# Bayesian Deep Learning for Discrete Choice

Daniel F. Villarraga[1] and Ricardo A. Daziano*[1]

[1]School of Civil and Environmental Engineering, Cornell University, USA

## Short summary

Discrete choice models are used to analyze individual decision-making in contexts such as consumer behavior, transportation choices, and political elections. Whereas traditional discrete choice models offer high economic interpretability, these models often underperform in predictive tasks compared to deep learning. Deep learning models remain largely underutilized in discrete choice due to concerns about their lack of interpretability, unstable parameter estimates, and the absence of established methods for uncertainty quantification. In this work, we introduce a deep learning model architecture that exploits approximate Bayesian inference methods via Stochastic Gradient Langevin Dynamics. Our proposed model collapses to behaviorally informed hypotheses when data is limited, mitigating overfitting and instability in underspecified settings while retaining the flexibility to capture complex nonlinear relationships when sufficient data is available. We demonstrate the value of our approach in a Monte Carlo simulation study, evaluating both predictive metrics –such as out-of-sample balanced accuracy– and inferential metrics –such as empirical coverage for marginal rates of substitution interval estimates. Additionally, we present results from two empirical case studies: one using revealed mode choice data in NYC, and the other based on the widely used Swiss train choice stated preference data.
**Keywords**: Bayesian deep learning; discrete choice modeling; stochastic gradient Langevin dynamics; uncertainty quantification; value of travel time

## 1 Introduction

Discrete choice models aim to explain how individuals choose among discrete alternatives by estimating the effect of attributes such as travel time, cost, and service characteristics on economic value and choice probabilities. Unlike purely predictive machine learning approaches, the central objective of discrete choice analysis is behavioral inference, including the estimation of microeconomic quantities such as elasticities, marginal utilities, and willingness to pay, including valuation of travel time savings.

Traditional econometric specifications in the logit and probit families provide interpretable functions of parameter estimates and well-established methods for statistical inference (Train, 2009). However, these models impose strong structural assumptions, often relying on linear compensatory effects that may oversimplify complex valuation and interactions present in real-world behavioral data.

Recent research has explored the use of deep learning models to improve predictive accuracy in discrete choice problems (Wang et al., 2021). Neural networks can approximate highly nonlinear functions and automatically learn interactions among attributes. Nevertheless, adoption in discrete choice analysis remains limited for several reasons. First, generic neural network architectures lack explicit connections to behavioral theory, making interpretation beyond prediction difficult. Second, deep models trained on relatively small datasets —-common in transportation studies– can produce unstable parameter estimates. Third, most existing applications focus on point prediction and do not provide uncertainty quantification for behavioral outputs such as willingness to pay.

Bayesian deep learning offers a potential solution to these challenges. Instead of relying on a single set of estimated parameters, Bayesian methods represent uncertainty through a posterior distribution over model parameters (Wilson & Izmailov, 2020). This posterior distribution can be propagated to behavioral quantities, enabling credible intervals for economic measures such as marginal rates of substitution. However, current Bayesian neural network approaches fail to incorporate the behavioral structure underlying discrete choice models, leaving the integration of utility-based econometric frameworks largely unaddressed.

This paper proposes a deep learning architecture designed specifically for discrete choice modeling within an approximate Bayesian framework. The model combines a knowledge-informed utility component with nonlinear neural blocks whose magnitude is explicitly controlled. When the data provide limited evidence for nonlinear effects, the model naturally collapses toward the simpler behavioral specification. Posterior inference is performed using stochastic gradient Langevin dynamics (SGLD), which supports posterior sampling during training.

The performance of the proposed approach is evaluated through a Monte Carlo simulation study and two empirical case studies: multinomial mode choice in New York City and a stated-preference train route choice experiment from Switzerland. The results show that the proposed architecture provides well-calibrated credible intervals for marginal rates of substitution while maintaining competitive predictive accuracy.

## 2 Methodology

### *Standard discrete choice framework*

Let $u_{ij} = v_{ij} + \epsilon_{ij}$ denote latent utility for individual $i$ and alternative $j$, where $v_{ij}$ is the representative utility and $\epsilon_{ij}$ captures unobserved factors. A common behavioral specification assumes a linear form $v_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta}$, where $\mathbf{x}_{ij}$ are attributes and $\boldsymbol{\beta}$ are marginal utilities. Choice probabilities are derived from assumptions on the joint distribution of the error terms $\epsilon_{ij}$, with most empirical applications in transportation economics relying on conditional logit kernels. The linear specification of $v_{ij}$ implies compensatory trade-offs across attributes, so marginal rates of substitution (MRS) are given by the parameter ratios, with parameters typically obtained using the maximum likelihood estimator (MLE).

### *Deep learning architecture*

Our contribution is a discrete-choice-specific neural architecture designed to (i) preserve a behaviorally interpretable baseline (linear compensatory effects), (ii) allow flexible nonlinearities when supported by the data, and (iii) support approximate Bayesian inference with calibrated uncertainty on *behavioral* outputs (e.g., marginal utilities and MRS). Our key design choice is to separate the representative utility into an expert-informed component and two nonlinear components with distinct economic interpretations. For individual $i$ and alternative $j$, the representative utility is

$$v_{ij} = \alpha_j + \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{q}_{ij}^\top \boldsymbol{\gamma} + \sigma_{IIA}\, \tilde{f}_{\Theta_f}(\mathbf{x}_{ij}, \mathbf{q}_{ij}) + \sigma_{nonIIA}\, \tilde{g}_{\Theta_g}(\mathbf{x}_i, \mathbf{q}_i)_j, \tag{1}$$

where $\mathbf{q}_i$ are individual characteristics. We construct alternative-specific embeddings $\mathbf{q}_{ij}$ from $\mathbf{q}_i$ so that shared covariates enter utilities in an identifiable and interpretable way across alternatives. The term $\alpha_j + \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{q}_{ij}^\top \boldsymbol{\gamma}$ is the *knowledge-informed* component and is the target of economic interpretation in low-data regimes. To allow substitution patterns across alternatives that differ from those implied by the independence of irrelevant alternatives (IIA) property of the conditional logit model, the nonlinear component is decomposed into (a) an *IIA-consistent* block $\tilde{f}_{\Theta_f}(\cdot)$ applied independently to each alternative and (b) a *non-IIA* block $\tilde{g}_{\Theta_g}(\cdot)$ that takes as input the concatenation of all alternatives'

attributes (and $\mathbf{q}_i$) and outputs a $J$-dimensional vector. This decomposition separates nonlinear taste heterogeneity and attribute interactions (captured by the IIA-consistent block) from flexible cross-alternative substitution patterns (captured by the non-IIA block). In practice, both $\tilde{f}$ and $\tilde{g}$ are implemented as small fully connected networks. To prevent nonlinear components from overwhelming the behavioral baseline, we normalize the outputs of the nonlinear blocks using non-affine batch normalization and introduce explicit scale parameters $\sigma_{IIA}$ and $\sigma_{nonIIA}$ that control their magnitudes. This makes prior specification targeted and interpretable: regularization (or Bayesian priors) can be placed primarily on the *strength of nonlinearities* rather than indiscriminately on all network weights. When the data do not justify complex patterns, posterior mass concentrates near small $\sigma_{IIA}$ and $\sigma_{nonIIA}$, and the model effectively collapses to the knowledge-informed specification; when data are informative, the nonlinear components are activated in a controlled way.
This structure is particularly compatible with posterior sampling methods such as SGLD because it reduces pathological non-identifiability common in generic deep networks for discrete choice, while still allowing uncertainty in $\boldsymbol{\beta}$ and in the induced behavioral quantities to be represented through posterior draws.

### *Approximate Bayesian inference*

To quantify uncertainty in model parameters and behavioral outputs, we adopt stochastic gradient Langevin dynamics (SGLD, Welling & Teh, 2011). SGLD modifies standard stochastic gradient descent by effectively injecting Gaussian noise into the parameter updates:

$$\Delta\theta_t = \frac{\alpha_t}{2}\nabla \log p(\theta_t|D) + \eta_t \tag{2}$$

where $\eta_t \sim \mathcal{N}(0, \alpha_t)$. The resulting parameter iterates then approximate samples from the posterior distribution. Behavioral quantities such as marginal utilities can therefore be evaluated for each draw, producing posterior distributions and credible intervals. For example, the marginal utility with respect to attribute $k$ is

$$\frac{\partial u_{ij}}{\partial x_{ijk}} = \beta_k + \sigma_{IIA}\frac{\partial f}{\partial x_{ijk}} + \sigma_{nonIIA}\frac{\partial g}{\partial x_{ijk}}. \tag{3}$$

Posterior samples of these derivatives allow direct computation of marginal rates of substitution and value-of-time measures.

## 3 Results and discussion

### *Monte Carlo study*

A full Monte Carlo study was conducted to evaluate the empirical coverage of credible intervals for marginal rates of substitution. The data-generating process includes nonlinear utility components and three alternatives. Datasets of 1,000 and 10,000 observations were generated across 100 replications. Table 1 reports the empirical coverage of 95% credible intervals for four marginal rates of substitution of interest.

Table 1: Empirical coverage of marginal rates of substitution.

| Model | N=1000 | N=10000 |
|---|---|---|
| Proposed model | 97% | 96% |
| Fully connected NN | 81% | 93% |
| Conditional logit | 74% | 62% |

The proposed architecture achieves coverage close to the nominal credibility level across dataset sizes. In contrast, the fully connected neural network performs poorly in smaller samples, while the conditional logit model fails to capture nonlinear relationships in the data.

***Mode choice in New York City***

The first empirical case study examines multinomial mode choice for home-based work trips in New York City. After cleaning the dataset, 3,277 observations remained with three alternatives: transit, car, and non-motorized travel. The proposed model yields distributions of the value of travel time savings (VOTT) that are consistent with behavioral expectations, with positive willingness to pay for travel time reductions and magnitudes comparable to those typically reported in the literature. Whereas fully connected neural networks can achieve similar levels of predictive accuracy, we show how they frequently produce economically implausible VOTT distributions, including large negative values, reflecting unstable marginal utilities. In contrast, our proposed architecture maintains behaviorally coherent economic measures while achieving out-of-sample predictive accuracy of approximately 78%, which is comparable to the best-performing models.

***Swiss train route choice***

The second case study uses a widely studied stated-preference dataset on Swiss train route choices (Vrtic & Axhausen, 2002). Using posterior samples, the model estimates individual-level distributions of the value of travel time savings (VOTT). The proposed architecture yields economically plausible VOTT estimates consistent with ranges reported in previous studies while maintaining stable marginal utilities across posterior draws. In addition, the model achieves out-of-sample predictive accuracy that exceeds benchmarks reported in the literature for this dataset. In contrast, fully connected neural networks perform substantially worse in predictive accuracy and frequently generate unstable behavioral estimates, highlighting the importance of incorporating discrete-choice structure.

## 4 Conclusions

This paper presents a deep learning architecture designed for discrete choice analysis within a Bayesian framework. The model integrates behavioral structure with neural network flexibility through controlled nonlinear components that allow the model to revert to simpler hypotheses when the data provide limited evidence for complex relationships. Approximate Bayesian inference using stochastic gradient Langevin dynamics enables posterior sampling of model parameters and credible intervals for behavioral quantities such as marginal rates of substitution and values of travel time. Simulation results demonstrate well-calibrated empirical coverage of credible intervals, whereas empirical case studies show competitive predictive performance and behaviorally plausible economic estimates. These findings suggest that combining theory-informed model design with Bayesian deep learning provides a promising direction for integrating modern machine learning techniques into discrete choice analysis while preserving the interpretability and inferential capabilities central to transportation research.

## Acknowledgements

This research was supported by the National Science Foundation (Award SES-2342215) and the Fulbright Scholarship Program.

## References


Train, K. E. (2009). *Discrete choice methods with simulation* (2nd ed.). Cambridge University Press. doi: 10.1017/CBO9780511805271

Vrtic, M., & Axhausen, K. W. (2002). The impact of tilting trains in Switzerland: A route choice model of regional-and long distance public transport trips. *Arbeitsberichte Verkehrs-und Raumplanung*, *128*.

Wang, S., Mo, B., Hess, S., & Zhao, J. (2021). Comparing hundreds of machine learning classifiers and discrete choice models in predicting travel behavior: an empirical benchmark. *arXiv preprint arXiv:2102.01130*.

Welling, M., & Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (icml-11)* (pp. 681–688).

Wilson, A. G., & Izmailov, P. (2020). Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, *33*, 4697–4708.