Neural-Network Mixed Logit Choice Model: Statistical and Optimality Guarantees

(Authors' names blinded for peer review)

The modeling of consumer choice behaviors, particularly their interactions with product and consumer features, has been a fundamental topic in revenue management, marketing, and economics. Pioneered by seminal works on the Multinomial Logit (MNL) model [17, 18, 16], discrete choice models become essential tools for predicting consumer behaviors and forecast demands. These models inform managerial decisions in various applications, such as pricing, assortment optimization, and transportation planning. To address consumer taste heterogeneity and substitution effects, a variety of choice models have been developed, including but not limited to nested logit [20], mixed logit [5, 2, 22, 21, 10], ranked-list models [4, 15, 8, 12], Exponomial [1], Markov chain models [3], and tree-based models [6, 7, 9].

This paper focuses on the mixed logit, which characterizes choice probability as a mixture of MNLs. Each mixture corresponds to a consumer type associated with a taste vector, representing the linear coefficient for the features. The *mixture distribution* of these taste vectors captures market heterogeneity. The mixed logit can flexibly represent any choice behaviors consistent with random utility maximization [21] while retaining the interpretability of MNL, which quantifies the sensitivity of features to utility.

Early literature has observed that the mixed logit model, with linear utility, resembles a one-hidden-layer neural network [19]. In this study, we revisit the effectiveness of representing the mixed logit using a one-hidden-layer neural net. Specifically, we study NN-mixed logit, which utilizes one hidden layer comprising N neurons to approximate the mixture distribution as an equally weighted distribution on N consumer types. Each neuron corresponding to one consumer type is linked with a taste vector representing its taste preferences. These neurons compute the MNL choice probability using Softmax activations. The output layer calculates the average of these MNL probabilities, serving as the predicted choice probability from the equally weighted mixture distribution.

From the modeling perspective, it is essential to address two fundamental statistical problems of the neural network: expressiveness and generalization. The expressiveness depicted as the approximation error measures how well NN-mixed logit can approximate the mixed logit model. Given

the Universal Approximation Theorem [11], the approximation error would decrease as the number of parameters N increases. While overparameterization benefits expressiveness, it may risk the generalization capability that assesses how well NN-mixed logit generalizes to unseen data given a finite number of training data.

From the learning perspective, we are interested in optimizing the estimation of the taste vectors, which presents significant challenges in the existing literature. Typically estimation methods, such as maximum likelihood or least squares, result in a non-convex optimization problem, even when the number of consumer types is fixed. As a result, gradient descent [10] and expectation-maximization [14] lack global optimality guarantees. Recent work [13] made promising advances by reformulating the problem as a constrained convex optimization over the class of choice probability vectors that a mixed logit can represent. This enables the use of the globally convergent Frank-Wolfe algorithm. Nonetheless, to guarantee such theoretical convergence properties, the algorithm is required to solve a non-convex subproblem at each iteration that can be computationally challenging. In our work, the taste vectors contained in the neural network are learned using a noisy gradient descent algorithm. We hope to address two optimization questions: whether the algorithm can converge to the global optimal mixture distribution, and if so, how fast it can converge.

We summarize our main contributions as follows.

(1) From a statistical standpoint, the neural network model offers universal expressiveness without suffering the curse of dimensionality while also demonstrating strong generalization to unseen data. Specifically, we show that any mixture distribution can be approximated by an NN-mixed logit with N consumer types, with an $O(1/\sqrt{N})$ error in the predicted choice probability vector, measured by the expected 2-norm. The constant is universal, indicating that the model does not suffer from the curse of dimensionality. Moreover, we prove that the Rademacher complexity of the mixture model class is independent of the width of the hidden layer, provided that the expected norm of the taste vector or the entropy of the mixture distribution is appropriately controlled. Thereby, overparameterization does not undermine generalization on unseen data.

(2) From an optimization standpoint, using a novel discrete-time mean-field analysis, we demonstrate that the noisy gradient descent algorithm converges exponentially fast to the global optimal solution of the norm- and entropy-regularized estimation problem. This convergence is subject to three error terms: the first proportional to the polynomial of the step size, resulting

from the randomness in the gradient oracle, the second due to the entropy regularization, and the third inversely proportional to the number of consumer types, resulting from the approximation error of finitely many consumer types. We bound the gap between the infinite-width model and the finite-width model, which decays inverse proportionally to the width of the hidden layer.

(3) We empirically demonstrate our approach's superior in-sample and out-of-sample performance compared to other benchmarks using synthetic and real datasets. The numerical results also validate our theoretical findings, in terms of expressiveness, width-independent generalization capability, and convergence behavior.

To conclude, we revisited the neural network representation for the mixed logit choice model in this paper. We demonstrated that it can approximate any mixture distribution without suffering from the curse of dimensionality or overfitting. Moreover, we showed that it can be learned by noisy gradient descent with guaranteed global convergence. To the best of our knowledge, it is the first work that theoretically guarantees to recover the global optimal parameters without using the panel data structure. These findings underscore the potential of even shallow neural network representations, coupled with efficient training algorithms, to effectively learn complex choice models with statistical and optimality guarantees.

References

[1] Alptekinoğlu A, Semple JH (2016) The exponomial choice model: A new alternative for assortment and price optimization. *Operations Research* 64(1):79–93.

[2] Bhat CR (1997) An endogenous segmentation mode choice model with an application to intercity travel. *Transportation science* 31(1):34–48.

[3] Blanchet J, Gallego G, Goyal V (2016) A markov chain approximation to choice modeling. *Operations Research* 64(4):886–905.

[4] Block HD, Marschak J (1959) Random orderings and stochastic theories of response .

[5] Cardell NS, Dunbar FC (1980) Measuring the societal impacts of automobile downsizing. *Transportation Research Part A: General* 14(5-6):423–434.

[6] Chen N, Gallego G, Tang Z (2021) Estimating discrete choice models with random forests. *INFORMS International Conference on Service Science*, 184–196 (Springer).

[7] Chen YC, Mišić VV (2022) Decision forest: A nonparametric approach to modeling irrational choice. *Management Science* 68(10):7090–7111.

[8] Farias VF, Jagabathula S, Shah D (2013) A nonparametric approach to modeling choice with limited data. *Management science* 59(2):305–322.

[9] Feng Q, Shanthikumar JG, Xue M (2023) Rational choice models: The temporal tree representation. *Available at SSRN 4653687* .

[10] Hensher DA, Greene WH (2003) The mixed logit model: the state of practice. *Transportation* 30:133–176.

[11] Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. *Neural networks* 2(5):359–366.

[12] Jagabathula S, Rusmevichientong P (2017) A nonparametric joint assortment and price choice model. *Management Science* 63(9):3128–3145.

[13] Jagabathula S, Subramanian L, Venkataraman A (2020) A conditional gradient approach for nonparametric estimation of mixing distributions. *Management Science* 66(8):3635–3656.

[14] Laird N (1978) Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* 73(364):805–811.

[15] Mahajan S, Van Ryzin G (2001) Stocking retail assortments under dynamic consumer substitution. *Operations research* 49(3):334–351.

[16] Malhotra NK (1984) The use of linear logit models in marketing research. *Journal of Marketing research* 21(1):20–31.

[17] McFadden D (1972) Conditional logit analysis of qualitative choice behavior .

[18] McFadden D (1981) Econometric models of probabilistic choice. *Structural analysis of discrete data with econometric applications* 198272.

[19] McFadden D (2001) Economic choices. *American economic review* 91(3):351–378.

[20] McFadden D, Karlqvist A, Lundqvist L, Snickars F, Weibull JW (1978) Spatial interaction theory and planning models. *Modeling the choice of residential location* 75–96.

[21] McFadden D, Train K (2000) Mixed mnl models for discrete response. *Journal of Applied Econometrics* 15(5):447–470.

[22] Revelt D, Train K (1998) Mixed logit with repeated choices: households' choices of appliance efficiency level. *Review of economics and statistics* 80(4):647–657.