaviour modelling researchers

- Search terms "machine learning" OR "deep learning" OR "artificial neural network" on Transportation Journals
- Presumptive gain in model performance (accuracy, prediction, etc.) and usability
- Automation of everything
Trends

Increasing appeal of artificial neural network (ANN) based models among travel behaviour modelling researchers

- Search terms "machine learning" OR "deep learning" OR "artificial neural network" on Transportation Journals
- Presumptive gain in model performance (accuracy, prediction, etc.) and usability
- Automation of everything

Skepticism and assumptions:

- “A black-box model”
- “Gains in predictive accuracy, but loss in general interpretability”
Increasing appeal of artificial neural network (ANN) based models among travel behaviour modelling researchers

- Search terms "machine learning" OR "deep learning" OR "artificial neural network" on Transportation Journals
- Presumptive gain in model performance (accuracy, prediction, etc.) and usability
- Automation of everything

Skepticism and assumptions:

- “A black-box model”
- “Gains in predictive accuracy, but loss in general interpretability”

Common misconception: Compared to logit as yet another variant of a statistical modelling
Where are we now?

The hype cycle

- **Introduction**
- **Visibility**
- **Abandonment**
- **Inflated expectations**
- **Delusion**
- **Productivity peak**
- **Learning curve**

Hype Phase  |  Realization Phase
Introduction - What is deep learning?
Introduction - What is deep learning?

In the context of DCA, **deep learning** is:

- the optimization of complex (behavioural) models by emulating the process of human learning and memory
- Combine concepts from behavioural science and information theory
In Neuroscience:

- “The human brain is constantly making predictions (inferences) and updating its beliefs to minimize uncertainty” [Friston, 2009]
- **Helmholtz Machine** [Dayan et al., 1995] → precursor to deep learning
Introduction

In Information theory:

- Decision makers are subject to \textit{information costs} in the learning process
- Information can be measured by \textit{entropy}
- Choice probabilities are given by the \textit{Boltzmann distribution} [Anas, 1983]

Information is how much \textit{uncertainty} is there about an event
Overview of this presentation

Topic I: Deep learning optimization
Topic II: Modelling behaviour heterogeneity through DNNs
Outline

1. Introduction

2. Topic I: Deep learning optimization
   - Multiple Discrete-Continuous Models
   - Econometric analysis

3. Topic II: Modelling behaviour heterogeneity through DNNs
   - Residual Logit model
   - Empirical Examples

4. Summary
   - Recent papers

Wong, M. DL Optimization in Choice Models
Overview
Applications in technology

Deep learning optimization

- Dealing with intractable problems in optimization through DL algos
- Breakthroughs in machine learning algorithms, e.g. generative models (GANs, VAE, RBM)
Deep learning optimization

- Dealing with intractable problems in optimization through DL algs
- Breakthroughs in machine learning algorithms, e.g. generative models (GANs, VAE, RBM)
- Applications: Recommender systems, image superimposition aka ‘deepfakes’, etc.
DL optimization as a psychological process
Integration of psychological factors into decision models

Examples:
- Decision protocols, choice sets, unobserved taste variations, latent information

Explicitly defined in most choice models:
- Choice set sampling
- Latent Class Model
- ICLV Model
- Random Regret Minimization

However, the true underlying behavioural processes are unknown, and constantly changing.
Examples

Applications in discrete choice modelling

Multiple Discrete-Continuous Models

- MDCEV
- Mixed Integer Linear Programming
Examples
Applications in discrete choice modelling

Deep learning optimization approach

- We introduced a *generative model framework* to solve the intractable objective function in MDC choice model estimation

Examples
Applications in discrete choice modelling
DL Optimization as a psychological process
Integration of psychological factors into decision models

Free-energy lower bound objective function:

\[ F = -\ln p(D) + D_{KL}(q(s)||p(s|D)) \]

Generative model that yields an optimal solution for \( q(s) \) by gradient ascent over \( F \)

- Inference becomes an optimization problem
- Solution is by optimizing an approximate lower bound

- \( D \): data
- \( s \): unobserved states
- \(-\ln p(D)\): Self-information or Entropy
- \( D_{KL} \): Kullback-Leibler divergence
Understanding the free-energy lower bound

Thermodynamic intuition

- Information is modelled as state changes utilized in thermodynamic systems
- Difference between two ‘energy’ states
- Corresponds to the utility gain in economic choice models with an information processing constraint [Ortega and Braun, 2013]

Variational Bayesian inference

- Optimization problem can be framed in terms of well-known variational Bayesian inference used in deep learning
- Provides a lower bound on an intractable function
Variational density and estimation

Assumption: approximating variational density can be factorized

\[ q(s) = \prod_h q_\theta(s_h) \approx \prod_h p(s_h|D) \]

In contrast to mixture models where densities are additive:

\[ q(s) = \sum_h a_h q_\theta(s_h); \quad \sum_h a_h = 1 \]

The goal is to solve:

\[ q_\theta^*(s) = \arg\min_{q_\theta} D_{KL}(q_\theta^*(s)\|p(s|D)) \]

\[ = \arg\max \left[ \mathbb{E}_q\{\ln p(D, s)\} - \mathbb{E}_q\{\ln q_\theta^*(s)\} \right] \]
Econometric analysis

Elasticity

\[ \varepsilon = \frac{Jp_n(x)x_n}{p_n(x)} = \frac{\partial p_n(x)}{p_n(x)} \cdot \frac{x_n}{\partial x_n} \]

The Jacobian matrix \( Jp_n(x) \) of each observation \( n \), for each fixed input vector \( x \) is defined as the backpropagation derivative w.r.t \( p_n \):

Let output \( p_n(x) = g(W_1 \cdot h(W_0 \cdot x_n)) \), then

\[
Jp_n(x) = \frac{\partial p_n(x)}{\partial x_n} = \left[ \frac{\partial p(x)_1}{\partial h_1}, \ldots, \frac{\partial p(x)_1}{\partial h_s} \right] \cdot \left[ \frac{\partial h_1}{\partial x_1}, \ldots, \frac{\partial h_1}{\partial x_n} \right] = \binom{\frac{\partial h_1}{\partial x_1}, \ldots, \frac{\partial h_1}{\partial x_n}}{\vdots} \cdot \binom{\frac{\partial h_s}{\partial x_1}, \ldots, \frac{\partial h_s}{\partial x_n}} \cdot \frac{\partial h}{\partial x_n}
\]
Econometric analysis

Elasticity

walk: $\mu = -0.08$, $\sigma = 0.34$

cycling: $\mu = 0.22$, $\sigma = 0.4$

driving: $\mu = 0.63$, $\sigma = 0.53$

walk: $\mu = -0.22$, $\sigma = 0.4$

cycling: $\mu = 0.4$, $\sigma = 0.53$

driving: $\mu = -0.63$, $\sigma = 0.53$

Transit: $\mu = 0.42$, $\sigma = 0.47$

Driving + Transit: $\mu = -0.53$, $\sigma = 0.61$
## Econometric analysis

### variable pair correlation

<table>
<thead>
<tr>
<th>Variable pair</th>
<th>Original data</th>
<th>H=5</th>
<th>H=25</th>
<th>H=100</th>
</tr>
</thead>
<tbody>
<tr>
<td>mode-purpose</td>
<td>-0.0961</td>
<td>-0.1149</td>
<td>-0.1002</td>
<td>-0.0954</td>
</tr>
<tr>
<td>mode-distance</td>
<td>-0.1884</td>
<td>-0.4439</td>
<td>-0.2269</td>
<td>-0.1919</td>
</tr>
<tr>
<td>mode-time</td>
<td>0.0349</td>
<td>-0.0244</td>
<td>0.0667</td>
<td>0.0382</td>
</tr>
<tr>
<td>purpose-distance</td>
<td>-0.1396</td>
<td>-0.2846</td>
<td>-0.1549</td>
<td>-0.1453</td>
</tr>
<tr>
<td>purpose-time</td>
<td>-0.1039</td>
<td>-0.3866</td>
<td>-0.1504</td>
<td>-0.1052</td>
</tr>
<tr>
<td>distance-time</td>
<td>0.4777</td>
<td>0.8247</td>
<td>0.5715</td>
<td>0.4907</td>
</tr>
<tr>
<td>avg. difference</td>
<td>-</td>
<td>0.07</td>
<td>-0.004</td>
<td>-0.001</td>
</tr>
</tbody>
</table>
Outline

1 Introduction

2 Topic I: Deep learning optimization
   - Multiple Discrete-Continuous Models
   - Econometric analysis

3 Topic II: Modelling behaviour heterogeneity through DNNs
   - Residual Logit model
   - Empirical Examples

4 Summary
   - Recent papers

Wong, M. DL Optimization in Choice Models
Feedforward deep neural networks

Problems

- Highly non-convex (identification problem)
- Arbitrary hyperparameter choices
- Inflexible
- Difficulty obtaining econometric parameters, elasticities

A typical example of a feedforward deep neural network
Multi-layer perceptron (MLP) model

Feedforward model:

\[ h_1 = \sigma(\langle \omega_1, x \rangle + b_1) \]
\[ h_2 = \sigma(\langle \omega_2, h_1 \rangle + b_2) \]
\[ \vdots \]
\[ p(i|C) = \frac{e^{\langle \beta_i, h_L \rangle_i + \varepsilon_i}}{\sum_{j \in C} e^{\langle \beta_j, h_L \rangle_j + \varepsilon_j}} \]

- \( x \): inputs
- \( \sigma \): activation function; sigmoid, relu, etc.
- \( \omega, b \): neural network parameters
- \( \langle \cdot \rangle \): matrix multiplication

\( \beta \) becomes an arbitrary parameter (obtain no econometric information about \( x \))
Problems with modelling deep neural nets

Assumptions

- Ability to capture non-linearity by introducing multiple neurons and multiple non-linear transform layers

Overfitting

- Increasing the number of layers (and increasing non-linearity) may lead to worse performance than a simpler model (e.g. MNL)
- Contradicts logical intuition that a \((N + 1)\)-layer neural net should, in theory, perform better than a \(N\)-layer neural net

Usability

- Not configured for discrete choice analysis
- How to relate it to behaviour and information theory?
ResLogit: A data driven deep neural net model for discrete choice

Conceptualization

- Increasing model depth may lead to performance degradation [He, 2015]
- Instead of stacking layers, use shortcut residuals that learns the unobserved utility component

- 6th International Choice Modelling Conference
  - Special Session 6: Machine Learning and Spatialtemporal Choice Modelling
- TRB Annual Meeting 2020
- Manuscript ready for journal submission within the month
Network Architecture

(a) MLP; (b) MNL; (c) ResLogit
Framing Residual NN as a dynamical system

A dynamical system is defined as:

\[ h_t = f(h_{t-1}; \omega_t) + x_t \]

In a ResLogit model:

\[
\begin{align*}
    h_{t=1} &= f(V; \omega_1) + V; \quad V = \sum_m \beta_{im} x_m \\
    h_{t=2} &= f(h_{t=1}; \omega_2) + h_{t=1} \\
    &\vdots \\
    U &= V + f(h_{T-1}; \omega_T) + f(h_{T-2}; \omega_{T-1}) + \ldots + f(V; \omega_1) + \varepsilon \\
    p(i|C) &= \frac{e^{v_i + \varepsilon_i}}{\sum_{j \in C} e^{v_j + \varepsilon_j}}
\end{align*}
\]

\(\beta\) parameters retain econometric interpretability
Model interpretability

What about $\omega$?

\[
f(V; \omega) = -\ln(1 + \exp(\langle \omega, V \rangle))
\]

\[
= -\ln \left( 1 + \exp \left( \begin{bmatrix} v_1 \\ \vdots \\ v_i \end{bmatrix} \begin{bmatrix} \omega_{11} & \cdots & \omega_{1j} \\ \vdots & \ddots & \vdots \\ \omega_{i1} & \cdots & \omega_{ij} \end{bmatrix} \right) \right)
\]

$\omega$ is a $j \times j$ matrix that represents the correlation between utilities of each alternative.

When $\omega$ is an **Identity matrix**, the ResLogit model collapses into a standard MNL model.

- Identity matrix meaning correlation between utilities $= 0$
- Fully IIA-RUM MNL model
Red bus/blue bus example

Residual function: \( f_1 = -\ln(1 + \exp(\langle \omega, V \rangle)) \);
Residual matrix:

\[
\begin{bmatrix}
\text{car} & \text{rbus} & \text{bbus} \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
\end{bmatrix}
\rightarrow
\begin{bmatrix}
\text{car} & \text{rbus} & \text{bbus} \\
0 & -1 & -1 \\
-1 & 0 & 1.5 \\
-1 & 1.5 & 0 \\
\end{bmatrix}
\]
Red bus/blue bus example

Residual function: \( f_1 = -\ln(1 + \exp(\langle \omega, V \rangle)) \);

Residual matrix:

\[
\omega_{\text{init}} = \begin{bmatrix}
\text{car} & \text{rbus} & \text{bbus} \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
\end{bmatrix} \rightarrow \omega_{\text{final}} = \begin{bmatrix}
\text{car} & \text{rbus} & \text{bbus} \\
0 & -1 & -1 \\
-1 & 0 & 1.5 \\
-1 & 1.5 & 0 \\
\end{bmatrix}
\]

<table>
<thead>
<tr>
<th>Choice</th>
<th>( v_j )</th>
<th>( f_1(v_1, \ldots, v_j) )</th>
<th>( v_j + f_1(v_1, \ldots, v_j) )</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scenario 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bus</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>car</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td><strong>Scenario 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>red bus</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.33</td>
</tr>
<tr>
<td>blue bus</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.33</td>
</tr>
<tr>
<td>car</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.33</td>
</tr>
<tr>
<td><strong>Scenario 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>red bus</td>
<td>1</td>
<td>-0.974</td>
<td>0.026</td>
<td>0.23</td>
</tr>
<tr>
<td>blue bus</td>
<td>1</td>
<td>-0.974</td>
<td>0.026</td>
<td>0.23</td>
</tr>
<tr>
<td>car</td>
<td>1</td>
<td>-0.127</td>
<td>0.873</td>
<td>0.54</td>
</tr>
</tbody>
</table>
Properties of ResLogit

Relation to GEV model formulation

In GEV models [McFadden, 1978], the probability is given by:

\[
P(i|C) = \frac{y_i G_i(y_1, \ldots, y_j)}{\mu G(y_1, \ldots, y_j)} = \frac{e^{v_i + \ln G_i}}{\sum_{j \in C} e^{v_j + \ln G_j}}
\]

- \( G \) term used to accommodate different forms of heterogeneity
  - Nested Logit, Cross-Nested Logit, etc.

ResLogit follows a ‘Universal Logit’ formulation [McFadden, 1975]
- Random terms remain i.i.d extreme value distributed
- non-RUM (cross-alternative attributes are used)
Properties of ResLogit

On Information Theory

Shannon information capacity $C$:
A bound on the information rate of data with the minimum error

$$C = -\ln \left(1 + \frac{S}{N}\right)$$

utility = observed utility + information capacity

- Rational Inattention Model [Fosgerau et al., 2017] in discrete choice
- “Individuals must choose among discrete alternatives with imperfect information about their values” [Matějka and McKay, 2015]
ResLogit vs MLP model estimation
Training and validation curves

Left: MNL vs MLP-DNN; Right: MNL vs ResLogit
## ResLogit vs MLP model performance

Loglikelihood and validation accuracy

<table>
<thead>
<tr>
<th>Model Depth</th>
<th>MNL Training LL</th>
<th>MLP 2</th>
<th>MLP 4</th>
<th>MLP 8</th>
<th>MLP 16</th>
<th>ResLogit 2</th>
<th>ResLogit 4</th>
<th>ResLogit 8</th>
<th>ResLogit 16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training LL</td>
<td>-38790</td>
<td>-37208</td>
<td>-37820</td>
<td>-38496</td>
<td>-42240</td>
<td>-40217</td>
<td>-32342</td>
<td>-30592</td>
<td>-31887</td>
</tr>
<tr>
<td>Validation LL</td>
<td>-16145</td>
<td>-15583</td>
<td>-15894</td>
<td>-16736</td>
<td>-16667</td>
<td>-13675</td>
<td>-13583</td>
<td>-12870</td>
<td>-13121</td>
</tr>
<tr>
<td>Validation accuracy</td>
<td>72.01%</td>
<td>72.57%</td>
<td>71.91%</td>
<td>69.83%</td>
<td>69.5%</td>
<td>75.18%</td>
<td>76.15%</td>
<td>77.29%</td>
<td>76.73%</td>
</tr>
</tbody>
</table>
## ResLogit vs MLP model performance

Loglikelihood and validation accuracy

<table>
<thead>
<tr>
<th>Model Depth</th>
<th>MNL</th>
<th>MLP 2</th>
<th>MLP 4</th>
<th>MLP 8</th>
<th>MLP 16</th>
<th>ResLogit 2</th>
<th>ResLogit 4</th>
<th>ResLogit 8</th>
<th>ResLogit 16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training LL</td>
<td>-38790</td>
<td>-37208</td>
<td>-37820</td>
<td>-38496</td>
<td>-42240</td>
<td>-40217</td>
<td>-32342</td>
<td>-30592</td>
<td>-31887</td>
</tr>
<tr>
<td>Validation LL</td>
<td>-16145</td>
<td>-15583</td>
<td>-15894</td>
<td>-16736</td>
<td>-16667</td>
<td>-13675</td>
<td>-13583</td>
<td>-12870</td>
<td>-13121</td>
</tr>
<tr>
<td>Validation accuracy</td>
<td>72.01%</td>
<td>72.57%</td>
<td>71.91%</td>
<td>69.83%</td>
<td>69.5%</td>
<td>75.18%</td>
<td>76.15%</td>
<td>77.29%</td>
<td>76.73%</td>
</tr>
</tbody>
</table>
Outline

1 Introduction

2 Topic I: Deep learning optimization
   - Multiple Discrete-Continuous Models
   - Econometric analysis

3 Topic II: Modelling behaviour heterogeneity through DNNs
   - Residual Logit model
   - Empirical Examples

4 Summary
   - Recent papers
Summary

- Deep learning (or machine learning) is not just another glorified probabilistic classification model.
- We presented an intuitive method that allows deep learning optimization of intractable problems.
- Representing behavioural models through a residual neural network structure.
- Numerous applications of deep learning in discrete choice analysis to be explored.
Recent papers

Deep learning optimization applications in choice modelling

Mixed Logit


Population synthesis


Multiple Discrete-Continuous data

Information processing constraints in travel behaviour modelling: A generative learning approach

Wong M., Farooq, B.

Abstract: Travel decisions tend to exhibit sensitivity to uncertainty and information processing constraints. These behavioural conditions can be characterized by a generative learning process. We propose a data-driven generative model version of rational inattention theory to emulate these behavioural representations. We outline the methodology of the generative model and the associated learning process as well as provide an intuitive explanation of how this process captures the value of prior information in the choice utility specification. We demonstrate the effects of information heterogeneity on a travel choice, analyze the econometric interpretation, and explore the properties of our generative model. Our findings indicate a strong correlation with rational inattention behaviour theory, which suggest that individuals may ignore certain exogenous variables and rely on prior information for evaluating decisions under uncertainty. Finally, the principles demonstrated in this study can be formulated as a generalized entropy and utility based multinomial logit model.
Questions and discussions

Melvin Wong  
Postdoctoral Research Fellow  
Laboratory of Innovations in Transportation (LITrans)  
email: melvin.wong@ryerson.ca