NBER WORKING PAPER SERIES

AMORTIZED INFERENCE FOR CORRELATED DISCRETE CHOICE MODELS VIA EQUIVARIANT NEURAL NETWORKS

Easton K. Huch
Michael P. Keane

Working Paper 35037
http://www.nber.org/papers/w35037

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2026, Revised September 2026

We have no financial interests to disclose. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2026 by Easton K. Huch and Michael P. Keane. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Amortized Inference for Correlated Discrete Choice Models via Equivariant Neural Networks
Easton K. Huch and Michael P. Keane
NBER Working Paper No. 35037
April 2026, Revised September 2026
JEL No. C10, C13, C15, C25, C35, C45

ABSTRACT

Discrete choice models are fundamental tools in management science, economics, and marketing for understanding and predicting decision-making. Logit-based models are dominant in applied work, largely due to their convenient closed-form expressions for choice probabilities. However, they impose restrictive assumptions on the stochastic utility component, constraining our ability to capture realistic substitution patterns. We propose an amortized inference approach that relies on a neural network emulator to approximate choice probabilities for general error distributions, including those with correlated errors. We develop a specialized neural network architecture designed to respect the invariance properties of discrete choice models. We provide group-theoretic foundations for the architecture, including a proof of universal approximation given a minimal set of invariant features. Once trained, the emulator enables rapid likelihood evaluation and gradient computation. We use Sobolev training, augmenting the likelihood loss with a gradient-matching penalty, so that the emulator learns both choice probabilities and their derivatives. We show that emulator-based maximum likelihood estimators are consistent and asymptotically normal under mild approximation conditions, and we provide sandwich standard errors that remain valid for a pseudo-true parameter even with imperfect likelihood approximation. Simulations show significant gains over the GHK simulator in accuracy and speed.

Easton K. Huch
Johns Hopkins University
Carey Business School
ehuch@jhu.edu

Michael P. Keane
Johns Hopkins University
Carey Business School
and NBER
mkeane14@jhu.edu

1 Introduction

Discrete choice models are a widely used tool in management science, economics, marketing, and other fields for understanding how individuals and organizations make decisions among finite sets of alternatives (McFadden, 1974). These models provide a structural framework that delivers interpretable parameters—such as willingness-to-pay, demand elasticities, and measures of perceived similarity between products—while also enabling predictions about effects of policy changes, such as price changes and product introductions/deletions.

The dominant discrete choice model in applied work is the multinomial logit (MNL), which assumes the utility that consumer i obtains from alternative j takes the form $U_j = v_j + \epsilon_j$ where v_j is the deterministic component of utility, typically written as a linear function $\boldsymbol{x}_j^\top \boldsymbol{\beta}$ of alternative j's attributes $\boldsymbol{x}_j$, and ϵ_j is a Type I extreme value (Gumbel) error that is independent and identically distributed (iid) across alternatives $j = 1, \ldots, K$. This setup yields simple closed-form choice probabilities: the probability of choosing alternative j is given by the "softmax" function $\exp(v_j)/\sum_{k=1}^{K} \exp(v_k)$. The resulting computational convenience has made MNL the default choice in countless applications.

However, this convenience comes at a well-known cost. The iid Gumbel error assumption implies the restrictive independence of irrelevant alternatives (IIA) property: the odds ratio between any two alternatives is unaffected by the presence or attributes of other alternatives, which can lead to unrealistic substitution patterns. A popular generalization of MNL is the mixed logit model (mixed-MNL), in which the utility weight $\boldsymbol{\beta}$ is allowed to be heterogeneous across consumers. Estimation is straightforward via simulation methods (Train, 2009). The mixed-MNL relaxes IIA at the aggregate level, but IIA still holds at the level of consumer i's individual choices given the corresponding random utility weight, $\boldsymbol{\beta}_i$.

The multinomial probit (MNP) is a well-known alternative to MNL that relaxes IIA and allows for flexible substitution patterns by assuming $\epsilon_1, \ldots, \epsilon_K$ come from a multivariate normal distribution with cross-alternative correlations. Despite this advantage, MNP has seen relatively limited adoption, largely due to its computational demands. MNP choice probabilities have no closed form and require evaluating multivariate normal rectangle probabilities. Estimation relies on more sophisticated simulation methods than needed for mixed-MNL: the GHK simulator in a classical setting (Geweke, 1989; Hajivassiliou and McFadden, 1998; Keane, 1994) or MCMC in a Bayesian setting (McCulloch and Rossi, 1994).

We propose a fundamentally different approach: Rather than simulating choice probabilities anew for each likelihood evaluation, we train a neural network emulator to directly approx-

imate the choice probability function. This strategy—known as amortized inference—shifts the computational burden from inference time to a one-time training phase. Once trained, the emulator provides rapid deterministic approximations to choice probabilities via simple function calls. The amortized inference framework has proven highly successful in multiple scientific domains for inference in computationally intensive simulation models (Lueckmann et al., 2019; Cranmer et al., 2020). Here, we adapt it to discrete choice models, extending beyond MNL and MNP to general correlated error distributions. One especially interesting case is an extension of MNL to allow for correlated Gumbel errors, thus relaxing IIA at the individual level. We refer to this as the correlated-Gumbel model.

In the econometrics literature, Norets (2012) employs a related strategy, approximating the expected value function in dynamic discrete choice models via a neural network. He applies this strategy to a specific parametric model, so changes in the model require modifying and retraining the network. In contrast, our emulator operates on the generic choice problem, given only the deterministic utilities and the error distribution, enabling modifications to the parametric model (e.g., the functional form of utility) without retraining the emulator.

Our methodology relies on several new contributions. First, we develop a neural network architecture specifically designed for choice models. It respects their fundamental invariance properties, which we formalize as symmetries (group actions) of the choice-probability map.[1] Choice probabilities are invariant to affine utility transformations (location and scale shifts), and equivariant with respect to permutations of choice alternatives. Our architecture incorporates a preprocessing transformation that reduces the size of the feature space by enforcing location and scale invariance. The processed features are then passed to a per-alternative encoder module based on the DeepSet architecture (Zaheer et al., 2017). The output is then concatenated across alternatives and processed through equivariant layers that respect the permutation equivariance property.

Second, we establish the theoretical foundations of our architecture. We prove that it can universally approximate choice probabilities on compact subsets of the parameter space, outside a measure-zero exceptional set. This result connects our architecture to the theory of orbit separation under group actions, extending the universal approximation results of Blum-Smith et al. (2025) on symmetric matrices to the joint space of utility vectors and centered covariance matrices. Blum-Smith et al. (2025) employ Galois theory to prove generic orbit separation, while we provide a direct proof based on an invariant reconstruction argument.

[1] If these properties are not embedded in the architecture, the NN must learn them during training. As a continuum of models generate equivalent choice probabilities, this would slow learning considerably.

Third, we establish the statistical properties of maximum likelihood estimators (MLEs) formed using an emulator approximation of the true discrete choice probabilities. We show that if the emulator approximates the true log likelihood sufficiently well—specifically, if the average approximation error is $o_p(n^{-1})$—then the emulator-based estimator inherits the consistency and asymptotic normality of the exact MLE. When this condition is not met, we show that valid inference can still be obtained for a pseudo-true parameter via sandwich standard errors, treating the emulator as a working model in the quasi-ML framework.

By careful design, our preprocessing transformations and architecture enforce smoothness of choice probabilities with respect to the inputs: the deterministic utilities and corresponding scale (covariance) matrix. We use Sobolev training (Czarnecki et al., 2017) to match emulator gradients to those of the log choice probabilities. Together, these design choices enable reliable use of automatic differentiation for downstream model-fitting and inference tasks.

With a pretrained emulator, generalizing from logit to probit (or another error distribution) requires only the replacement of closed-form softmax probabilities with emulator evaluations of choice probabilities when forming the likelihood. In the probit case, given a fixed computational budget, we show via simulations that an ML estimator using our amortized inference procedure matches or exceeds the performance of one using the GHK algorithm in terms of estimation error and coverage rates. Furthermore, our approach can easily handle models other than MNP where an efficient simulation algorithm like GHK is unavailable.

For concreteness, much of our exposition focuses on MNP. However, the emulator approach and supporting theory are largely agnostic to the assumed parametric form of the errors. Generalizing to other error specifications (e.g., Gumbel or multivariate-t) is straightforward, requiring only modest changes to the training data generation and emulator inputs.

The remainder of this paper is organized as follows. Section 2 summarizes related literature on choice modeling. Section 3 describes the family of correlated discrete choice models we consider and their properties. Section 4 presents the emulator architecture and training procedure. Section 5 establishes the theoretical properties of the architecture and emulator-based estimators. Section 6 presents simulation results, and Section 7 concludes.

2 Related Literature

The literature on applying machine learning methods to discrete choice can be divided into two broad streams. The first stream maintains the classic random utility model (RUM) of choice, in which choice probabilities are generated by a population of rational consumer

types with different (but stable) preference orderings over the universe of choice objects, as explained by Bloch and Marschak (1960) and McFadden and Richter (1990).[2] MNL, MNP, and mixed-MNL are all members of the RUM class provided one maintains the RUM assumption that the utility of an option depends only on its own characteristics. Key papers in this strand start from this basic MNL structure and use neural networks to generalize the functional form of utility: Bentz and Merunka (2000), Sifringer et al. (2020), Wang et al. (2020), Han et al. (2022), and Singh et al. (2023). Two important recent papers extend this work: Aouad and Desir (2025) develop an architecture that implements the mixed logit model with a flexible distribution of taste heterogeneity (RUMnet), and Bagheri et al. (2025) develop another architecture that generalizes the Gumbel error assumption (RUM-NN). Both papers use softmax-smoothed sample averages to approximate choice probabilities within the loss function, resulting in increased computation time relative to pure logit-based models.

Aouad and Desir (2025) build on the generic approximation property of mixed-MNL models shown in McFadden and Train (2000), which relies on a flexible basis expansion of the utility function and a nonparametric mixing distribution. In practice, these model attributes are unknown to analysts. Furthermore, this approximation property is not unique to mixed-MNL. Rather, as noted in both Aouad and Desir (2025) and McFadden and Train (2000), it holds more generally for a larger class of hierarchical discrete choice models, including mixed-MNP models. McFadden and Train (2000) use this approximation property to justify adoption of mixed-MNL with a parametric mixing distribution as a computationally convenient alternative to other models lacking closed-form choice probabilities, such as MNP. But this justification of mixed-MNL becomes less compelling given a practical, general-purpose alternative like the emulators we propose here.

The second stream dispenses with the RUM structure, often motivated by a desire to relax restrictive MNL assumptions like IIA and to allow more flexible substitution patterns. Some papers in this stream view discrete choice as a general classification problem that is amenable to machine learning methods. This is exemplified by Wang and Ross (2018), Lhéritier et al. (2019), Rosenfeld et al. (2020), Chen and Mišić (2022), and Chen et al. (2025). Others maintain an MNL structure at the top level—that is, choice probabilities are determined by a vector of alternative-specific utilities that enter a softmax function—but a neural net is used to construct the alternative-specific utility functions in flexible ways that deviate from RUM assumptions. For instance, the utility of alternative j is allowed to depend on

[2]Suppose there are $\bar{K}$ objects in the universe and a consumer is presented with a choice set (or "assortment") that contains $K \leq \bar{K}$ elements. A key implication of RUM is that the utility a consumer derives from product k is invariant to the choice set, ruling out context or assortment effects.

attributes of other alternatives to generate context effects. This is exemplified by Wang et al. (2021), Wong and Farooq (2021), Cai et al. (2022), Pfannschmidt et al. (2022), and Berbeglia and Venkataraman (2025).

These two streams of research present a fundamental tension between model flexibility and model interpretability: While RUM models are favored for their interpretability and grounding in economic theory, the dominant RUM models in empirical work (linear MNL and mixed-MNL) impose strong and potentially unrealistic constraints on deterministic utilities and substitution patterns; in particular, they assume IIA at the level of individual choices. The first stream seeks to address this problem by adding flexibility to RUM models. But many of these proposals maintain the assumption of independent logit errors, and thus they still constrain substitution patterns to obey IIA. The exceptions, namely, RUMnet and RUM-NN, allow for more general substitution patterns via flexible error distributions. But they do so at the expense of computational efficiency, as their approach requires expensive sample-average approximations to the likelihood function (a problem that our approach avoids, as we discuss below). The second stream abandons the RUM structure altogether. While this leads to very flexible models, it makes it difficult or impossible to obtain reliable inferences for many economically meaningful quantities, such as consumer welfare, willingness-to-pay measures, demand elasticities, and substitution effects.

MNP models constitute a notable exception to the above tradeoff as they allow for flexible substitution patterns via *interpretable* covariance relationships. Moreover, if additional flexibility is desired, they can be extended to allow the deterministic utilities to follow the functional form of a neural network (Hruschka, 2007). Conversely, if researchers desire a more parsimonious or interpretable model, analysts can constrain the covariance structure via penalization methods as in Jiang et al. (2025) or via factor structures as we illustrate in Section 6. Despite these virtues, empirical applications of MNP are relatively sparse in the literature due, in large part, to the difficulty of evaluating MNP choice probabilities.

More broadly, MNP models are just one member of a large class of RUM models with correlated errors. Other models in this class can be generated by assuming errors of the form $\boldsymbol{\epsilon} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\epsilon}^*$, where $\boldsymbol{\epsilon}^*$ is a vector of exchangeable errors, $\boldsymbol{\Sigma}$ is a scale matrix, and $\boldsymbol{\Sigma}^{1/2}$ satisfies $\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\Sigma}^{1/2})^\top = \boldsymbol{\Sigma}$. This is the general class of models we study. For example, taking $\epsilon_j^* \overset{iid}{\sim}$ Gumbel$(0, 1)$ provides a natural way to generalize MNL and mixed-MNL to allow correlated errors. This "correlated-Gumbel" model relaxes IIA at the individual level. As another example, we could assume $\boldsymbol{\epsilon}^* \sim$ Multivariate-$t(\mathbf{0}, \mathbf{I}, \nu)$ to capture heavy-tailed behavior via the degrees-of-freedom parameter ν. Bagheri et al. (2025) also consider correlated error

distributions formed in this manner, but they need to resort to expensive simulation methods to approximate the choice probabilities because an efficient algorithm (GHK) is available only in the Gaussian case. Our approach allows one to circumvent this computational barrier using a neural net emulator for general discrete choice probabilities.

The primary contribution of this work is an amortized inference framework, based on our emulator, that produces accurate, reusable, and computationally efficient estimates of choice probabilities for RUM models with general error distributions, including those featuring nontrivial correlation structures. The framework presents a potential resolution to the flexibility–interpretability tradeoff highlighted above, enabling practical estimation of flexible choice models without sacrificing the economically meaningful insights and parsimony of RUM models. The framework is supported by strong theoretical justification, including a universal approximation guarantee and asymptotic inference results under mild approximation conditions. Moreover, the framework is largely complementary to recent advances in machine learning methods, including the two streams discussed above. In many cases, these methods could be enhanced with emulator-based likelihood evaluations, resulting in flexible models with interpretable substitution patterns and manageable computational demands.

3 Problem Setup

This section summarizes the problem setup, including the invariance properties of discrete choice models and our inferential goals.

3.1 Discrete Choice Models

We consider a decision-maker choosing among K mutually exclusive alternatives. The decision-maker assigns a latent utility U_j to each alternative $j \in \{1, \ldots, K\}$ and selects the alternative with the highest utility:

$$Y = \underset{j \in \{1,\ldots,K\}}{\arg\max}\, U_j.$$

The latent utilities decompose into deterministic and stochastic components:

$$U_j = v_j + \epsilon_j, \quad j = 1, \ldots, K,$$

where v_j is the deterministic (systematic) utility that depends on observable characteristics, and ϵ_j is a random error capturing unobserved factors. As explained in Section 1, the deterministic utility typically takes a linear form $v_j = \boldsymbol{x}_j^\top \boldsymbol{\beta}$, where $\boldsymbol{x}_j$ is a vector of alternative-specific attributes and $\boldsymbol{\beta}$ is a parameter vector to be estimated.

We introduce a scale matrix, $\boldsymbol{\Sigma}$, with the following structure:[3]

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \mathbf{0}_{K-1}^\top \\ \mathbf{0}_{K-1} & \boldsymbol{\Omega}, \end{bmatrix} \tag{1}$$

where $\boldsymbol{\Omega} \in \mathbb{R}^{(K-1)\times(K-1)}$ and $\mathbf{0}_{K-1}$ is a zero vector of dimension $K-1$. The scale matrix transforms the errors as

$$\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_K)^\top = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\epsilon}^*,$$

where $\boldsymbol{\epsilon}^*$ is a vector of exchangeable errors and $\boldsymbol{\Sigma}^{1/2}$ is chosen according to a predefined rule such that $\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\Sigma}^{1/2})^\top = \boldsymbol{\Sigma}$. For general distributions, we specify that $\boldsymbol{\Sigma}^{1/2}$ is the (unique) symmetric positive semidefinite square-root matrix. However, in special cases described in Definition 1, the choice of factor does not matter, so we are free to select another rule, such as the lower Cholesky factor.

Under the above setup, we obtain the MNP family with $\boldsymbol{\epsilon}^* \sim \mathcal{N}(\mathbf{0}_K, \mathbf{I}_K)$ and $\boldsymbol{\Omega}$ unconstrained, and we obtain the MNL as the special case $\boldsymbol{\Omega} = \mathbf{I}_{K-1}$ and $\epsilon_j^* \overset{iid}{\sim}$ Gumbel$(a, 1)$, where $a \in \mathbb{R}$ is arbitrary.[4] We refer to the model obtained by allowing $\boldsymbol{\Omega}$ to differ from $\mathbf{I}_{K-1}$ while maintaining iid Gumbel ϵ_j^* as the "correlated-Gumbel" model. This model relaxes IIA at the individual level. We impose the following assumption on $\boldsymbol{\epsilon}^*$.

Assumption 1. The error vector $\boldsymbol{\epsilon}^*$ admits a density, $f_{\boldsymbol{\epsilon}^*} : \mathbb{R}^K \to [0, \infty)$, with respect to Lebesgue measure. Moreover, $\text{Cov}(\boldsymbol{\epsilon}^*) = C\,\mathbf{I}_K$ for some $C > 0$, and the elements of $\boldsymbol{\epsilon}^*$ are exchangeable so that $f_{\boldsymbol{\epsilon}^*}(\mathbf{P}_\pi\boldsymbol{\epsilon}') = f_{\boldsymbol{\epsilon}^*}(\boldsymbol{\epsilon}')$ for all $\boldsymbol{\epsilon}' \in \mathbb{R}^K$ and any permutation matrix, $\mathbf{P}_\pi$.

In the case of the MNP model, $f_{\boldsymbol{\epsilon}^*}$ represents the probability density function (PDF) of the $\mathcal{N}(\mathbf{0}_K, \mathbf{I}_K)$ distribution. Assumption 1 places weak assumptions on $f_{\boldsymbol{\epsilon}^*}$, allowing for a wide variety of error distributions, including Gumbel, multivariate-t, and many others.

We denote the choice probability for alternative j as

$$P_j(\boldsymbol{v}, \boldsymbol{\Sigma}^{1/2}) = \Pr(U_j \geq U_k \text{ for all } k \neq j) = \Pr(\epsilon_k - \epsilon_j \leq v_j - v_k \text{ for all } k \neq j),$$

where $\boldsymbol{v} = (v_1, \ldots, v_K)^\top$. In practice, we may also choose to include additional parameters governing the distribution of $\boldsymbol{\epsilon}^*$, such as the degrees-of-freedom parameter, ν, described in

[3] Since only utility differences affect choices, MNP is often parameterized in differences relative to a base alternative (Keane, 1992). This is convenient because Gaussianity is preserved under differencing. For non-Gaussian errors this generally fails, so we instead use the level representation in (1).

[4] The constant a cancels when we consider the distribution of utility differences. To obtain mean-zero errors, we can set a equal to the negative of the Euler–Mascheroni constant (typically denoted γ).

Section 2, but these parameters are suppressed in the notation for simplicity. The choice probability can be expressed as the following integral:

$$P_j(\boldsymbol{v}, \boldsymbol{\Sigma}^{1/2}) = \int_{\mathcal{R}_j} f_{\boldsymbol{\epsilon}^*}(\boldsymbol{\epsilon}^*)\, d\boldsymbol{\epsilon}^*, \tag{2}$$
$$\mathcal{R}_j = \left\{\boldsymbol{\epsilon}^* \in \mathbb{R}^K : v_j + \left[\boldsymbol{\Sigma}^{1/2}\right]_{j\bullet}\boldsymbol{\epsilon}^* \geq v_k + \left[\boldsymbol{\Sigma}^{1/2}\right]_{k\bullet}\boldsymbol{\epsilon}^* \text{ for all } k \neq j\right\},$$

where $\mathcal{R}_j$ is the region of the error space where alternative j is chosen and $\left[\boldsymbol{\Sigma}^{1/2}\right]_{j\bullet}$ denotes the jth row of $\boldsymbol{\Sigma}^{1/2}$ so that $\left[\boldsymbol{\Sigma}^{1/2}\right]_{j\bullet}\boldsymbol{\epsilon}^* = \epsilon_j$. In general, this integral is not analytically tractable. In MNP models, in particular, it does not have a closed-form solution for $K \geq 3$, necessitating numerical methods, such as the GHK algorithm, which simulates these probabilities via recursive conditioning. For each evaluation, GHK requires R simulation draws, with the approximation error decreasing at rate $O(R^{-1/2})$.

3.2 Invariance Properties

Choice probabilities satisfy three fundamental invariance properties.

Location invariance. Adding a constant to all utilities does not change choice probabilities:

$$P_j(\boldsymbol{v} + C\mathbf{1}_K, \boldsymbol{\Sigma}^{1/2}) = P_j(\boldsymbol{v}, \boldsymbol{\Sigma}^{1/2}) \tag{3}$$

where $\mathbf{1}_K$ is the K-vector of ones and $C \in \mathbb{R}$ may be random. This follows because choices depend only on utility differences.

Scale invariance. Scaling $\boldsymbol{v}$ and $\boldsymbol{\Sigma}^{1/2}$ preserves choice probabilities:

$$P_j(\alpha\boldsymbol{v}, \alpha\boldsymbol{\Sigma}^{1/2}) = P_j(\boldsymbol{v}, \boldsymbol{\Sigma}^{1/2}) \quad \text{for all } \alpha > 0. \tag{4}$$

This follows because the scaled model is equivalent to the original model with $\boldsymbol{v}$ and $\boldsymbol{\epsilon}$ replaced by $\alpha\boldsymbol{v}$ and $\alpha\boldsymbol{\epsilon}$, respectively.

Permutation equivariance. Relabeling alternatives permutes the choice probabilities correspondingly. For any permutation π of $\{1, \ldots, K\}$ with associated permutation matrix $\mathbf{P}_\pi$:

$$P_j(\boldsymbol{v}, \boldsymbol{\Sigma}^{1/2}) = P_{\pi^{-1}(j)}(\mathbf{P}_\pi\boldsymbol{v}, \mathbf{P}_\pi\boldsymbol{\Sigma}^{1/2}).$$

This property holds under the symmetric positive semidefinite square-root matrix because

$$\mathbf{P}_\pi\boldsymbol{v} + \left(\mathbf{P}_\pi\boldsymbol{\Sigma}\mathbf{P}_\pi^\top\right)^{1/2}\boldsymbol{\epsilon}^* \;=\; \mathbf{P}_\pi\boldsymbol{v} + \mathbf{P}_\pi\boldsymbol{\Sigma}^{1/2}\mathbf{P}_\pi^\top\boldsymbol{\epsilon}^* \;\overset{d}{=}\; \mathbf{P}_\pi\boldsymbol{v} + \mathbf{P}_\pi\boldsymbol{\Sigma}^{1/2}\boldsymbol{\epsilon}^*.$$

A naive attempt at constructing an emulator may not respect these invariance properties. Translating, rescaling, or permuting the alternatives could produce different emulator predictions, resulting in slow learning and poor generalization because the emulator would need to learn these properties manually. In Section 4, we propose a solution that automatically satisfies these properties based on two design components: a preprocessing step that fixes the location and scale, and an equivariant architecture that ensures permutation equivariance without sacrificing expressivity.

The above three invariance properties are satisfied for all discrete choice models we consider. A subset of these models also satisfy the following special property.

Definition 1. A discrete choice model is *factorization-invariant* if $\boldsymbol{\Sigma}_1^{1/2}(\boldsymbol{\Sigma}_1^{1/2})^\top = \boldsymbol{\Sigma}_2^{1/2}(\boldsymbol{\Sigma}_2^{1/2})^\top$ implies $P_j(\boldsymbol{v}, \boldsymbol{\Sigma}_1^{1/2}) = P_j(\boldsymbol{v}, \boldsymbol{\Sigma}_2^{1/2})$ for all $\boldsymbol{v} \in \mathbb{R}^K$ and $j \in [K]$.

We present this property because, as we show below, if it holds it allows a simplification of the neural net architecture. Notably, MNP models satisfy this property, so any permissible factorization rule defining $\boldsymbol{\Sigma} \mapsto \boldsymbol{\Sigma}^{1/2}$ produces an equivalent model. More generally, this property is satisfied by the broader class of models with elliptically symmetric errors, including multivariate-t.

We note, however, that our method applies much more broadly to any model that satisfies Assumption 1. For general distributions satisfying Assumption 1 (but not necessarily Definition 1), such as *iid* Gumbel, different factorizations of the same $\boldsymbol{\Sigma}$ can induce different higher-order moments of the transformed errors, resulting in distinct choice probabilities. Our proposed method remains applicable provided we employ the symmetric positive semidefinite square-root matrix. For this reason, the remainder of the paper suppresses the dependence on the factorization rule, denoting the choice probabilities as $P_j(\boldsymbol{v}, \boldsymbol{\Sigma})$.

3.3 Inferential Goals

Given observations $\{(y_i, \mathbf{X}_i)\}_{i=1}^n$, where $y_i \in \{1, \ldots, K\}$ is the observed choice and $\mathbf{X}_i$ represents the covariates, we seek to estimate a finite-dimensional parameter vector $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$. The parameter vector includes the coefficients governing the deterministic utilities, the free parameters governing $\boldsymbol{\Omega}$ (and hence $\boldsymbol{\Sigma}$), and, where applicable, additional parameters governing $f_{\boldsymbol{\epsilon}^*}$, such as the degrees-of-freedom parameter ν. Thus, for observation i, the model implies a deterministic utility vector $\boldsymbol{v}_i(\mathbf{X}_i, \boldsymbol{\theta})$ and scale matrix $\boldsymbol{\Sigma}_i(\mathbf{X}_i, \boldsymbol{\theta})$. Section 4.1 describes how these model-implied objects are subsequently transformed into the normalized inputs supplied to the emulator. It is worth noting that the emulator itself does not require

identification: it maps any specified $(\boldsymbol{v}, \boldsymbol{\Sigma})$ into choice probabilities. Identification matters only when using observed choices to estimate the parameters generating $(\boldsymbol{v}, \boldsymbol{\Sigma})$; we return to this issue in Section 5.2. The (averaged) log-likelihood function is

$$\ell_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \log P_{y_i} \left\{ \boldsymbol{v}_i(\mathbf{X}_i, \boldsymbol{\theta}), \boldsymbol{\Sigma}_i(\mathbf{X}_i, \boldsymbol{\theta}) \right\},$$

Maximum likelihood estimation requires maximizing $\ell_n(\boldsymbol{\theta})$ over Θ, which in turn requires repeated evaluation of $\ell_n(\boldsymbol{\theta})$ and its gradient $\nabla_{\boldsymbol{\theta}} \log P_j(\boldsymbol{v}, \boldsymbol{\Sigma})$. Similarly, Bayesian methods such as Hamiltonian Monte Carlo require many repeated evaluations of both $\ell_n(\boldsymbol{\theta})$ and $\nabla_{\boldsymbol{\theta}} \log P_j(\boldsymbol{v}, \boldsymbol{\Sigma})$ throughout the estimation process.

Our goal is to construct a neural network emulator that provides fast, accurate approximations of the true choice probabilities, along with analytic gradients via automatic differentiation. The emulator is trained once on simulated data spanning the relevant input space, after which it can be used for rapid inference on many datasets and models.

4 Emulator Design

This section outlines the emulator preprocessing transformations, neural network architecture, training procedure, and inference process. The emulator must satisfy four desiderata:

1. **Respect invariance properties.** The emulator should respect the invariance properties described in Section 3.2: invariance with respect to location shifts and scale transformations, and equivariance with respect to permutations.
2. **Provide smooth approximations.** The emulator should be differentiable with respect to its inputs, enabling gradient-based optimization and automatic differentiation.
3. **Generalize across specifications.** A single trained emulator should work for diverse utility and covariance structures.
4. **Be computationally efficient.** Emulator evaluation should be fast enough to enable routine use in estimation.

We achieve these goals through a combination of preprocessing transformations and a carefully designed neural network architecture.

4.1 Preprocessing: Centering and Scaling

We address location and scale invariance through preprocessing transformations that project the inputs onto a canonical subspace.

Centering. Define the centering matrix $\mathbf{M} = \mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top$, which we employ to project the vector of utilities, $\boldsymbol{U} \in \mathbb{R}^K$, onto the subspace orthogonal to $\mathbf{1}_K$, resulting in transformed utilities $\tilde{\boldsymbol{U}} = \mathbf{M}(\boldsymbol{v} + \boldsymbol{\epsilon}) = \boldsymbol{v} + \boldsymbol{\epsilon} - \mathbf{1}_K(\bar{v} + \bar{\epsilon}) = \boldsymbol{U} - \mathbf{1}_K\bar{U}$, where $\bar{v} = \frac{1}{K}\sum_{j=1}^K v_j$, $\bar{\epsilon} = \frac{1}{K}\sum_{j=1}^K \epsilon_j$, and $\bar{U} = \bar{v} + \bar{\epsilon}$. These transformed utilities sum to zero within each realization: $\sum_{j=1}^K \tilde{U}_j = 0$.

By the location invariance property in (3), this transformation does not alter the choice probabilities. Further, under the assumed structure of $\boldsymbol{\Sigma}$ in (1), the transformation $\boldsymbol{\Sigma} \mapsto \mathbf{M}\boldsymbol{\Sigma}\mathbf{M}^\top$ can be inverted as $\boldsymbol{\Omega} = \mathbf{D}\mathbf{M}\boldsymbol{\Sigma}\mathbf{M}^\top\mathbf{D}^\top - \mathbf{1}_{K-1}\mathbf{1}_{K-1}^\top$, where $\mathbf{D}$ is the differencing matrix $\mathbf{D} = \begin{bmatrix} -\mathbf{1}_{K-1} & \mathbf{I}_{K-1} \end{bmatrix}$. Consequently, distinct models remain separated and we can represent a transformed model with the following modified parameters:

$$\tilde{\boldsymbol{v}} = \mathbf{M}\boldsymbol{v} = \boldsymbol{v} - \mathbf{1}_K\bar{v}, \qquad \tilde{\boldsymbol{\Sigma}} = \mathbf{M}\boldsymbol{\Sigma}\mathbf{M}.$$

Scaling. We normalize by the trace of $\tilde{\boldsymbol{\Sigma}}$, producing transformed utilities $\boldsymbol{U}^* = \sqrt{C^*}\tilde{\boldsymbol{U}}$, where $C^* = K/\mathrm{tr}\big(\tilde{\boldsymbol{\Sigma}}\big)$. By the scale invariance property in (4), this normalization does not affect the choice probabilities. Further, the map $(\tilde{\boldsymbol{v}}, \tilde{\boldsymbol{\Sigma}}) \mapsto \left(C^*, \sqrt{C^*}\tilde{\boldsymbol{v}}, C^*\tilde{\boldsymbol{\Sigma}}\right)$ is invertible; thus, the final transformed model can be represented with parameters:

$$C^* = K/\mathrm{tr}\big(\tilde{\boldsymbol{\Sigma}}\big), \qquad \boldsymbol{v}^* = \sqrt{C^*}\,\tilde{\boldsymbol{v}}, \qquad \boldsymbol{\Sigma}^* = C^*\,\tilde{\boldsymbol{\Sigma}}.$$

We can omit the case $\mathrm{tr}(\tilde{\boldsymbol{\Sigma}}) = 0$ as this results in known, deterministic choices.[5] The complete transformation produces normalized parameters $(C^*, \boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$ satisfying:

1. $\sum_{j=1}^K v_j^* = 0$ (centered deterministic utilities),
2. $\boldsymbol{\Sigma}^*\mathbf{1}_K = \mathbf{0}$ (centered scale matrix),
3. $\mathrm{tr}(\boldsymbol{\Sigma}^*) = K$ (scale-normalized),
4. $\boldsymbol{\Sigma}^*$ is positive semidefinite with rank at most $K - 1$.

This normalization reduces the size of the emulator's input space, which accelerates learning and improves generalization. Importantly, this transformation commutes with permutation matrices: for any permutation matrix $\mathbf{P}_\pi$, applying the transformation to $(\mathbf{P}_\pi\boldsymbol{v}, \mathbf{P}_\pi\boldsymbol{\Sigma}\mathbf{P}_\pi^\top)$ yields $(\mathbf{P}_\pi\boldsymbol{v}^*, \mathbf{P}_\pi\boldsymbol{\Sigma}^*\mathbf{P}_\pi^\top)$. Thus, the preprocessing preserves permutation equivariance.

In factorization-invariant models (Definition 1), the pair $(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$ is sufficient to identify the

[5] Assumption 1 implies $\mathrm{Cov}(\boldsymbol{U} - \mathbf{1}_K\bar{U}) = C\mathbf{M}\boldsymbol{\Sigma}\mathbf{M} = C\tilde{\boldsymbol{\Sigma}}$ for some $C > 0$. Thus, $\mathrm{tr}(\tilde{\boldsymbol{\Sigma}}) = 0$ implies that $\epsilon_j = \bar{\epsilon}$ with probability one for all $j \in [K]$.

choice probabilities, so the scale factor C^* is not needed. For general exchangeable errors (e.g., Gumbel errors), we need two additional pieces of information: C^* and an indicator for which alternative corresponds to base alternative (index 1) in the level parameterization defined in (1). The latter is required to invert the preprocessing transformation to an equivalent model in the original parameterization; without it, the preprocessing transformation can result in information loss.[6]

4.2 Neural Network Architecture

After preprocessing, we construct a neural network that maps $(C^*, \boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$ to choice probabilities while maintaining permutation equivariance. The architecture consists of three components: a per-alternative encoder, permutation-equivariant layers, and an output layer. Below, we interpret $\boldsymbol{\Sigma}^*$ as a covariance matrix to simplify the exposition, referring to its diagonal and off-diagonal elements as variances and covariances, respectively.

Per-alternative encoder. For each alternative $j \in \{1, \ldots, K\}$, we construct a representation $\boldsymbol{z}_j \in \mathbb{R}^{d_z}$ that captures how alternative j relates to all other alternatives. This representation is built from two complementary DeepSet networks—a diagonal DeepSet and an off-diagonal DeepSet—whose outputs are combined with alternative j's own features. The diagonal DeepSet processes pairwise relationships between j and each other alternative, while the off-diagonal DeepSet summarizes the covariance structure among alternatives other than j. In the descriptions below, the features labeled as "base inputs" are sufficient for universal approximation of choice probabilities (see Section 5.1). We include additional features to improve the expressivity of the emulator, as we find this improves its predictions.

Diagonal DeepSet. For each pair (j, k) with $k \neq j$, we construct a feature vector $\boldsymbol{d}_{jk}$ containing differentiable functions of the utilities and covariances associated with alternatives j and k. The base inputs are the utilities v_j^* and v_k^*, the variances Σ_{jj}^* and Σ_{kk}^*, and the covariance Σ_{jk}^*. From these, we derive additional feature transformations that facilitate learning, including standard deviations $\sigma_j = \sqrt{\Sigma_{jj}^*}$ and $\sigma_k = \sqrt{\Sigma_{kk}^*}$, the correlation $\rho_{jk} = \Sigma_{jk}^*/(\sigma_j \sigma_k)$, and the standardized utility difference (z-score):

$$z_{jk} = \frac{v_j^* - v_k^*}{\sqrt{\Sigma_{jj}^* + \Sigma_{kk}^* - 2\Sigma_{jk}^*}}.$$

[6] The inversion formula $\boldsymbol{\Omega} = \mathbf{D}\mathbf{M}\boldsymbol{\Sigma}\mathbf{M}^\top\mathbf{D}^\top - \mathbf{1}_{K-1}\mathbf{1}_{K-1}^\top$ implicitly requires knowledge of which alternative occupies the first position. This information allows us to recover $\boldsymbol{\Sigma}$. Although we cannot recover $\boldsymbol{v}$, we can recover $\tilde{\boldsymbol{v}} = \boldsymbol{v}^*/\sqrt{C^*}$, which differs from $\boldsymbol{v}$ by an additive constant and, hence, results in equivalent choices.

For models that are not factorization-invariant, we must also include an indicator function, $1_{\{k=1\}}$, to ensure that the preprocessing transformation does not result in information loss.[7] Appendix C.1 provides the full set of input features for each module of the encoder. It also describes transformations we applied to these features to reduce sensitivity to outliers.

The diagonal DeepSet processes these features using the architecture of Zaheer et al. (2017):

$$\boldsymbol{h}_j^{\text{diag}} = \rho_{\text{diag}}\left(\sum_{k\neq j} \phi_{\text{diag}}(\boldsymbol{d}_{jk})\right),$$

where ϕ_{diag} is a multi-layer perceptron (MLP) applied to each pair, the sum aggregates over all $k \neq j$, and ρ_{diag} is another MLP that produces a learned nonlinear representation of the sum. We use the notation ϕ_{diag} and ρ_{diag} following Zaheer et al. (2017); the MLPs ρ_{diag} and ρ_{off} (defined below) can be distinguished from the correlation ρ_{jk} by their subscripts.

Singh et al. (2023) also employ a variant of the DeepSet architecture. However, their method assumes iid errors and operates on problem-specific features, which requires their neural network to be modified and retrained to accommodate new data structures. In contrast, our architecture relies on generic RUM model inputs, such as deterministic utilities.

Off-diagonal DeepSet. For each pair (k, l) with $k < l$ and $k, l \neq j$, we construct a feature vector $\boldsymbol{o}_{kl}$ containing differentiable, symmetric functions comparing alternatives k and l; we require symmetry because the pairs do not possess a natural ordering. The base input is Σ^*_{kl}, and we include other derived features, such as the correlation ρ_{kl}, squared utility difference $(v^*_k - v^*_l)^2$, the squared z-score z^2_{kl}, and the variance sum $\Sigma^*_{kk} + \Sigma^*_{ll}$.

The off-diagonal DeepSet has an analogous structure to that of the diagonal DeepSet:

$$\boldsymbol{h}_j^{\text{off}} = \rho_{\text{off}}\left(\sum_{\substack{k<l \\ k,l\neq j}} \phi_{\text{off}}(\boldsymbol{o}_{kl})\right).$$

Combining network. The outputs of both DeepSets are concatenated with additional "pass-through features" $\boldsymbol{s}_j$, such as C^*, $1_{\{j=1\}}$, v^*_j, σ_j, and summary statistics of alternative j's covariances with other alternatives. Of these features, only C^* is required for universal approximation, but it can be omitted for factorization-invariant models. These concate-

[7] Section 4.1 describes how the transformed parameters can be inverted to an equivalent model with parameters $\tilde{\boldsymbol{v}}$ and $\boldsymbol{\Sigma}$. However, this transformation can only be applied if we know which alternative occupies the first position. The indicator function $1_{\{k=1\}}$ provides exactly this information.

nated features are then passed to the combining MLP, ζ, which produces the per-alternative representation $\boldsymbol{z}_j = \zeta\big(\boldsymbol{h}_j^{\text{diag}}, \boldsymbol{h}_j^{\text{off}}, \boldsymbol{s}_j\big)$.

Permutation-equivariant layers. The per-alternative representations are stacked into a matrix $\mathbf{Z} \in \mathbb{R}^{K \times d_z}$ and processed through linear permutation-equivariant layers (Zaheer et al., 2017), allowing information exchange across alternatives. These layers take the form

$$L(\mathbf{Z}) = \sigma\left(\mathbf{Z}\mathbf{A} + \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\mathbf{Z}\mathbf{B} + \mathbf{1}_K\boldsymbol{c}^\top\right),$$

where $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$ and $\boldsymbol{c} \in \mathbb{R}^{d_{\text{out}}}$ are learnable parameters, and σ is a smooth activation function. The first term applies a per-alternative transformation, the second aggregates information across alternatives, and the third adds a shared bias (intercept).

Output layer. The final layer produces one logit per alternative:

$$\textbf{logit} = \mathbf{Z}^{(L)}\boldsymbol{a},$$

where $\mathbf{Z}^{(L)}$ is the output of the last equivariant layer and $\boldsymbol{a} \in \mathbb{R}^{d_L}$ so that the output **logit** is a vector in $\mathbb{R}^K$. We omit the other terms because they produce a constant shift that does not affect the output probabilities.

Choice probabilities are obtained by applying the softmax function to the output layer:

$$\hat{P}_j = \frac{\exp(\text{logit}_j)}{\sum_{k=1}^K \exp(\text{logit}_k)}.$$

For any vector of valid choice probabilities $(P_1, \ldots, P_K)$ on the interior of the probability simplex Δ^{K-1}, there exists a corresponding vector of logits yielding $(P_1, \ldots, P_K)$ via softmax, so this parameterization entails no loss of generality. Figure 1 illustrates the architecture.

Activation functions. Throughout the architecture, we use smooth (infinitely differentiable) activation functions. This ensures that the emulator output is a smooth function of the inputs, enabling accurate gradient computation for downstream model fitting. We found that the Swish activation function $\text{Swish}(x) = x/(1+e^{-x})$ works well empirically (Ramachandran et al., 2018), though other smooth activation functions could also be applied, such as tanh or softplus (Dugas et al., 2000).

Computational complexity. A naive implementation of the per-alternative encoder that recomputes the off-diagonal DeepSet sum $\sum_{k<l:\, k,l\neq j} \phi^{\text{off}}(o_{kl})$ separately for each alternative

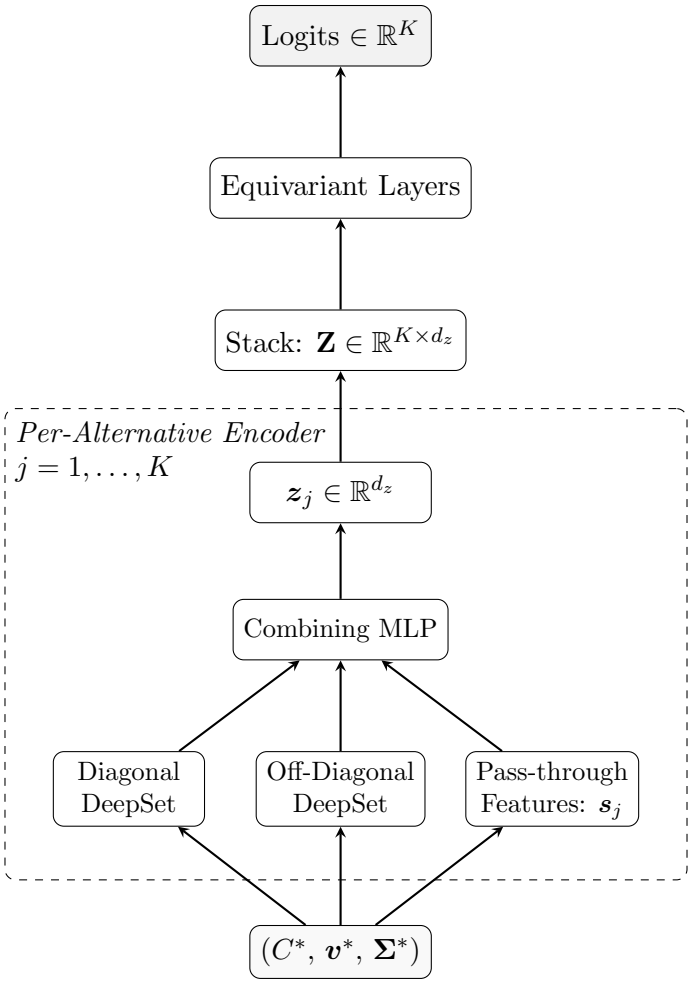


Figure 1: Architecture of the neural network emulator. For each alternative j, the per-alternative encoder processes diagonal features (relating j to each other alternative), off-diagonal features (pairwise features among alternatives other than j), and alternative j's own features. These are combined via a combining MLP to produce representation $\boldsymbol{z}_j$. Representations for all alternatives are stacked and processed through permutation-equivariant layers to produce the final logits.

j would require $O(K^3)$ operations. But most of these computations are redundant. Instead, we compute each pair contribution $\phi^{\text{off}}(o_{kl})$ once, form the total sum over all pairs, and then subtract the terms involving alternative j to obtain the required sum for that alternative. This reduces the complexity to $O(K^2)$.

Network size. The architecture can be compact. In our simulations, we use networks with 8–64 hidden units and 0–2 hidden layers per component. In general, larger values of K require larger networks for accurate approximation.

Relation to Blum-Smith et al. (2025). Our architecture builds on the DS-CI architecture of Blum-Smith et al. (2025), which is designed to approximate permutation-invariant scalar

functions of symmetric matrices. In contrast, our emulator approximates the full vector of choice probabilities, which depends jointly on the systematic utility vector $\boldsymbol{v}$ and the matrix $\boldsymbol{\Sigma}$ governing the error distribution. If the alternatives are relabeled, the choice probabilities must be relabeled in the same way, so the choice-probability map is permutation-equivariant rather than permutation-invariant. These differences motivate three architectural changes. First, the equivariance property requires us to construct a separate representation for each alternative j, distinguishing it from its rivals $k \neq j$ in the per-alternative encoder. Second, these separate representations and the presence of $\boldsymbol{v}$ require us to pass v_k^* and Σ_{jk}^* to the Diagonal DeepSet (in addition to the diagonals, Σ_{kk}^*, which DS-CI also requires). Third, whereas Blum-Smith et al. (2025) require a "linking feature" that combines diagonal and off-diagonal elements, our centering procedure makes this feature redundant. However, we require a different scalar-valued feature, C^*, for universal approximation (except on factorization-invariant models). Rather than processing C^* in its own submodule, we simply pass it as an additional feature to the combining MLP.

4.3 Training Procedure

We train the emulator using simulated data spanning a diverse range of utility and covariance configurations, employing Sobolev training to ensure accurate approximation of both choice probabilities and their gradients.

Data generation. To generate each training example for our simulations in Section 6 involving the MNP model, we follow the procedure below:

1. **Error covariance.** Draw a covariance matrix $\boldsymbol{\Omega}$ from a Wishart$(\mathbf{I}_{K-1}/(K+1), K+1)$ distribution and form $\boldsymbol{\Sigma}$ according to (1).
2. **Raw deterministic utilities.** Draw independent utilities $\boldsymbol{v}' \sim \mathcal{N}(\mathbf{0}, 4\mathbf{I}_K)$.
3. **Deterministic utility covariance.** Draw a second covariance matrix $\boldsymbol{\Omega}'$ (independently from the same distribution as $\boldsymbol{\Omega}$) to form $\boldsymbol{\Sigma}'$ as in (1), draw $\omega \sim \text{Uniform}(0, 1)$, and form the convex combination $\boldsymbol{\Sigma}'' = \omega\boldsymbol{\Sigma} + (1-\omega)\boldsymbol{\Sigma}'$.
4. **Transformed deterministic utilities.** Transform the utilities to have covariance structure correlated with $\boldsymbol{\Sigma}''$: $\boldsymbol{v} = (\boldsymbol{\Sigma}'')^{1/2}\boldsymbol{v}'$. This yields $\text{Cov}(\boldsymbol{v}) = 4\,\boldsymbol{\Sigma}''$.[8]
5. **Preprocessing.** Apply the preprocessing transform to obtain $(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$ from $(\boldsymbol{v}, \boldsymbol{\Sigma})$.

[8] If desired, we can also center $\boldsymbol{v}$ as this results in an equivalent model.

6. **Simulated choices.** Simulate 10^6 choices from the MNP model with utilities $\boldsymbol{v}$ and scale matrix $\boldsymbol{\Sigma}$ (or an equivalent parameterization).

Steps 3 and 4 ensure diversity in the relationship between deterministic utilities and error covariances. When $\omega = 1$, the utilities are perfectly aligned with the covariance structure; when $\omega = 0$, they are independent. Intermediate values produce partial correlation.

For error distributions that are not factorization-invariant, we must additionally compute C^* in step 5 and replace the MNP model with the assumed error distribution in step 6. The other choices listed above, including the specific distributions and number of choices (10^6), may also be modified based on the problem domain and available computational resources.

Loss function. We employ Sobolev training (Czarnecki et al., 2017), which augments the standard likelihood loss with a gradient-matching penalty. The total loss is:

$$\mathcal{L}(\boldsymbol{\phi}) = \mathcal{L}_{\text{CE}}(\boldsymbol{\phi}) + \lambda_{\text{grad}} \mathcal{L}_{\text{grad}}(\boldsymbol{\phi}),$$

where $\boldsymbol{\phi}$ represents the network parameters and $\lambda_{\text{grad}} > 0$ controls the weight of the gradient-matching term. The cross-entropy component is the multinomial negative log-likelihood (up to constants):

$$\mathcal{L}_{\text{CE}}(\boldsymbol{\phi}) = -\sum_{j=1}^{K} \hat{P}_j^{\text{sim}} \log \hat{P}_j(C^*, \boldsymbol{v}^*, \boldsymbol{\Sigma}^*; \boldsymbol{\phi}),$$

where $\hat{P}_j^{\text{sim}} = n_j / \sum_{k=1}^{K} n_k$ is the simulated choice frequency and $\hat{P}_j(C^*, \boldsymbol{v}^*, \boldsymbol{\Sigma}^*; \boldsymbol{\phi})$ is the emulator choice probability of alternative j.

For the gradient-matching component, we follow the stochastic Sobolev training approach of Czarnecki et al. (2017), which avoids computing full Jacobian matrices by instead matching directional derivatives along random directions. Let $\boldsymbol{d} = (\boldsymbol{d}_{\boldsymbol{v}}, \text{vech}(\mathbf{D}_{\mathbf{L}}))$ denote a random direction, where $\boldsymbol{d}_{\boldsymbol{v}} \in \mathbb{R}^K$, $\mathbf{D}_{\mathbf{L}} \in \mathbb{R}^{(K-1)\times(K-1)}$ is lower-triangular, and $\text{vech}(\cdot)$ denotes the half-vectorization (lower triangular elements). The gradient-matching loss is:

$$\mathcal{L}_{\text{grad}}(\boldsymbol{\phi}) = \sum_{j=1}^{K} \hat{P}_j^{\text{sim}} \left(\nabla_{\boldsymbol{d}} \log \hat{P}_j - \nabla_{\boldsymbol{d}} \log P_j^{\text{target}} \right)^2,$$

where $\nabla_{\boldsymbol{d}}$ denotes the directional derivative along $\boldsymbol{d}$. Weighting by the simulated choice frequencies $\hat{P}_j^{\text{sim}}$ focuses learning on alternatives that are frequently chosen, whose gradients most influence parameter estimation in downstream inference tasks.

Constraint-respecting directions. The inputs $(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$ lie on the constrained set:

$$\mathcal{X}_K = \left\{ (\boldsymbol{v}^*, \boldsymbol{\Sigma}^*) : \sum_{j=1}^{K} v_j^* = 0,\ \boldsymbol{\Sigma}^* \mathbf{1}_K = \mathbf{0},\ \operatorname{tr}(\boldsymbol{\Sigma}^*) = K,\ \boldsymbol{\Sigma}^* \succeq \mathbf{0} \right\}. \tag{5}$$

To ensure these constraints are respected, we compute directions in the latent space of $\boldsymbol{v}$ and $\mathbf{L} = \operatorname{chol}(\boldsymbol{\Omega})$, the lower Cholesky factor of $\boldsymbol{\Omega}$.[9] We generate directions as follows:

1. Sample $\boldsymbol{d}_{\boldsymbol{v}} \sim \mathcal{N}(\mathbf{0}_K, \mathbf{I}_K)$.
2. Sample a lower-triangular matrix, $\mathbf{D}_{\mathbf{L}}$, with nonzero entries in row j distributed as $\mathcal{N}(\mathbf{0}_j, \mathbf{I}_j / j)$.
3. Set $\boldsymbol{d} = (\boldsymbol{d}_{\boldsymbol{v}}, \operatorname{vech}(\mathbf{D}_{\mathbf{L}}))$ and normalize: $\boldsymbol{d} \leftarrow \boldsymbol{d} / \|\boldsymbol{d}\|$.

In step 2, we divide the variances by j to account for the number of nonzero entries in each row, maintaining a comparable scale across rows and columns of $\boldsymbol{\Omega} = \mathbf{L}\mathbf{L}^\top$. If desired, we can remove directions that do not contribute to choice probabilities, but doing so does not affect the directional derivatives.[10]

Computing directional derivatives. Define the raw values $\boldsymbol{r} = (\boldsymbol{v}, \mathbf{L})$ and let $\boldsymbol{t}$ denote the transformation from $\boldsymbol{r}$ to $(C^*, \boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$. We approximate the emulator's directional derivatives via finite differences as

$$\nabla_{\boldsymbol{d}} \log \hat{P}_j = \frac{\log \hat{P}_j \left\{ \boldsymbol{t}(\boldsymbol{r} + \epsilon \boldsymbol{d}); \boldsymbol{\phi} \right\} - \log \hat{P}_j \left\{ \boldsymbol{t}(\boldsymbol{r} - \epsilon \boldsymbol{d}); \boldsymbol{\phi} \right\}}{2\epsilon},$$

where $\epsilon > 0$ is a small step size (we use $\epsilon = 10^{-5}$). This approach is computationally efficient, requiring only two additional forward passes per direction. Alternatively, one could compute the directional derivative via automatic differentiation as in Czarnecki et al. (2017). Working with the latent parameters, $\boldsymbol{r} = (\boldsymbol{v}, \mathbf{L})$, guarantees that the resulting values of $\boldsymbol{v}^*$ and $\boldsymbol{\Sigma}^*$ obey the constraints of $\mathcal{X}_K$.

Target directional derivatives are computed from pre-stored Jacobians. During data generation, we compute and store the full Jacobians $\nabla_{\boldsymbol{r}} \log P_j^{\text{target}}$. During training, the target

[9] Note that $\mathbf{L}$ does not necessarily correspond with $\boldsymbol{\Omega}^{1/2}$. These matrix factors play different roles. Whereas $\boldsymbol{\Omega}^{1/2}$ is used within the choice model to create covariance relationships, $\mathbf{L}$ is used as a computational device to ensure that the gradients respect the positive semidefiniteness of $\boldsymbol{\Omega}$ and the resulting constraints on the set $\mathcal{X}_K$.

[10] For example, we can set $\boldsymbol{d}_{\boldsymbol{v}} \leftarrow \mathbf{M}\boldsymbol{d}_{\boldsymbol{v}}$. However, the removed direction is orthogonal to the Jacobian of the choice probabilities, so this operation has no effect on the directional derivatives.

directional derivative is obtained via the inner product $\nabla_{\boldsymbol{d}} \log P_j^{\text{target}} = \boldsymbol{d}^\top \nabla_{\boldsymbol{r}} \log P_j^{\text{target}}$.

Target gradient computation. Computing target gradients for Sobolev training requires differentiating through the choice simulation process. Since discrete choices are inherently non-differentiable, we employ a smooth relaxation using a temperature-scaled softmax. For each Monte Carlo draw r with utility vector $\boldsymbol{U}^{(r)} = \boldsymbol{v}^* + \boldsymbol{\epsilon}^{(r)}$, we approximate the hard choice indicator with:

$$\tilde{Y}_j^{(r)} = \frac{\exp(\tau U_j^{(r)})}{\sum_{k=1}^K \exp(\tau U_k^{(r)})},$$

where $\tau > 0$ is a temperature parameter. As $\tau \to \infty$, this converges to the hard choice indicator; for finite τ, it provides a differentiable approximation. We use $\tau = 5$ in our experiments, which provides a balance between gradient accuracy (low bias) and computational stability. The target log-probability is then:

$$\log P_j^{\text{target}} = \log\left(\frac{1}{R}\sum_{r=1}^{R} \tilde{Y}_j^{(r)}\right),$$

and the target gradient $\nabla_{\boldsymbol{r}} \log P_j^{\text{target}}$ is obtained via automatic differentiation through this expression. This approach is closely related to the "Gumbel–softmax trick" frequently used for gradient estimation with neural networks (Jang et al., 2017; Maddison et al., 2017).

Training details. Our training procedure makes use of a *replay buffer*, a common technique in deep learning (Mnih et al., 2015). It stores precomputed training examples including raw inputs $(\boldsymbol{v}, \mathbf{L})$, simulated choice frequencies, and target Jacobians. At each training step, we sample inputs from the replay buffer, sample directions, compute the loss, and update the network parameters via stochastic gradient descent. The replay buffer accelerates training by allowing reuse of expensive high-precision simulations and gradient computations. We set the gradient penalty weight $\lambda_{\text{grad}} = 10^{-4}$ in our experiments; empirically, we find that this value provides sufficiently well-behaved derivatives.

Multi-K training. Remarkably, none of the weight vectors and matrices has a dimension depending on K. The dimension K enters the neural network only through the input data dimension; thus, we can design a single emulator capable of processing multiple values of K simultaneously. The training procedures described above can then be adapted to iterate over a range of K values, say $K \in \{3, 4, 5\}$, and the resulting emulator can then learn to approximate choice probabilities for a varying number of choices (see Section 6.3). Under this configuration, we recommend including K as an element of $\boldsymbol{s}_j$.

4.4 Inference with the Trained Emulator

Once trained, the emulator enables rapid inference. Given a new dataset $\{(y_i, \mathbf{X}_i)\}_{i=1}^n$ and candidate parameter vector $\boldsymbol{\theta}$:

1. Compute the deterministic utility vector $\boldsymbol{v}_i(\mathbf{X}_i, \boldsymbol{\theta})$ and scale matrix $\boldsymbol{\Sigma}_i(\mathbf{X}_i, \boldsymbol{\theta})$.
2. Apply the preprocessing transformation to obtain $(C_i^*, \boldsymbol{v}_i^*, \boldsymbol{\Sigma}_i^*)$.
3. Evaluate the emulator to obtain estimated choice probabilities $\hat{P}_{ij}$ for all alternatives.
4. Compute the log likelihood $\hat{\ell}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \log \hat{P}_{i,y_i}$.

As the preprocessing transformation and neural network are both smooth functions, the gradients of the emulator log likelihood with respect to $\boldsymbol{\theta}$ can be computed exactly via automatic differentiation. This enables the use of gradient-based optimizers for maximum likelihood estimation, as well as gradient-based sampling methods such as Hamiltonian Monte Carlo for Bayesian inference. The smoothness of the emulator also facilitates computation of standard errors, which can be estimated via emulator gradients (see Sections 5.2–5.3).

5 Theory

This section establishes the theoretical foundations for the emulator architecture and the statistical properties of emulator-based discrete choice estimators.

5.1 Universal Approximation

We first show that our architecture can universally approximate choice probabilities. The key insight is that the base inputs processed by the per-alternative encoder's diagonal and off-diagonal DeepSets generically determine $(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$ up to the relevant symmetry group (i.e, permutations of rivals). In order to prove this result, we define, for each alternative j, the invariant function $g_j : \mathcal{X}_K \to (\text{Multiset}, \text{Multiset})$ by:

$$\mathcal{T}_j(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*) = \{\{(v_k^*, \Sigma_{kk}^*, \Sigma_{jk}^*) : k \neq j\}\}, \tag{6}$$

$$\mathcal{O}_j(\boldsymbol{\Sigma}^*) = \{\{\Sigma_{kl}^* : 1 \leq k < l \leq K,\, k \neq j,\, l \neq j\}\}, \tag{7}$$

$$g_j(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*) = (\mathcal{T}_j(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*), \mathcal{O}_j(\boldsymbol{\Sigma}^*)).$$

The multiset $\mathcal{T}_j$ collects information about all the rivals of focal alternative j. It consists of $K-1$ triples, each containing alternative k's utility, variance, and covariance with alternative

j, for all $k \neq j$. It is a multiset (rather than a set or ordered list) because the order does not matter, but repeated elements do matter. The multiset $\mathcal{O}_j$ collects information about the similarity of these rivals. It consists of the off-diagonal covariances among alternatives other than j. These two multisets, which serve as base inputs to the encoder, generically contain enough information to reconstruct the full normalized utility/covariance object $(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$, up to relabeling of the rivals, as we now show:[11]

Theorem 1 (Generic Separation)**.** *For each alternative $j \in \{1, \ldots, K\}$, there exists a closed, measure-zero set $\mathcal{B}_j \subset \mathcal{X}_K$ such that for any $(\boldsymbol{v}_1^*, \boldsymbol{\Sigma}_1^*), (\boldsymbol{v}_2^*, \boldsymbol{\Sigma}_2^*) \in \mathcal{X}_K \setminus \mathcal{B}_j$:*

$$g_j(\boldsymbol{v}_1^*, \boldsymbol{\Sigma}_1^*) = g_j(\boldsymbol{v}_2^*, \boldsymbol{\Sigma}_2^*) \implies (\boldsymbol{v}_1^*, \boldsymbol{\Sigma}_1^*) \text{ and } (\boldsymbol{v}_2^*, \boldsymbol{\Sigma}_2^*) \text{ are in the same } S_{K-1}^{(j)}\text{-orbit},$$

where S_K is the permutation group on $\{1, \ldots, K\}$ and $S_{K-1}^{(j)} = \{\pi \in S_K : \pi(j) = j\}$ is the subgroup of permutations that fixes alternative j.

Proof sketch. The proof, which we give in detail in Appendix A.8, proceeds in three steps:

Step 1: Genericity conditions. We define two genericity conditions on $(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$. First, the triples $(v_k^*, \Sigma_{kk}^*, \Sigma_{jk}^*)$ for $k \neq j$ are pairwise distinct as elements of $\mathbb{R}^3$. Second, for each rival $k \neq j$, there is a unique subcollection of $\mathcal{O}_j$, containing one covariance for each rival other than k, whose elements sum to $-\Sigma_{kk}^* - \Sigma_{kj}^*$. The row-sum constraint $\boldsymbol{\Sigma}^* \mathbf{1}_K = \mathbf{0}$ implies that the row-k subcollection, namely $\{\{\Sigma_{k\ell}^* : \ell \neq k, j\}\}$, has exactly this sum.

Step 2: Invariant reconstruction. When both genericity conditions hold, we can recover $(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$ from g_j up to the relabeling of alternatives other than j. First, order the triples in $\mathcal{T}_j$ lexicographically,[12] thereby assigning labels to the rivals. Next, for each rival k, use the row-sum restriction and the second genericity condition to identify the unique subcollection of $\mathcal{O}_j$ corresponding to the covariances between k and the other rivals. Finally, recover each rival–rival covariance $\Sigma_{k\ell}^*$ as the common element in the recovered subcollections for rivals k and ℓ. The remaining entries v_j^* and Σ_{jj}^* are pinned down by the normalization constraints $\sum_{\ell=1}^K v_\ell^* = 0$ and $\boldsymbol{\Sigma}^* \mathbf{1}_K = \mathbf{0}$.

Step 3: Measure-zero exceptional set. The set $\mathcal{B}_j$ where either condition fails is a finite union of zero sets of nontrivial analytic functions on $\mathcal{X}_K$ that was defined in (5). Compare

[11] It is worth noting that, while $\mathcal{T}_j$ and $\mathcal{O}_j$ do not explicitly include v_j^* or Σ_{jj}^*, these entries are pinned down by the normalization constraints $\sum_{\ell=1}^K v_\ell^* = 0$ and $\boldsymbol{\Sigma}^* \mathbf{1}_K = \mathbf{0}$.

[12] Lexicographic ordering sorts triples by their first coordinate, breaking ties by the second coordinate, and then by the third. Here it is used only to impose an arbitrary canonical labeling of the rivals.

with Definition 12 in Appendix A.7. By the principle that proper analytic subvarieties have measure zero (Mityagin, 2020), $\mathcal{B}_j$ has measure zero. □

This generic separation result, combined with the Stone–Weierstrass theorem, yields universal approximation. To show this, define the enlarged space $\overline{\mathcal{X}}_K = \mathcal{X}_K \times \mathbb{R}$, where the new dimension provides values of C^*, and form an enlarged measure-zero set in a similar manner: $\overline{\mathcal{B}}_j = \mathcal{B}_j \times \mathbb{R}$. To accommodate models that are not factorization invariant, we also must modify the function $\mathcal{T}_j$ as follows:

$$\overline{\mathcal{T}}_j(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*) = \{\{(v_k^*, \Sigma_{kk}^*, \Sigma_{jk}^*, 1_{\{k=1\}}) : k \neq j\}\}, \tag{8}$$

where we have added the element $1_{\{k=1\}}$. This additional element guarantees that the preprocessing transformation in Section 4.1 can be inverted to an equivalent model, allowing us to apply it without information loss. We then define the augmented invariant function

$$\overline{g}_j(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*, C^*) = (\overline{\mathcal{T}}_j(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*), \mathcal{O}_j(\boldsymbol{\Sigma}^*), C^*),$$

which, based on Theorem 1, separates parameters $(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*, C^*)$ up to permutations fixing alternatives 1 and j (which may coincide) outside of the measure-zero set $\overline{\mathcal{B}}_j$. We then have the following universal approximation property.

Theorem 2 (MLP Universal Approximation)**.** *Under Assumption 1, for each alternative j, the choice probability function P_j can be uniformly approximated on any compact subset of $\overline{\mathcal{X}}_K \setminus \overline{\mathcal{B}}_j$ by a multi-layer perceptron taking the components of $\overline{g}_j$ as input.*

Proof sketch. We apply the proof strategy of Blum-Smith et al. (2025). The choice probability P_j is continuous and $S_{K-2}^{(1,j)}$-invariant, where $S_{K-2}^{(1,j)}$ is the subgroup of permutations fixing alternatives 1 and j. Define continuous functions encoding the components of $\overline{g}_j$ (except the indicators, $1_{\{k=1\}}$, which are used only for determining indices). By Theorem 1, these functions separate points in the quotient space $(\overline{\mathcal{X}}_K \setminus \overline{\mathcal{B}}_j)/S_{K-2}^{(1,j)}$. In this quotient space, two normalized choice environments are treated as the same point if they differ only by an element of $S_{K-2}^{(1,j)}$, that is, by a relabeling of the alternatives other than 1 and j. Since $(\overline{\mathcal{X}}_K \setminus \overline{\mathcal{B}}_j)$ is a subspace of a finite-dimensional Euclidean space, it is Hausdorff. Moreover, since $S_{K-2}^{(1,j)}$ is finite and acts by coordinate permutations, standard results imply that the quotient space is also Hausdorff (Bredon, 1972). Hence, by the Stone–Weierstrass theorem, the algebra generated by functions encoding $\overline{g}_j$ is dense in the continuous functions on compact subsets

of this quotient space.[13] Pulling back to $(\overline{\mathcal{X}}_K \setminus \overline{\mathcal{B}}_j)$ and applying the universal approximation theorem for MLPs (Cybenko, 1989; Hornik et al., 1989) yields the result.[14] See Appendix A.8 for details. □

Theorem 2 shows that no information beyond the invariant summaries $\overline{g}_j$ is needed to approximate P_j on compact subsets of $\overline{\mathcal{X}}_K \setminus \overline{\mathcal{B}}_j$. The remaining question is whether the particular architecture in Section 4.2 is rich enough to approximate the required functions of this summary. The following corollary answers this question:

Corollary 1 (Architecture Universal Approximation)**.** *Let $\overline{\mathcal{B}} = \cup_{j=1}^{K} \overline{\mathcal{B}}_j$. Under Assumption 1, for each alternative j, the choice probability function P_j can be uniformly approximated on any compact subset of $\overline{\mathcal{X}}_K \setminus \overline{\mathcal{B}}_j$ by the per-alternative encoder. Further, the vector of choice probabilities $(P_1, \ldots, P_K)$ can be uniformly approximated on any compact subset of $\overline{\mathcal{X}}_K \setminus \overline{\mathcal{B}}$ by the neural network architecture of Section 4.2.*

Proof. Theorem 2 shows that, for each alternative j, P_j can be uniformly approximated on compact subsets of $\overline{\mathcal{X}}_K \setminus \overline{\mathcal{B}}_j$ by a continuous function of $\overline{g}_j$. The per-alternative encoder is constructed to process precisely these values: the diagonal DeepSet processes the values in (8), the off-diagonal DeepSet processes those in (7), and the combining MLP processes the corresponding representations with C^*. The first statement of corollary 1 therefore follows from Theorem 2 and the universal approximation property of DeepSets (Zaheer et al., 2017). For the second statement, first note that the preceding argument applies to $\overline{\mathcal{B}} = \cup_{j=1}^{K} \overline{\mathcal{B}}_j$ for all K alternatives simultaneously. Further, Segol and Lipman (2020) show that linear equivariant layers composed with nonlinear activation functions can uniformly approximate continuous permutation-equivariant functions on compact subsets of the feature space. Thus, the full architecture can uniformly approximate the vector of choice probabilities on compact subsets of $\overline{\mathcal{X}}_K \setminus \overline{\mathcal{B}}$. □

Remark 1. The exceptional set $\overline{\mathcal{B}}$ has measure zero, so it does not affect approximation in practice: for any probability distribution over $\overline{\mathcal{X}}_K$ that is absolutely continuous with respect to Lebesgue measure, inputs lie in $\overline{\mathcal{B}}$ with probability zero.

[13] The usual Stone–Weierstrass theorem applies to compact Hausdorff spaces. This condition is not automatic after passing to a quotient: when multiple original points are combined into a single quotient-point, it can become impossible to separate distinct quotient-points by disjoint open neighborhoods.

[14] Here "pulling back" means composing a quotient-space function with the natural projection from the original space to the quotient. If q maps each $(C^*, \boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$ to its equivalence class under rival relabeling, then an approximation $\widetilde{F}$ on the quotient gives the corresponding approximation $\widetilde{F} \circ q$ on $\mathcal{X}_K \setminus \mathcal{B}_j$.

Remark 2. The values C^* and $1_{k=1}$ are not required for models that are factorization invariant. In this special case, the proofs can be simplified because it is sufficient to show generic separation up to the symmetry group $S_{K-1}^{(j)}$.

Remark 3. Corollary 1 guarantees that the per-alternative encoder alone is sufficient for universal approximation. The inclusion of the final equivariant layers ensures that the estimated probabilities sum to one and improves the expressivity of the neural network.

Remark 4. For very large choice sets, in order to reduce computation time, one may opt for a simpler architecture that applies the per-alternative encoder to only the chosen alternative with no equivariant layers, resulting in a single output, P_j. This modification maintains the universal approximation property; however, it relaxes the sum-to-one constraint in return for computational efficiency.

Remark 5. Corollary 1 can be extended to hold simultaneously for a finite set of K values under a suitable extension of the space $\overline{\mathcal{X}}_K$ and its exceptional sets. Section 6.3 presents numerical results showing that the proposed architecture can provide accurate approximations for multiple values of K simultaneously. This makes it simple to train the emulator to handle contexts where the choice set size varies across observations.

5.2 Consistency and Asymptotic Normality

We now establish that emulator-based estimators inherit the asymptotic properties of exact maximum likelihood estimators under appropriate conditions.

Let $\{(y_i, \mathbf{X}_i)\}_{i=1}^n$ be iid observations from the true model with parameter $\boldsymbol{\theta}_0 \in \Theta$. Define the true and emulator log-likelihood contributions as

$$
\begin{aligned}
m(y, \mathbf{X}; \boldsymbol{\theta}) &= \log P_y \left\{ \boldsymbol{v}(\mathbf{X}, \boldsymbol{\theta}), \boldsymbol{\Sigma}(\mathbf{X}, \boldsymbol{\theta}) \right\}, \\
\hat{m}_n(y, \mathbf{X}; \boldsymbol{\theta}) &= \log \hat{P}_y \left\{ C^*(\mathbf{X}, \boldsymbol{\theta}), \boldsymbol{v}^*(\mathbf{X}, \boldsymbol{\theta}), \boldsymbol{\Sigma}^*(\mathbf{X}, \boldsymbol{\theta}); \boldsymbol{\phi}_n \right\},
\end{aligned}
$$

with corresponding score functions

$$
\psi(y, \mathbf{X}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} m(y, \mathbf{X}; \boldsymbol{\theta}), \qquad \hat{\psi}_n(y, \mathbf{X}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \hat{m}_n(y, \mathbf{X}; \boldsymbol{\theta}).
$$

The true and emulator (averaged) log likelihoods are $\ell_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n m(y_i, \mathbf{X}_i; \boldsymbol{\theta})$ and $\hat{\ell}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \hat{m}_n(y_i, \mathbf{X}_i; \boldsymbol{\theta})$. Define the MLE and emulator-based estimator as

$$
\tilde{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta} \in \Theta} \ell_n(\boldsymbol{\theta}), \qquad \hat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta} \in \Theta} \hat{\ell}_n(\boldsymbol{\theta}).
$$

We require that the emulator log likelihood converges to the true log likelihood.

Assumption 2 (Emulator Approximation Quality)**.** For some $\alpha \geq 0$, the emulator satisfies

$$\sup_{\boldsymbol{\theta} \in \Theta} \left| \hat{\ell}_n(\boldsymbol{\theta}) - \ell_n(\boldsymbol{\theta}) \right| = o_p(n^{-\alpha}).$$

The n referenced in Assumption 2 is the sample size of the observed data. Consequently, this assumption involves an asymptotic regime in which the emulator becomes increasingly accurate as $n \to \infty$, which is achievable by progressively (a) increasing the number of simulated training examples and (b) increasing the neural network complexity (i.e., the number of layers and hidden units). Appendix B.1 provides more primitive sufficient conditions for Assumption 2 with references to closely related work on convergence rates of neural networks and their usage in econometric theory. Notably, for models with analytic choice probabilities, we expect this assumption to require only $O(n^{\alpha+\delta})$ training examples for any $\delta > 0$.

We also require sufficient regularity of the true model.

Assumption 3 (Regularity)**.** The model satisfies:

(i) $\boldsymbol{\theta}_0$ is the unique maximizer of $\mathbb{E}\{m(y, \mathbf{X}; \boldsymbol{\theta})\}$ in the compact set Θ with $\boldsymbol{\theta}_0 \in \text{int}(\Theta)$ and $\mathbb{E}\{|m(y, \mathbf{X}; \boldsymbol{\theta}_0)|\} < \infty$;

(ii) $m(y, \mathbf{X}; \boldsymbol{\theta})$ is twice continuously differentiable in $\boldsymbol{\theta}$ for all $(y, \mathbf{X})$;

(iii) $\mathbb{E}\{\sup_{\boldsymbol{\theta} \in \Theta} \|\psi(y, \mathbf{X}; \boldsymbol{\theta})\|^2\} < \infty$;

(iv) $\mathbb{E}\{\sup_{\boldsymbol{\theta} \in \Theta} \|\nabla^2_{\boldsymbol{\theta}} m(y, \mathbf{X}; \boldsymbol{\theta})\|\} < \infty$;

(v) the Fisher information $\mathbf{J}(\boldsymbol{\theta}_0) = \mathbb{E}\left\{\psi(y, \mathbf{X}; \boldsymbol{\theta}_0)\psi(y, \mathbf{X}; \boldsymbol{\theta}_0)^\top\right\}$ is positive definite.

The unique-maximizer requirement in condition (i) is an identification condition. It presumes a parameterization in which the usual location and scale indeterminacies of random utility models have been eliminated. In particular, the location normalization must rule out observationally equivalent parameterizations that differ only by a common additive shift in utilities, while the scale normalization must rule out parameterizations that differ only by a common positive rescaling of utilities. The latter may be imposed, for example, by normalizing one utility coefficient, such as the price coefficient, or by imposing an appropriate restriction on the error covariance parameters. Conditional on these normalizations, the data must contain sufficient variation to identify the remaining parameters; for example, variation in alternative-specific prices can be essential (Keane, 1992).

Theorem 3 (Consistency)**.** *Under Assumptions 2 and 3,* $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$ *as* $n \to \infty$.

Proof. By Assumption 2 with $\alpha \geq 0$, $\sup_{\boldsymbol{\theta}} |\hat{\ell}_n(\boldsymbol{\theta}) - \ell_n(\boldsymbol{\theta})| \xrightarrow{p} 0$. Further, the regularity conditions of Assumption 3 are sufficient to apply the uniform law of large numbers (ULLN) so that $\sup_{\boldsymbol{\theta}} |\ell_n(\boldsymbol{\theta}) - \mathbb{E}\{m(y, \mathbf{X}; \boldsymbol{\theta})\}| \xrightarrow{p} 0$. The triangle inequality gives uniform convergence of $\hat{\ell}_n$ to $\mathbb{E}\{m(y, \mathbf{X}; \boldsymbol{\theta})\}$, which is uniquely maximized at $\boldsymbol{\theta}_0$ by Assumption 3(i). By the consistency theorem for extremum estimators (Newey and McFadden, 1994, Theorem 2.1), $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$. □

Under the stronger condition $\alpha \geq 1$, the emulator-based estimator is asymptotically normal.

Theorem 4 (Asymptotic Normality)**.** *Under Assumptions 2 and 3 with* $\alpha \geq 1$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left\{\mathbf{0}, \mathbf{J}(\boldsymbol{\theta}_0)^{-1}\right\}.$$

The proof, given in Appendix B.2, shows that $\hat{\boldsymbol{\theta}}_n$ is asymptotically equivalent to $\tilde{\boldsymbol{\theta}}_n$.

Remark 6 (Approximate Maximizers). If a maximum is not uniquely attained, one can work with approximate maximizers satisfying $\hat{\ell}_n(\hat{\boldsymbol{\theta}}_n) \geq \sup_{\boldsymbol{\theta}} \hat{\ell}_n(\boldsymbol{\theta}) - o_p(n^{-\alpha})$. The consistency and asymptotic normality results continue to hold; see Newey and McFadden (1994, Theorem 2.1) and Appendix B.2.

To estimate $\mathbf{J}(\boldsymbol{\theta}_0)$, define the true and emulator outer-product-of-scores estimators:

$$\mathbf{J}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \psi(y_i, \mathbf{X}_i; \boldsymbol{\theta})\psi(y_i, \mathbf{X}_i; \boldsymbol{\theta})^\top,$$
$$\hat{\mathbf{J}}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \hat{\psi}_n(y_i, \mathbf{X}_i; \boldsymbol{\theta})\hat{\psi}_n(y_i, \mathbf{X}_i; \boldsymbol{\theta})^\top.$$

Assumption 4 (Gradient Approximation Quality)**.** The emulator satisfies

$$\sup_{\boldsymbol{\theta} \in \Theta} \left\| \hat{\mathbf{J}}_n(\boldsymbol{\theta}) - \mathbf{J}_n(\boldsymbol{\theta}) \right\| \xrightarrow{p} 0 \quad \text{as } n \to \infty.$$

Assumption 4 is justified by Sobolev training, which penalizes discrepancies between emulator and target gradients. The penalization is required because training on log-likelihood values alone does not guarantee convergence of gradients. As with Assumption 2, see Appendix B.1 for more primitive sufficient conditions and additional discussion.

Proposition 1 (Consistent Estimation of Fisher Information)**.** *Under Assumptions 2, 3, and 4,* $\hat{\mathbf{J}}_n(\hat{\boldsymbol{\theta}}_n) \xrightarrow{p} \mathbf{J}(\boldsymbol{\theta}_0)$.

Proof. By the triangle inequality, $\|\hat{\mathbf{J}}_n(\hat{\boldsymbol{\theta}}_n) - \mathbf{J}(\boldsymbol{\theta}_0)\| \leq \|\hat{\mathbf{J}}_n(\hat{\boldsymbol{\theta}}_n) - \mathbf{J}_n(\hat{\boldsymbol{\theta}}_n)\| + \|\mathbf{J}_n(\hat{\boldsymbol{\theta}}_n) - \mathbf{J}(\hat{\boldsymbol{\theta}}_n)\| + \|\mathbf{J}(\hat{\boldsymbol{\theta}}_n) - \mathbf{J}(\boldsymbol{\theta}_0)\|$. The first term vanishes by Assumption 4. For the second term, the regularity conditions of Assumption 3 are sufficient to apply the ULLN to $\mathbf{J}_n$; thus, this term also converges in probability to zero. Finally, the third term vanishes by continuity of $\mathbf{J}$ and consistency of $\hat{\boldsymbol{\theta}}_n$. □

The emulator score $\hat{\psi}_n(y_i, \mathbf{X}_i; \boldsymbol{\theta})$ is computable via automatic differentiation through the smooth neural network. Standard errors for $\hat{\boldsymbol{\theta}}_n$ are obtained as the square roots of the diagonal elements of $\hat{\mathbf{J}}_n(\hat{\boldsymbol{\theta}}_n)^{-1}/n$.

An alternative estimator based on the Hessian, $\tilde{\mathbf{J}}_n = -\frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \hat{m}_n(y_i, \mathbf{X}_i; \hat{\boldsymbol{\theta}}_n)$, requires second-derivative accuracy not directly targeted by first-order Sobolev training. The score-based estimator $\hat{\mathbf{J}}_n(\hat{\boldsymbol{\theta}}_n)$ is therefore preferred.

5.3 Inference Under Misspecification

When the emulator does not perfectly approximate the true likelihood, we can still obtain valid inference by treating the emulator as a working model, following the quasi-maximum likelihood framework (White, 1982). In this section, we treat the neural network as fixed with n, dropping the n subscript from the observation log likelihood, $\hat{m}(y, \mathbf{X}; \boldsymbol{\theta})$, and score function, $\hat{\psi}(y, \mathbf{X}; \boldsymbol{\theta})$. Define the pseudo-true parameter as the population maximizer of the emulator log likelihood:

$$\boldsymbol{\theta}_0^\dagger = \arg\max_{\boldsymbol{\theta} \in \Theta} \mathbb{E}\left\{\hat{m}(y, \mathbf{X}; \boldsymbol{\theta})\right\}.$$

When the emulator closely approximates the true likelihood, $\boldsymbol{\theta}_0^\dagger \approx \boldsymbol{\theta}_0$. The asymptotic distribution depends on two matrices:

$$\mathbf{A}(\boldsymbol{\theta}) = -\mathbb{E}\left\{\nabla^2_{\boldsymbol{\theta}} \hat{m}(y, \mathbf{X}; \boldsymbol{\theta})\right\}, \qquad \mathbf{B}(\boldsymbol{\theta}) = \mathbb{E}\left\{\hat{\psi}(y, \mathbf{X}; \boldsymbol{\theta}) \hat{\psi}(y, \mathbf{X}; \boldsymbol{\theta})^\top\right\}.$$

Under correct specification (i.e., a perfect emulator), the information matrix equality gives $\mathbf{A}(\boldsymbol{\theta}_0) = \mathbf{B}(\boldsymbol{\theta}_0) = \mathbf{J}(\boldsymbol{\theta}_0)$. Under misspecification, these matrices differ, necessitating the sandwich covariance form. Under regularity conditions stated in Appendix B.3, we obtain the following results.

Theorem 5 (Consistency Under Misspecification)**.** *Under Assumptions 5 and 6,* $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0^\dagger$.

Theorem 6 (Asymptotic Normality Under Misspecification)**.** *Under Assumptions 5 and 6,*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0^\dagger) \xrightarrow{d} \mathcal{N}\left\{\mathbf{0}, \mathbf{A}(\boldsymbol{\theta}_0^\dagger)^{-1}\mathbf{B}(\boldsymbol{\theta}_0^\dagger)\mathbf{A}(\boldsymbol{\theta}_0^\dagger)^{-\top}\right\}.$$

The proofs follow standard M-estimator arguments; see Appendix B.4. The sandwich covariance $\mathbf{V} = \mathbf{A}(\boldsymbol{\theta}_0^\dagger)^{-1}\mathbf{B}(\boldsymbol{\theta}_0^\dagger)\mathbf{A}(\boldsymbol{\theta}_0^\dagger)^{-\top}$ can be consistently estimated by $\hat{\mathbf{V}}_n = \hat{\mathbf{A}}_n^{-1}\hat{\mathbf{B}}_n\hat{\mathbf{A}}_n^{-\top}$, where

$$\hat{\mathbf{A}}_n = -\frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \hat{m}(y_i, \mathbf{X}_i; \hat{\boldsymbol{\theta}}_n), \qquad \hat{\mathbf{B}}_n = \frac{1}{n}\sum_{i=1}^{n} \hat{\psi}(y_i, \mathbf{X}_i; \hat{\boldsymbol{\theta}}_n)\hat{\psi}(y_i, \mathbf{X}_i; \hat{\boldsymbol{\theta}}_n)^\top.$$

Both quantities are computable via automatic differentiation. Note that $\hat{\mathbf{B}}_n$ corresponds to $\hat{\mathbf{J}}_n(\hat{\boldsymbol{\theta}}_n)$ from the previous section under an asymptotic regime with a fixed emulator.

Theorem 6 provides valid inference for $\boldsymbol{\theta}_0^\dagger$, not necessarily $\boldsymbol{\theta}_0$. If we employ an asymptotic regime where the emulator varies with n, then the pseudo-true parameter also varies with n, so we denote it as $\boldsymbol{\theta}_n^\dagger$. When $\alpha \in [0, 1)$ in Assumption 2, we have $\boldsymbol{\theta}_n^\dagger \xrightarrow{p} \boldsymbol{\theta}_0$ as $n \to \infty$, but this convergence may be too slow relative to $\sqrt{n}$ sampling variability. The sandwich standard errors remain valid for $\boldsymbol{\theta}_n^\dagger$, but confidence intervals should not be interpreted as targeting $\boldsymbol{\theta}_0$. Even if $\mathbf{A}(\boldsymbol{\theta}_0^\dagger)$ and $\mathbf{B}(\boldsymbol{\theta}_0^\dagger)$ both approximate $\mathbf{J}(\boldsymbol{\theta}_0)$, the centering at $\boldsymbol{\theta}_n^\dagger \neq \boldsymbol{\theta}_0$ introduces asymptotic bias. Valid inference for $\boldsymbol{\theta}_0$ requires $\alpha \geq 1$ (Theorem 4).

Similarly, Proposition 1 provides conditions under which $\hat{\mathbf{B}}_n \xrightarrow{p} \mathbf{J}(\boldsymbol{\theta}_0)$. However, our first-order Sobolev training procedure provides no guarantee that the emulator Hessian will approximate the Hessian of the true model; thus, $\hat{\mathbf{A}}_n$ need not converge to any quantity from the true model. However, Theorem 6 requires only that $\mathbf{A}(\boldsymbol{\theta}_0^\dagger)$ be nonsingular, not that it match the true model. When the emulator additionally provides accurate second derivatives, the sandwich covariance matrix reduces to $\mathbf{J}(\boldsymbol{\theta}_0)^{-1}$.

6 Simulation Study

We evaluate the performance of emulator-based MNP estimators across a range of sample sizes, numbers of alternatives, and model specifications. We compare their statistical performance and computational requirements to estimators based on the GHK simulator.

6.1 Simulation Design

Model specification. We consider MNP models with $K \in \{3, 5, 10\}$ alternatives and $p = 2$ covariates. Alternative K serves as the reference with utility normalized to zero, representing an outside option. For each non-reference alternative $j \in \{1, \ldots, K-1\}$, the deterministic utility is $v_j = \boldsymbol{x}_j^\top \boldsymbol{\beta}$, where $\boldsymbol{x}_j \in \mathbb{R}^p$ is a vector of alternative-specific attributes and $\boldsymbol{\beta} = (0, 1)^\top$ is the true coefficient vector. We consider two specifications for $\boldsymbol{\Sigma}$, denoting them as Dense and Factor.

The Dense specification assumes a dense structure for $\boldsymbol{\Sigma}_{1:(K-1),1:(K-1)}$, the upper $(K-1) \times (K-1)$ submatrix of $\boldsymbol{\Sigma}$. For identification, we set $\Sigma_{11} = 1$ and $\boldsymbol{\Sigma}_{\cdot K} = \boldsymbol{\Sigma}_{K\cdot}^\top = \mathbf{0}_K$; the latter equation sets the variance and covariances of the zero-utility option to zero.

The Factor specification assumes a one-factor structure for $\boldsymbol{\Sigma}_{1:(K-1),1:(K-1)}$ of the form $\boldsymbol{\Sigma}_{1:(K-1),1:(K-1)} = \boldsymbol{\Psi} + \boldsymbol{\gamma}\boldsymbol{\gamma}^\top$, where $\boldsymbol{\gamma} \in \mathbb{R}^{K-1}$ and $\boldsymbol{\Psi} \in \mathbb{R}^{(K-1)\times(K-1)}$ is a diagonal matrix of uniquenesses. For identification, we impose an anchored normalization that fixes the first uniqueness and first factor loading:

$$\Psi_{11} = 0, \qquad \gamma_1 = 1.$$

As with the Dense specification, we set $\boldsymbol{\Sigma}_{\cdot K} = \boldsymbol{\Sigma}_{K\cdot}^\top = \mathbf{0}_K$. These constraints imply $\boldsymbol{\Sigma}_{11} = 1$, fixing the scale of the model, and $\boldsymbol{\Sigma}_{1j} = \gamma_j$ for $j \geq 2$, so the factor loadings directly encode covariances with the first alternative. Given these constraints, we parameterize the model in terms of $K-2$ free uniquenesses $(\Psi_{22}, \ldots, \Psi_{K-1,K-1})$ and $K-2$ free loadings $(\gamma_2, \ldots, \gamma_{K-1})$. This anchored normalization is designed to eliminate flat directions in the likelihood surface.

Table 1 shows the number of nonredundant parameters by type for the Dense and Factor specifications. With the anchored normalization, the Factor specification is exactly identified for $K = 3$ and overidentified for $K \geq 4$. For $K = 10$, the Factor specification results in fewer than half as many parameters as the Dense specification (18 vs. 46).

Data generation. For each replication, we generate covariates $\boldsymbol{x}_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ for observations $i = 1, \ldots, n$ and alternatives $j = 1, \ldots, K-1$, with $\boldsymbol{x}_{iK} = \mathbf{0}$ for the reference

Covariance ($\boldsymbol{\Sigma}$) Specification	K	Coefficients	Variances (Uniquenesses)	Covariances (Loadings)	Total
Dense	3	2	1	1	4
	5	2	3	6	11
	10	2	8	36	46
Factor	3	2	1	1	4
	5	2	3	3	8
	10	2	8	8	18

Table 1: Parameter counts by $\boldsymbol{\Sigma}$ specification (Dense or Factor) and K.

alternative.

For the Dense specification, we draw the covariance matrix according to a Wishart($\mathbf{I}_{K-1}, K+10$) distribution and then normalize it so that $\Sigma_{11} = 1$, resulting in a PDF given by

$$\frac{\Gamma(s) \cdot |\boldsymbol{\Sigma}_{1:(K-1),1:(K-1)}|^5}{2^s \Gamma_{K-1}\left(\frac{K+10}{2}\right) \cdot \left\{\frac{1}{2}\text{tr}(\boldsymbol{\Sigma}_{1:(K-1),1:(K-1)})\right\}^s},$$

where $s = (K+10)(K-1)/2$ and $\Gamma_{K-1}(\cdot)$ denotes the $(K-1)$-dimensional multivariate gamma function (see Appendix C.3).

For the Factor specification, we sample the free uniquenesses on the log scale, drawing $\log(\Psi_{jj}) \overset{iid}{\sim} \mathcal{N}(0, 0.1^2)$ for $j = 2, \ldots, K-1$, and exponentiate to obtain $\Psi_{jj} > 0$. For the factor loadings, we draw $\gamma_j \overset{iid}{\sim} \mathcal{N}(0, 0.2^2)$ for $j = 2, \ldots, K-1$. The fixed values $\Psi_{11} = 0$ and $\gamma_1 = 1$ complete the parameterization.

Given the covariates and true parameters, we compute deterministic utilities and simulate choices $y_i \in \{1, \ldots, K\}$ according to one of two data-generating processes (DGPs): Probit or Emulator. The Probit RUM generates choices from the exact MNP model, drawing errors $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and selecting the utility-maximizing alternative. The Emulator RUM generates choices by sampling directly from the categorical distribution defined by the emulator's predicted choice probabilities. When coupled with the Emulator RUM, the emulator-based estimation method is correctly specified, enabling a comparison of emulator-based estimators under both correct specification (Emulator RUM) and approximate inference (Probit RUM).

Estimation. To ensure positive definiteness of $\boldsymbol{\Sigma}$, we estimate the Dense specification on the scale of the Cholesky factor, with the diagonal elements optimized on the log scale to enforce positivity. Similarly, we estimate the Factor specification with the free uniquenesses

on the log scale, optimizing $\log(\Psi_{jj})$ for $j = 2, \ldots, K-1$.

We form estimators using two methods, Emulator and GHK, corresponding to whether we approximate the log-likelihood function using a pretrained emulator or the GHK simulator. We fit both methods to data generated by the Probit RUM; we also fit the Emulator method to data generated by the Emulator RUM to analyze its performance under correct specification. We apply the GHK method with $R \in \{10, 50, 250\}$ draws, representing low, moderate, and high computational effort.

Both methods, Emulator and GHK, perform regularized maximum likelihood estimation. The regularization consists of two components: a barrier penalty and a pseudo-prior. The barrier penalty is applied to $\log(\Psi_{jj})$ for the Factor specification and to the log-diagonal Cholesky parameters for the Dense specification. For a latent parameter η, the barrier penalty is $\lambda_{\text{reg}} \cdot \log(\eta + 8)/n$. The pseudo-prior terms take the form $\lambda_{\text{reg}} \cdot \log\{p(\boldsymbol{\theta})\}/n$, where $p(\boldsymbol{\theta})$ is the density function of $\boldsymbol{\theta}$ under the data-generating process. We add these penalties together to form a combined penalty of the form $\lambda_{\text{reg}} \cdot r(\boldsymbol{\theta})/n$, where $r(\boldsymbol{\theta})$ sums the barrier and log-prior penalties and λ_{reg} is a regularization hyperparameter.

We then estimate $\boldsymbol{\theta}$ as the maximizer of the sum of the (approximate) log likelihood and $\lambda_{\text{reg}} \cdot r(\boldsymbol{\theta})/n$. We take $\lambda_{\text{reg}} = 0.1$ so that the penalty is small and asymptotically negligible. In practice, we find that these penalties lead to more reliable convergence of the model-fitting algorithm and help address weak identification of the covariance parameters, especially when n is small relative to K. We account for these penalties in the standard error computation by estimating $\text{Cov}(\hat{\boldsymbol{\theta}}_n)$ as $\left\{n \cdot \hat{\mathbf{J}}_n(\hat{\boldsymbol{\theta}}_n) - \lambda_{\text{reg}} \cdot \nabla^2_{\boldsymbol{\theta}}\, r(\hat{\boldsymbol{\theta}}_n)\right\}^{-1}$. Again, the impact of the penalization is asymptotically negligible because the term $n \cdot \hat{\mathbf{J}}_n(\hat{\boldsymbol{\theta}}_n)$ dominates asymptotically.

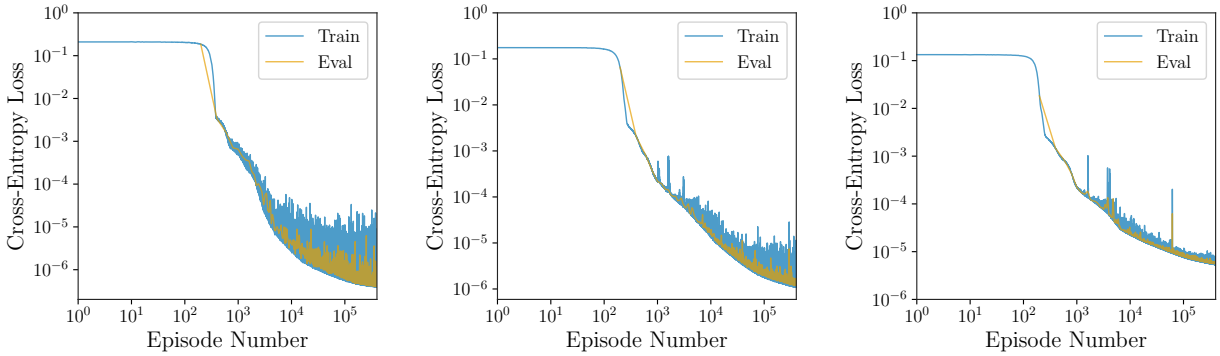
For all methods, we maximize the (approximate) regularized log likelihood using L-BFGS optimization (Liu and Nocedal, 1989). In practice, we found it necessary to limit the minimum step size (we used 10^{-6}) and allow small decreases in the objective function to enable the optimizer to escape local minima.

Experimental design. We consider sample sizes $n \in \{1{,}000, 10{,}000, 100{,}000\}$ and conduct 100 replications for each configuration. We compute performance metrics aggregated across all parameter types. We compute Monte Carlo standard errors for the coverage and average timings using scaled empirical standard deviations. For the remaining metrics, we apply bootstrap resampling (across repetitions) with 1,000 bootstrap samples to estimate the Monte Carlo standard errors.

Computational environment. All simulations were implemented in PyTorch (Paszke et al., 2019) and executed on a personal laptop with 48 GB RAM and an Apple M4 Pro processor. We fit the emulators using the Apple Metal Performance Shaders (MPS). To obtain comparable estimation timings, however, we performed the rest of the simulation on CPU.

6.2 Simulation Results

We fit the emulators for 400,000 training episodes using 400,000 pregenerated training examples, each consisting of 10^6 choices. Details on the optimizer, learning rate schedule, and network architecture are provided in Appendix C.2. This model-fitting procedure completed in 2.8, 4.8, and 14.0 hours for $K = 3, 5$, and 10, respectively, representing a modest level of computational effort.



(a) $K = 3$ **(b)** $K = 5$ **(c)** $K = 10$

Figure 2: Training and evaluation loss over 400,000 training episodes for emulators with $K \in \{3, 5, 10\}$ alternatives. The training loss (blue) decreases consistently throughout training. The evaluation loss (orange) also decreases, but it shows occasional periods of overfitting for $K = 5$. We address this overfitting by saving the emulator model weights that achieve the best evaluation loss.

Figure 2 displays the training and evaluation loss for the emulators over time. At each value of $K \in \{3, 5, 10\}$, the training loss decreases consistently over time even at the end of training, indicating that further training would likely lead to additional improvements. The evaluation loss closely tracks the training loss, implying that the emulator generalizes well with minimal overfitting to the simulated training probabilities. We compute the evaluation loss every 200 episodes; these periodic calculations produce significantly less visible variation in the evaluation loss compared to the training loss.

Table 2 compares the performance of the emulator and GHK estimation methods for $K = 3$. The performance metrics are root mean squared error (RMSE), root mean squared bias

n	DGP	Method	RMSE	RMS Bias	SE Ratio	Coverage	Time (s)
	Emulator	Emulator	0.157 (0.016)	0.012 (0.009)	1.01 (0.05)	0.94 (0.01)	0.11 (0.01)
	Probit	Emulator	0.208 (0.033)	0.050 (0.019)	0.95 (0.08)	0.95 (0.01)	0.11 (0.00)
1,000	Probit	GHK(10)	0.202 (0.031)	0.047 (0.019)	0.96 (0.08)	0.94 (0.02)	0.08 (0.00)
	Probit	GHK(50)	0.203 (0.032)	0.048 (0.018)	0.96 (0.09)	0.95 (0.01)	0.14 (0.00)
	Probit	GHK(250)	0.207 (0.033)	0.049 (0.019)	0.95 (0.08)	0.95 (0.01)	0.40 (0.01)
	Emulator	Emulator	0.052 (0.006)	0.007 (0.004)	0.93 (0.07)	0.93 (0.01)	0.36 (0.00)
	Probit	Emulator	0.054 (0.009)	0.008 (0.005)	0.98 (0.11)	0.94 (0.01)	0.36 (0.00)
10,000	Probit	GHK(10)	0.052 (0.008)	0.010 (0.003)	1.00 (0.09)	0.94 (0.01)	0.29 (0.00)
	Probit	GHK(50)	0.052 (0.009)	0.007 (0.004)	1.00 (0.10)	0.95 (0.01)	0.42 (0.00)
	Probit	GHK(250)	0.053 (0.009)	0.008 (0.005)	0.99 (0.11)	0.94 (0.01)	0.84 (0.01)
	Emulator	Emulator	0.016 (0.002)	0.001 (0.001)	0.97 (0.08)	0.94 (0.01)	1.69 (0.01)
	Probit	Emulator	0.016 (0.002)	0.001 (0.001)	1.00 (0.07)	0.97 (0.01)	1.73 (0.01)
100,000	Probit	GHK(10)	0.020 (0.002)	0.007 (0.001)	0.83 (0.06)	0.83 (0.02)	0.87 (0.01)
	Probit	GHK(50)	0.016 (0.002)	0.002 (0.001)	1.03 (0.08)	0.97 (0.01)	1.77 (0.02)
	Probit	GHK(250)	0.016 (0.002)	0.001 (0.001)	1.06 (0.07)	0.98 (0.01)	6.76 (0.09)

Table 2: Comparison of Emulator and GHK estimation methods for multinomial probit models with $K = 3$. The emulator performs competitively with GHK(50) and GHK(250) and requires computation time between that of GHK(10) and GHK(50). Performance metrics are averaged across all parameters (coefficients, variances, and covariances) with Monte Carlo standard errors shown in parentheses.

(RMS Bias), the ratio of estimated to empirical standard errors (SE Ratio), 95% confidence interval coverage rates (Coverage), and elapsed model-fitting time in seconds (Time (s)). The emulator performs well across all sample sizes, nearly matching GHK(50) and GHK(250) in statistical performance but at a lower computational cost in most comparisons. GHK(10) significantly undercovers at $n = 100{,}000$ and exhibits substantially higher RMS Bias than the other methods. This poor performance is likely an artifact of the GHK simulator providing biased estimates on the log scale due to Jensen's inequality; for sufficiently large n, this bias is eventually nonnegligible relative to $1/\sqrt{n}$ sampling error.

Table 3 displays similar results for $K = 5$. The emulator is competitive with GHK(50) and GHK(250) at all sample sizes. Compared to GHK(50), the compute time for the Emulator method is slightly lower for $n = 1{,}000$, approximately equal for $n = 10{,}000$, and slightly higher for $n = 100{,}000$.

Table 4 displays performance results for $K = 10$. The emulator performs comparably to GHK(250) with significantly less computation. At $n = 1{,}000$, the emulator fits the model in less than 3 seconds on average, outpacing even GHK(10). The computation time of the emulator falls between that of GHK(10) and GHK(50) at $n = 10{,}000$ and between GHK(50) and GHK(250) at $n = 100{,}000$. Compared to Tables 2 and 3, the standard error estimates

n	DGP	Method	RMSE	RMS Bias	SE Ratio	Coverage	Time (s)
	Emulator	Emulator	0.243 (0.017)	0.032 (0.012)	1.11 (0.06)	0.94 (0.01)	0.41 (0.01)
	Probit	Emulator	0.282 (0.020)	0.045 (0.017)	1.04 (0.04)	0.93 (0.01)	0.42 (0.01)
1,000	Probit	GHK(10)	0.278 (0.015)	0.074 (0.017)	0.95 (0.03)	0.89 (0.01)	0.28 (0.00)
	Probit	GHK(50)	0.280 (0.017)	0.051 (0.017)	1.02 (0.04)	0.92 (0.01)	0.58 (0.01)
	Probit	GHK(250)	0.284 (0.020)	0.046 (0.017)	1.03 (0.04)	0.93 (0.01)	1.94 (0.02)
	Emulator	Emulator	0.086 (0.006)	0.010 (0.004)	0.99 (0.05)	0.94 (0.01)	2.24 (0.02)
	Probit	Emulator	0.084 (0.005)	0.009 (0.004)	1.05 (0.04)	0.96 (0.01)	2.27 (0.02)
10,000	Probit	GHK(10)	0.095 (0.006)	0.032 (0.006)	0.88 (0.03)	0.89 (0.01)	0.85 (0.01)
	Probit	GHK(50)	0.086 (0.006)	0.013 (0.005)	1.02 (0.04)	0.95 (0.01)	2.24 (0.03)
	Probit	GHK(250)	0.084 (0.005)	0.010 (0.004)	1.06 (0.04)	0.96 (0.01)	4.86 (0.05)
	Emulator	Emulator	0.028 (0.004)	0.003 (0.002)	0.97 (0.09)	0.95 (0.01)	17.42 (0.14)
	Probit	Emulator	0.029 (0.002)	0.001 (0.001)	0.98 (0.05)	0.95 (0.01)	17.68 (0.16)
100,000	Probit	GHK(10)	0.056 (0.004)	0.024 (0.003)	0.49 (0.03)	0.60 (0.02)	3.73 (0.03)
	Probit	GHK(50)	0.030 (0.002)	0.004 (0.001)	0.93 (0.05)	0.92 (0.01)	8.29 (0.07)
	Probit	GHK(250)	0.028 (0.002)	0.002 (0.001)	1.00 (0.06)	0.95 (0.01)	34.70 (0.43)

Table 3: Comparison of Emulator and GHK estimation methods for multinomial probit models with $K = 5$. As in Table 2, the emulator performs competitively with GHK(50) and GHK(250) at a comparable or lower computational cost.

n	DGP	Method	RMSE	RMS Bias	SE Ratio	Coverage	Time (s)
	Emulator	Emulator	0.319 (0.014)	0.192 (0.011)	1.45 (0.07)	0.94 (0.01)	2.23 (0.02)
	Probit	Emulator	0.342 (0.016)	0.208 (0.012)	1.87 (0.38)	0.94 (0.01)	2.18 (0.02)
1,000	Probit	GHK(10)	0.388 (0.017)	0.216 (0.013)	0.29 (0.06)	0.21 (0.02)	5.90 (0.38)
	Probit	GHK(50)	0.363 (0.016)	0.215 (0.013)	0.67 (0.06)	0.53 (0.03)	13.24 (0.62)
	Probit	GHK(250)	0.351 (0.017)	0.214 (0.012)	1.44 (0.12)	0.91 (0.01)	14.63 (0.39)
	Emulator	Emulator	0.171 (0.009)	0.072 (0.007)	1.52 (0.08)	0.96 (0.01)	19.18 (0.27)
	Probit	Emulator	0.198 (0.011)	0.081 (0.010)	1.68 (0.13)	0.95 (0.01)	16.79 (0.21)
10,000	Probit	GHK(10)	0.235 (0.014)	0.123 (0.010)	1.32 (0.35)	0.79 (0.02)	8.70 (0.29)
	Probit	GHK(50)	0.203 (0.010)	0.090 (0.009)	1.27 (0.07)	0.91 (0.01)	34.10 (0.85)
	Probit	GHK(250)	0.199 (0.010)	0.079 (0.010)	1.40 (0.07)	0.95 (0.01)	47.98 (0.74)
	Emulator	Emulator	0.070 (0.004)	0.013 (0.003)	1.08 (0.04)	0.95 (0.01)	143.74 (1.18)
	Probit	Emulator	0.082 (0.005)	0.019 (0.004)	0.93 (0.03)	0.93 (0.01)	143.66 (1.28)
100,000	Probit	GHK(10)	0.138 (0.009)	0.073 (0.006)	0.48 (0.02)	0.60 (0.01)	31.11 (0.34)
	Probit	GHK(50)	0.085 (0.005)	0.027 (0.004)	0.89 (0.03)	0.90 (0.01)	80.62 (0.80)
	Probit	GHK(250)	0.081 (0.005)	0.020 (0.004)	0.97 (0.03)	0.93 (0.01)	348.26 (4.36)

Table 4: Comparison of Emulator and GHK estimation methods for multinomial probit models with $K = 10$. The emulator performs comparably to GHK(250) in terms of statistical performance but requires far less computation time.

are less reliable for the smaller sample sizes; the problem is amplified if we omit the penalty on $\boldsymbol{\Sigma}$. The standard errors improve substantially with the sample size, reaching SE Ratios of 0.93–1.08 for the Emulator and GHK(250) methods at $n = 100{,}000$.

n	DGP	Method	RMSE	RMS Bias	SE Ratio	Coverage	Time (s)
1,000	Emulator	Emulator	0.089 (0.004)	0.022 (0.006)	1.17 (0.05)	0.94 (0.01)	0.11 (0.00)
	Probit	Emulator	0.095 (0.004)	0.007 (0.005)	1.10 (0.04)	0.96 (0.01)	0.10 (0.00)
	Probit	GHK(10)	0.096 (0.004)	0.011 (0.005)	1.08 (0.04)	0.96 (0.01)	0.08 (0.00)
	Probit	GHK(50)	0.094 (0.004)	0.007 (0.005)	1.11 (0.04)	0.96 (0.01)	0.15 (0.00)
	Probit	GHK(250)	0.095 (0.004)	0.007 (0.005)	1.10 (0.04)	0.96 (0.01)	0.41 (0.01)
10,000	Emulator	Emulator	0.038 (0.002)	0.003 (0.002)	1.00 (0.04)	0.93 (0.01)	0.39 (0.00)
	Probit	Emulator	0.038 (0.002)	0.007 (0.003)	1.00 (0.05)	0.96 (0.01)	0.39 (0.00)
	Probit	GHK(10)	0.040 (0.002)	0.013 (0.002)	0.99 (0.05)	0.95 (0.01)	0.32 (0.00)
	Probit	GHK(50)	0.038 (0.002)	0.008 (0.003)	1.00 (0.05)	0.96 (0.01)	0.46 (0.01)
	Probit	GHK(250)	0.038 (0.002)	0.007 (0.003)	1.00 (0.05)	0.95 (0.01)	0.91 (0.01)
100,000	Emulator	Emulator	0.011 (0.001)	0.001 (0.001)	1.07 (0.05)	0.96 (0.01)	1.86 (0.01)
	Probit	Emulator	0.011 (0.001)	0.001 (0.001)	1.04 (0.05)	0.96 (0.01)	1.85 (0.01)
	Probit	GHK(10)	0.016 (0.001)	0.009 (0.001)	0.88 (0.04)	0.81 (0.02)	0.93 (0.01)
	Probit	GHK(50)	0.012 (0.001)	0.002 (0.001)	1.04 (0.04)	0.95 (0.01)	1.96 (0.02)
	Probit	GHK(250)	0.011 (0.001)	0.001 (0.001)	1.04 (0.05)	0.96 (0.01)	7.23 (0.09)

Table 5: Comparison of Emulator and GHK estimation methods for factor-structured multinomial probit models with $K = 3$. Performance metrics are averaged over the factor parameters (coefficients, uniquenesses, and factor loadings). The Emulator's performance mirrors that of GHK(250) at a much lower computational cost.

Table 5 displays analogous results for the Factor specification with $K = 3$. Performance metrics are averaged over the nonredundant elements of $\boldsymbol{\beta}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\gamma}$. The Emulator's performance mirrors that of GHK(250) but at a much lower computational cost.

Tables 6 and 7 display analogous results for 5 and 10 alternatives, respectively. In both cases, the Emulator performs comparably to GHK(50) and GHK(250) despite not being trained specifically for factor models. The Emulator is comparable to or faster than GHK(50) for $n \in \{1{,}000, 10{,}000\}$. At $n = 100{,}000$, the Emulator's computation time falls between that of GHK(50) and GHK(250).

Across all scenarios, the Emulator method performs competitively with or significantly better than GHK. We expect that the relative performance of the emulator would increase with (a) additional training and (b) implementation on specialized deep-learning hardware, such as GPUs. The latter is likely to be especially advantageous for the emulator because its computations are trivially parallelizable and highly optimized on modern GPUs.

n	DGP	Method	RMSE	RMS Bias	SE Ratio	Coverage	Time (s)
	Emulator	Emulator	0.128 (0.004)	0.016 (0.005)	1.20 (0.04)	0.97 (0.01)	0.32 (0.01)
	Probit	Emulator	0.132 (0.003)	0.012 (0.004)	1.19 (0.03)	0.97 (0.01)	0.32 (0.00)
1,000	Probit	GHK(10)	0.134 (0.004)	0.017 (0.005)	1.12 (0.03)	0.95 (0.01)	0.23 (0.00)
	Probit	GHK(50)	0.131 (0.003)	0.012 (0.004)	1.19 (0.03)	0.98 (0.01)	0.47 (0.01)
	Probit	GHK(250)	0.132 (0.003)	0.012 (0.004)	1.19 (0.03)	0.97 (0.01)	1.57 (0.02)
	Emulator	Emulator	0.059 (0.002)	0.007 (0.003)	1.01 (0.04)	0.95 (0.01)	1.98 (0.03)
	Probit	Emulator	0.059 (0.003)	0.007 (0.003)	1.04 (0.04)	0.96 (0.01)	1.96 (0.02)
10,000	Probit	GHK(10)	0.062 (0.003)	0.011 (0.003)	0.93 (0.04)	0.93 (0.01)	0.73 (0.01)
	Probit	GHK(50)	0.059 (0.003)	0.004 (0.003)	1.02 (0.04)	0.95 (0.01)	1.89 (0.03)
	Probit	GHK(250)	0.059 (0.003)	0.007 (0.003)	1.04 (0.04)	0.95 (0.01)	4.93 (0.07)
	Emulator	Emulator	0.019 (0.001)	0.002 (0.001)	1.01 (0.03)	0.96 (0.01)	15.97 (0.18)
	Probit	Emulator	0.019 (0.001)	0.002 (0.001)	1.04 (0.03)	0.96 (0.01)	15.96 (0.18)
100,000	Probit	GHK(10)	0.029 (0.001)	0.017 (0.001)	0.75 (0.03)	0.75 (0.02)	3.53 (0.04)
	Probit	GHK(50)	0.019 (0.001)	0.004 (0.001)	1.02 (0.03)	0.95 (0.01)	7.58 (0.08)
	Probit	GHK(250)	0.019 (0.001)	0.002 (0.001)	1.04 (0.03)	0.96 (0.01)	30.74 (0.45)

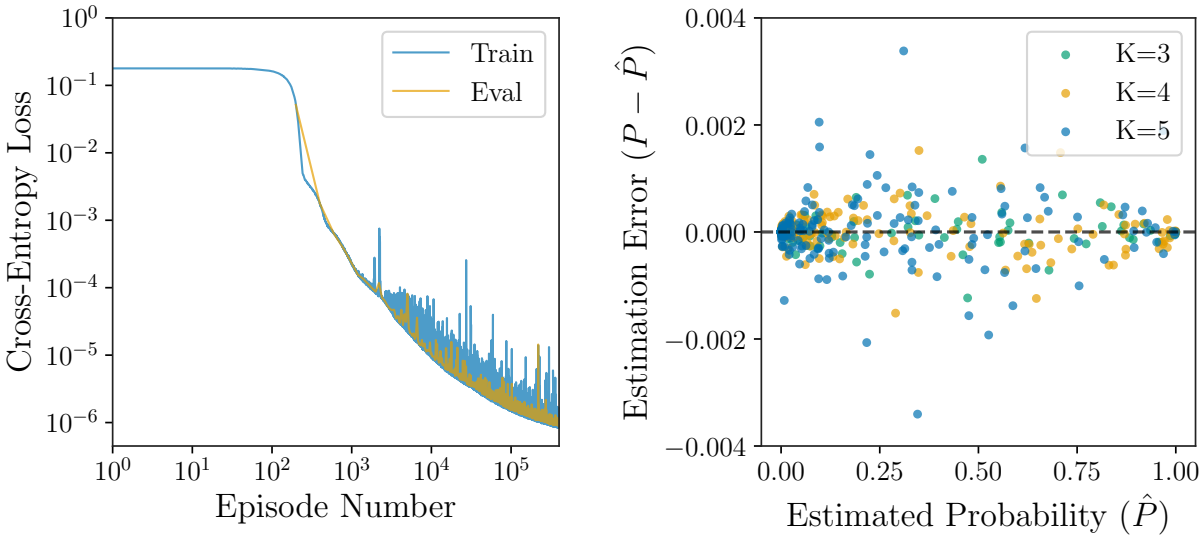
Table 6: Comparison of Emulator and GHK estimation methods for factor-structured multinomial probit models with $K = 5$. The emulator performs competitively with GHK(50) and GHK(250), often at a lower computational cost.

n	DGP	Method	RMSE	RMS Bias	SE Ratio	Coverage	Time (s)
	Emulator	Emulator	0.188 (0.004)	0.023 (0.004)	1.14 (0.02)	0.96 (0.00)	1.08 (0.01)
	Probit	Emulator	0.198 (0.005)	0.030 (0.005)	1.10 (0.02)	0.96 (0.01)	1.03 (0.01)
1,000	Probit	GHK(10)	0.220 (0.005)	0.059 (0.006)	0.92 (0.02)	0.91 (0.01)	0.76 (0.01)
	Probit	GHK(50)	0.205 (0.005)	0.039 (0.005)	1.06 (0.02)	0.95 (0.01)	2.46 (0.03)
	Probit	GHK(250)	0.199 (0.005)	0.030 (0.005)	1.11 (0.02)	0.96 (0.01)	5.80 (0.09)
	Emulator	Emulator	0.080 (0.002)	0.011 (0.003)	1.22 (0.04)	0.97 (0.01)	6.95 (0.06)
	Probit	Emulator	0.081 (0.002)	0.010 (0.002)	1.22 (0.04)	0.97 (0.01)	6.83 (0.06)
10,000	Probit	GHK(10)	0.103 (0.002)	0.049 (0.003)	0.95 (0.03)	0.88 (0.01)	3.92 (0.05)
	Probit	GHK(50)	0.087 (0.002)	0.019 (0.003)	1.14 (0.04)	0.96 (0.01)	14.67 (0.15)
	Probit	GHK(250)	0.084 (0.002)	0.012 (0.003)	1.20 (0.04)	0.97 (0.01)	23.93 (0.23)
	Emulator	Emulator	0.042 (0.002)	0.003 (0.002)	0.95 (0.04)	0.94 (0.01)	62.10 (0.51)
	Probit	Emulator	0.037 (0.001)	0.010 (0.002)	1.13 (0.05)	0.97 (0.01)	59.63 (0.55)
100,000	Probit	GHK(10)	0.071 (0.002)	0.047 (0.002)	0.65 (0.02)	0.59 (0.02)	19.50 (0.22)
	Probit	GHK(50)	0.039 (0.001)	0.012 (0.001)	1.04 (0.04)	0.93 (0.01)	48.68 (0.65)
	Probit	GHK(250)	0.036 (0.001)	0.004 (0.001)	1.11 (0.05)	0.96 (0.01)	190.72 (2.50)

Table 7: Comparison of Emulator and GHK estimation methods for factor-structured multinomial probit models with $K = 10$. The emulator performs comparably to GHK(250) in terms of statistical performance but requires far less computation time.

6.3 Multi-K Training Accuracy

We conclude this section by providing numerical results from training a neural network emulator to approximate choice probabilities for $K \in \{3, 4, 5\}$ simultaneously. We use the same neural network as we used in the previous section for $K = 5$. As above, we train for 400,000 episodes, but we loop over training examples with $K = 3$, 4, and 5 alternatives within each episode. The training process completes in 10.3 hours.



(a) Training Loss **(b)** Estimation Accuracy

Figure 3: (a) Training and evaluation loss over 400,000 training episodes for emulators with $K \in \{3, 4, 5\}$ alternatives. The training loss (blue) decreases consistently throughout training, and the evaluation loss (orange) tracks the training loss, indicating minimal overfitting. (b) Emulator probability estimates compared to estimation errors. The emulator simultaneously learns the choice probabilities for $K = 3$, 4, 5 with small error.

Figure 3a shows the training and evaluation loss steadily decreasing throughout the training procedure, approaching levels similar to those of Figure 2. Figure 3b plots the estimated choice probabilities against estimation errors for $K \in \{3, 4, 5\}$ with 400 simulated examples each. The small errors indicate that the model accurately approximates the choice probabilities across all three values of K simultaneously.

7 Conclusion

This paper proposes an amortized inference approach for discrete choice models. The approach uses a novel neural network architecture to approximate choice probabilities under general error distributions, including those featuring nontrivial cross-alternative correlations,

such as MNP models. This generality enables flexible modeling of empirical choice behavior, including complex substitution patterns, without losing the theoretical grounding or interpretability of RUM models.

Because the emulator operates on the generic choice problem, it is agnostic to the assumed parametric form of the deterministic utilities or the accompanying scale (covariance) matrix. Given a pretrained emulator, this generality enables drop-in adoption without the need for custom architecture changes or additional training. In particular, generalizing from logit to probit requires only the replacement of closed-form softmax probabilities with emulator evaluations. Other correlated error distributions are similarly straightforward.

Our proposed architecture and training process feature a preprocessing transformation, DeepSet modules (Zaheer et al., 2017), equivariant layers, smooth activation functions, and Sobolev training (Czarnecki et al., 2017). Together, these design choices produce smooth approximations of choice probabilities that respect their invariance properties and accelerate learning via parameter sharing. We show that the architecture possesses a universal approximation property, extending recent theoretical results for symmetric matrices (Blum-Smith et al., 2025) to the joint space of utility vectors and centered scale (covariance) matrices.

In Monte Carlo simulations, we find that emulator-based estimators perform as well or better than GHK-based estimators in terms of both statistical performance and computational requirements when estimating MNP models. These findings support our theoretical results showing that emulator-based estimators inherit the properties of exact MLEs under mild approximation conditions. We focus on MNP models in the simulation study to enable comparison to the GHK simulator, which is directly applicable only to Gaussian errors. The computational benefits of the amortized inference approach are amplified by the fact that emulator evaluations are trivially parallelizable and highly optimized on modern deep learning hardware, such as GPUs.

Several limitations merit discussion. Practitioners must ensure sufficient overlap between the distribution of inputs in their application and the training distribution. Additionally, our asymptotic theory requires the emulator approximation error to vanish at an appropriate rate, meaning applications with very large samples may require more extensively trained emulators. Lastly, models with correlated errors require identifying restrictions not needed for independent logit-based models. Although our approach enables simple and efficient computation of choice probabilities for these models, the data requirements and modeling decisions to ensure proper identification remain a separate challenge.

The framework also allows numerous extensions. Our emulator architecture and training procedure could be specialized to specific settings, such as factor-structured scale matrices or mixture distributions. We expect these settings could be handled by reconfiguring the modules in our architecture and, in some cases, generalizing the preprocessing transformation. Our emulators could also be used in tandem with other modern choice modeling architectures, such as those described in Section 2. Previous work has applied neural-network-assisted estimation approaches to other econometric models with intractable likelihood functions, such as dynamic discrete choice (Norets, 2012) and consumer search (Wei and Jiang, 2025), in the context of specific parametric models. The amortized inference approach presented here could likely be generalized to many of these settings, enabling efficient amortized inference without modifying or retraining the neural network.

Acknowledgments

The authors acknowledge the use of AI tools to assist with mathematical derivations, computer code, and writing. All content has been manually reviewed for accuracy.

References

Aouad, A. and Desir, A. (2025). Representing random utility choice models with nueral networks. *Management Science*, Articles in Advance.

Bagheri, N., Ghasri, M., and Barlow, M. (2025). A neural estimation framework for discrete choice models with arbitrary error distributions. *Journal of Choice Modelling*, 57:100583.

Bentz, Y. and Merunka, D. (2000). Neural networks and the multinomial logit for brand choice modelling: a hybrid approach. *Journal of Forecasting*, 19(3):177–200.

Berbeglia, G. and Venkataraman, A. (2025). The generalized stochastic preference choice model. *arXiv preprint arXiv:1803.04244*.

Bloch, H. and Marschak, J. (1960). Stochastic rationality and revealed stochastic preference. In *Contributions to probability and Statistics*. Stanford University Press.

Blum-Smith, B., Huang, N. T., Cuturi, M., and Villar, S. (2025). Functions on symmetric matrices and point clouds via lightweight invariant features from Galois theory. *SIAM Journal on Applied Algebra and Geometry*, 9(4):902–938.

Bredon, G. E. (1972). *Introduction to Compact Transformation Groups*, volume 46. Academic Press, New York.

Cai, Z., Wang, H., Talluri, K., and Li, X. (2022). Deep learning for choice modeling. *arXiv preprint arXiv:2208.09325*.

Chen, N., Gallego, G., and Tang, Z. (2025). The use of binary choice forests to model and estimate discrete choices. *arXiv preprint arXiv:1908.01109*.

Chen, Y.-C. and Mišić, V. V. (2022). Decision forest: A nonparametric approach to modeling irrational choice. *Management Science (INFORMS)*, 68(10):7090–7111.

Cocola, J. and Hand, P. (2020). Global convergence of Sobolev training for overparameterized neural networks. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 574–586. Springer.

Cranmer, K., Brehmer, J., and Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314.

Czarnecki, W. M., Osindero, S., Jaderberg, M., Swirszcz, G., and Pascanu, R. (2017). Sobolev training for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pages 4278–4287.

De Ryck, T., Lanthaler, S., and Mishra, S. (2021). On the approximation of functions by tanh neural networks. *Neural Networks*, 143:732–750.

Dugas, C., Bengio, Y., Bélisle, F., Nadeau, C., and Garcia, R. (2000). Incorporating second-order functional knowledge for better option pricing. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

E, W. and Wang, Q. (2018). Exponential convergence of the deep neural network approximation for analytic functions. *Science China Mathematics*, 61(10):1733–1740.

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57(6):1317–1339.

Hajivassiliou, V. A. and McFadden, D. L. (1998). The method of simulated scores for the estimation of LDV models. *Econometrica*, 66(4):863–896.

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

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

Hruschka, H. (2007). Using a heterogeneous multinomial probit model with a neural net extension to model brand choice. *Journal of Forecasting*, 26(2):113–127.

Jang, E., Gu, S., and Poole, B. (2017). Categorical reparameterization with Gumbel–softmax. In *International Conference on Learning Representations*.

Jiang, Z., Li, J., and Zhang, D. (2025). A high-dimensional choice model for online retailing. *Management Science (INFORMS)*, 71(4):3320–3339.

Keane, M. P. (1992). A note on identification in the multinomial probit model. *Journal of Business & Economic Statistics*, 10(2):193–200.

Keane, M. P. (1994). A computationally practical simulation estimator for panel data. *Econometrica*, 62(1):95–116.

Kingma, D. P. and Ba, J. L. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Lhéritier, A., Bocamazo, M., Delahaye, T., and Acuna-Agost, R. (2019). Airline itinerary choice modeling using machine learning. *Journal of Choice Modelling*, 31:198–209.

Liu, D. C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Lueckmann, J.-M., Bassetto, G., Karaletsos, T., and Macke, J. H. (2019). Likelihood-free inference with emulator networks. In Ruiz, F., Zhang, C., Liang, D., and Bui, T., editors, *Proceedings of The 1st Symposium on Advances in Approximate Bayesian Inference*, volume 96 of *Proceedings of Machine Learning Research*, pages 32–53. PMLR.

Maddison, C. J., Mnih, A., and Teh, Y. W. (2017). The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*.

McCulloch, R. and Rossi, P. E. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64(1):207–240.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, New York.

McFadden, D. and Richter, M. (1990). Stochastic rationality and revealed stochastic preference. In *Preferences, Uncertainty, and Optimality: Essays in Honor of Leo Hurwicz*. Westview Press.

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

Mityagin, B. S. (2020). The zero set of a real analytic function. *Mathematical Notes*, 107(3):529–530.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In Engle, R. F. and McFadden, D. L., editors, *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier.

Norets, A. (2012). Estimation of dynamic discrete choice models using artificial neural network approximations. *Econometric Reviews*, 31(1):84–106.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

Pfannschmidt, K., Gupta, P., Haddenhorst, B., and Hüllermeier, E. (2022). Learning context-dependent choice functions. *International Journal of Approximate Reasoning*, 140:116–155.

Ramachandran, P., Zoph, B., and Le, Q. V. (2018). Searching for activation functions. In *International Conference on Learning Representations*.

Rosenfeld, N., Oshiba, K., and Singer, Y. (2020). Predicting choice with set-dependent aggregation. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8220–8229. PMLR.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897.

Segol, N. and Lipman, Y. (2020). On universal equivariant set networks. In *International Conference on Learning Representations*.

Shen, X., Jiang, C., Sakhanenko, L., and Lu, Q. (2023). Asymptotic properties of neural network sieve estimators. *Journal of Nonparametric Statistics*, 35(4):839–868. PMID: 38169985.

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

Singh, A., Liu, Y., and Yoganarasimhan, H. (2023). Choice models and permutation invariance: Demand estimation in differentiated products markets. *arXiv preprint arXiv:2307.07090*.

Train, K. E. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge, 2nd edition.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, UK.

Wang, F. and Ross, C. L. (2018). Machine learning travel mode choices: Comparing the performance of an extreme gradient boosting model with a multinomial logit model. *Transportation Research Record*, 2672(47):35–45.

Wang, S., Mo, B., and Zhao, J. (2020). 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., and 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.

Wei, Y. M. and Jiang, Z. (2025). Estimating parameters of structural models using neural networks. *Marketing Science*, 44(1):102–128.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

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

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. (2017). Deep sets. In *Advances in Neural Information Processing Systems*, volume 30.

Supplementary Materials for "Amortized Inference for Correlated Discrete Choice Models via Equivariant Neural Networks"

Easton Huch
Postdoctoral Fellow, Johns Hopkins Carey Business School

Michael Keane
Carey Distinguished Professor, Johns Hopkins Carey Business School

Appendix A presents results and proofs regarding generic separation and universal approximation. Section A.8 proves the results given in Section 5.1. Appendix B provides proofs of the estimation theory results in Section 5.2. Appendix C provides details on the simulation study.

A Generic Separation and Universal Approximation

This appendix develops the group-theoretic foundations underlying the per-alternative encoder architecture presented in the main text. We establish that certain invariants generically separate orbits under permutation groups, yielding universal approximation results via the Stone–Weierstrass theorem.

Let $\mathbb{S}_K$ denote the space of symmetric real-valued $K \times K$ matrices. We consider an extended group G_K acting on $\mathbb{S}_K$ by permutation, scaling, and translation by matrices in the null space of double-centering. Section A.1 introduces the necessary notation and establishes basic properties of the centering matrix and the normalized double-centering function f.

Sections A.2–A.4 establish that the composition $h = g \circ f$, which first applies normalized double-centering and then extracts multisets of diagonal and off-diagonal elements, generically separates G_K-orbits. The exceptional set where separation fails is closed, G_K-invariant, and has measure zero.

Section A.5 considers the restriction to the centered, trace-normalized subspace $\mathbb{S}_K^* = \mathrm{Im}(f)$, where g generically separates S_K-orbits and the Stone–Weierstrass theorem yields universal approximation.

Section A.6 develops a variant for a distinguished index j, showing that an invariant g_j generically separates orbits under $S_{K-1}^{(j)}$, the subgroup fixing j.

Sections A.7–A.8 extend the theory to incorporate deterministic utilities, culminating in proofs of Theorem 1 and Theorem 2 from the main text. Throughout, $\mathbf{X}$ and $\mathbf{Y}$ represent scale matrices. As in Section 4.2, we sometimes refer to these matrices as covariance matrices although this interpretation is not strictly correct outside of MNP models.

A.1 Preliminaries

Throughout, let $K \geq 2$ be the number of alternatives. Let $\mathbf{I}_K$ denote the $K \times K$ identity matrix and $\mathbf{1}_K \in \mathbb{R}^K$ the vector of all ones.

Definition 2 (Centering Matrix)**.** The *centering matrix* $\mathbf{M} \in \mathbb{R}^{K \times K}$ is defined as $\mathbf{M} = \mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top$.

Proposition 2 (Properties of the Centering Matrix)**.** *The centering matrix* $\mathbf{M}$ *is symmetric* ($\mathbf{M}^\top = \mathbf{M}$)*, idempotent* ($\mathbf{M}^2 = \mathbf{M}$)*, and annihilates constant vectors* ($\mathbf{M}\mathbf{1}_K = \mathbf{0}$)*.*

Proof. Symmetry is immediate. For idempotency, $\mathbf{M}^2 = \mathbf{I}_K - \frac{2}{K}\mathbf{1}_K\mathbf{1}_K^\top + \frac{1}{K^2}\mathbf{1}_K(\mathbf{1}_K^\top\mathbf{1}_K)\mathbf{1}_K^\top = \mathbf{I}_K - \frac{2}{K}\mathbf{1}_K\mathbf{1}_K^\top + \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top = \mathbf{M}$. Finally, $\mathbf{M}\mathbf{1}_K = \mathbf{1}_K - \frac{1}{K}\mathbf{1}_K(\mathbf{1}_K^\top\mathbf{1}_K) = \mathbf{1}_K - \mathbf{1}_K = \mathbf{0}$. □

Definition 3 (Double-Centering Function)**.** Define the domain $\mathcal{U} = \{\mathbf{X} \in \mathbb{S}_K : \operatorname{tr}(\mathbf{MXM}) \neq 0\}$ and the function $f : \mathcal{U} \to \mathbb{S}_K$ by

$$f(\mathbf{X}) = \frac{K \cdot \mathbf{MXM}}{\operatorname{tr}(\mathbf{MXM})}.$$

Proposition 3 (Explicit Formula)**.** *For* $\mathbf{X} \in \mathbb{S}_K$*, let* $\bar{X}_{i\bullet} = \frac{1}{K}\sum_{k=1}^K X_{ik}$*,* $\bar{X}_{\bullet j} = \frac{1}{K}\sum_{k=1}^K X_{kj}$*, and* $\bar{X}_{\bullet\bullet} = \frac{1}{K^2}\sum_{i,j} X_{ij}$*. Then:*

(i) $[\mathbf{MXM}]_{ij} = X_{ij} - \bar{X}_{i\bullet} - \bar{X}_{\bullet j} + \bar{X}_{\bullet\bullet}$.

(ii) $\operatorname{tr}(\mathbf{MXM}) = \operatorname{tr}(\mathbf{X}) - K\bar{X}_{\bullet\bullet}$.

Proof. Part (i) follows by direct expansion. For part (ii), use idempotency and the cyclic property of trace: $\operatorname{tr}(\mathbf{MXM}) = \operatorname{tr}(\mathbf{M}^2\mathbf{X}) = \operatorname{tr}(\mathbf{MX}) = \operatorname{tr}(\mathbf{X}) - \frac{1}{K}\mathbf{1}_K^\top\mathbf{X}\mathbf{1}_K = \operatorname{tr}(\mathbf{X}) - K\bar{X}_{\bullet\bullet}$. □

Proposition 4 (Domain Properties)**.** *The set* $\mathbb{S}_K \setminus \mathcal{U} = \{\mathbf{X} \in \mathbb{S}_K : \operatorname{tr}(\mathbf{MXM}) = 0\}$ *is a hyperplane in* $\mathbb{S}_K$ *with Lebesgue measure zero. The domain* $\mathcal{U}$ *is open and* G_K*-invariant.*

Proof. By Proposition 3(ii), $\operatorname{tr}(\mathbf{MXM})$ is a nontrivial linear function, so its zero set is a measure-zero hyperplane. Openness follows from continuity. For G_K-invariance: scaling by $c \neq 0$ gives $\operatorname{tr}(\mathbf{M}(c\mathbf{X})\mathbf{M}) = c\operatorname{tr}(\mathbf{MXM}) \neq 0$; null-space translation by $\mathbf{N} = \boldsymbol{u}\mathbf{1}_K^\top + \mathbf{1}_K\boldsymbol{u}^\top$ satisfies $\mathbf{MNM} = \mathbf{0}$ since $\mathbf{M}\mathbf{1}_K = \mathbf{0}$; permutation preserves the trace by the cyclic property. □

Definition 4 (Extended Group Action)**.** Let S_K denote the symmetric group on K elements with associated permutation matrices $\mathbf{P}_\pi$, and let $\mathbb{R}^* = \mathbb{R} \setminus \{0\}$. Define the extended group $G_K = \mathbb{R}^K \rtimes (\mathbb{R}^* \times S_K)$, which acts on $\mathbb{S}_K$ by:

$$(c, \boldsymbol{u}, \pi) \cdot \mathbf{X} = c\,\mathbf{P}_\pi\mathbf{X}\mathbf{P}_\pi^\top + \boldsymbol{u}\mathbf{1}_K^\top + \mathbf{1}_K\boldsymbol{u}^\top.$$

Proposition 5 (f and G_K)**.** *The function* f *satisfies:*

(i) Equivariance under permutation: $f(\mathbf{P}_\pi\mathbf{X}\mathbf{P}_\pi^\top) = \mathbf{P}_\pi f(\mathbf{X})\mathbf{P}_\pi^\top$.

(ii) Invariance under scaling: $f(c\mathbf{X}) = f(\mathbf{X})$ *for* $c \neq 0$.

(iii) Invariance under null-space translation: $f(\mathbf{X} + \boldsymbol{u}\mathbf{1}_K^\top + \mathbf{1}_K\boldsymbol{u}^\top) = f(\mathbf{X})$.

Proof. For (i), since $\mathbf{P}_\pi\mathbf{1}_K = \mathbf{1}_K$, we have $\mathbf{P}_\pi\mathbf{M} = \mathbf{M}\mathbf{P}_\pi$, so $\mathbf{M}(\mathbf{P}_\pi\mathbf{X}\mathbf{P}_\pi^\top)\mathbf{M} = \mathbf{P}_\pi\mathbf{MXM}\mathbf{P}_\pi^\top$, and the trace is preserved. Part (ii) follows from $f(c\mathbf{X}) = \frac{cK\,\mathbf{MXM}}{c\cdot\operatorname{tr}(\mathbf{MXM})} = f(\mathbf{X})$. For (iii), $\mathbf{M}(\boldsymbol{u}\mathbf{1}_K^\top + \mathbf{1}_K\boldsymbol{u}^\top)\mathbf{M} = \mathbf{0}$ since $\mathbf{M}\mathbf{1}_K = \mathbf{0}$. □

Lemma 1 (Preimage Characterization)**.** *If* $f(\mathbf{X}) = f(\mathbf{Y})$ *for* $\mathbf{X}, \mathbf{Y} \in \mathcal{U}$*, then there exist* $c \in \mathbb{R}^*$ *and* $\boldsymbol{w} \in \mathbb{R}^K$ *such that* $\mathbf{X} = c\mathbf{Y} + \boldsymbol{w}\mathbf{1}_K^\top + \mathbf{1}_K\boldsymbol{w}^\top$.

Proof. Let $t_X = \operatorname{tr}(\mathbf{MXM})$ and $t_Y = \operatorname{tr}(\mathbf{MYM})$. The assumption $f(\mathbf{X}) = f(\mathbf{Y})$ implies $\mathbf{MXM} = c\,\mathbf{MYM}$ where $c = t_X/t_Y \in \mathbb{R}^*$.

For symmetric $\mathbf{Z}$, let $\boldsymbol{\mu}_\mathbf{Z} = \mathbf{Z}\mathbf{1}_K/K$ (row averages) and $m_\mathbf{Z} = \mathbf{1}_K^\top\mathbf{Z}\mathbf{1}_K/K^2$ (grand mean). Expanding $\mathbf{MZM} = \mathbf{Z} - \boldsymbol{\mu}_\mathbf{Z}\mathbf{1}_K^\top - \mathbf{1}_K\boldsymbol{\mu}_\mathbf{Z}^\top + m_\mathbf{Z}\mathbf{1}_K\mathbf{1}_K^\top$ and applying to both sides:

$$\mathbf{X} = c\mathbf{Y} + (\boldsymbol{\mu}_\mathbf{X} - c\boldsymbol{\mu}_\mathbf{Y})\mathbf{1}_K^\top + \mathbf{1}_K(\boldsymbol{\mu}_\mathbf{X} - c\boldsymbol{\mu}_\mathbf{Y})^\top - (m_\mathbf{X} - cm_\mathbf{Y})\mathbf{1}_K\mathbf{1}_K^\top.$$

Setting $\boldsymbol{w} = \boldsymbol{\mu}_{\mathbf{X}} - c\boldsymbol{\mu}_{\mathbf{Y}} - \frac{m_{\mathbf{X}} - cm_{\mathbf{Y}}}{2}\mathbf{1}_K$ gives the result. □

Definition 5 (Centered, Trace-Normalized Subspace)**.** Define $\mathbb{S}_K^* = \{\mathbf{Y} \in \mathbb{S}_K : \mathbf{Y}\mathbf{1}_K = \mathbf{0}, \operatorname{tr}(\mathbf{Y}) = K\}$.

Definition 6 (Invariant Function g)**.** For $\mathbf{Y} \in \mathbb{S}_K^*$, define $g(\mathbf{Y}) = (\mathcal{D}_K(\mathbf{Y}), \mathcal{O}_K(\mathbf{Y}))$ where $\mathcal{D}_K(\mathbf{Y}) = \{\{Y_{ii} : i = 1, \ldots, K\}\}$ is the multiset of diagonal entries and $\mathcal{O}_K(\mathbf{Y}) = \{\{Y_{ij} : 1 \leq i < j \leq K\}\}$ is the multiset of off-diagonal entries.

A.2 Generic Injectivity

We now define the genericity conditions under which g separates orbits.

Definition 7 (Genericity Conditions)**.** Let $\mathbf{Y} \in \mathbb{S}_K$. We say $\mathbf{Y}$ satisfies:

(a) *Condition (a)* if all entries are distinct: $Y_{ij} \neq Y_{kl}$ for $(i, j) \neq (k, l)$ with $i \leq j$, $k \leq l$.

(b) *Condition (b)* if for each i, there exists exactly one $(K-1)$-subset $\mathcal{S}$ of off-diagonal positions such that $\sum_{(k,l)\in\mathcal{S}} Y_{kl} = -Y_{ii}$.

Condition (b) exploits the row-sum constraint $\mathbf{Y}\mathbf{1}_K = \mathbf{0}$: the off-diagonal entries in row i sum to $-Y_{ii}$, so the canonical subset $\mathcal{S}_i^* = \{(i, j) : j \neq i\}$ always satisfies the equation. Condition (b) requires this subset to be unique among the $\binom{K(K-1)/2}{K-1}$ possible $(K-1)$-subsets.

Lemma 2 (Recovery Lemma)**.** *If $\mathbf{Y} \in \mathbb{S}_K^*$ satisfies conditions (a) and (b), then $g(\mathbf{Y})$ determines $\mathbf{Y}$ up to S_K-orbit.*

Proof. Given multisets $\mathcal{D}_K$ and $\mathcal{O}_K$, we reconstruct $\mathbf{Y}$ up to relabeling:

Step 1: Order the distinct diagonal elements as $Y_{(1)(1)} > \cdots > Y_{(K)(K)}$, assigning canonical labels to alternatives.

Step 2: For each (i), identify $\mathcal{R}_{(i)} \subset \mathcal{O}_K$ as the unique $(K-1)$-subset summing to $-Y_{(i)(i)}$ (exists by condition (b)).

Step 3: For $i \neq j$, recover $Y_{(i)(j)}$ as the unique element in $\mathcal{R}_{(i)} \cap \mathcal{R}_{(j)}$ (singleton by condition (a)).

Any $\mathbf{Y}'$ with $g(\mathbf{Y}') = g(\mathbf{Y})$ satisfying (a) and (b) yields the same reconstruction up to relabeling. □

A.3 The Exceptional Set

Definition 8 (Exceptional Set)**.** Define $\mathcal{E} = \mathcal{U}^C \cup \{\mathbf{X} \in \mathcal{U} : f(\mathbf{X})$ violates (a) or (b)$\}$.

Lemma 3 (Non-Triviality of Constraints)**.** *The following hold:*

(i) For distinct positions $(i,j) \neq (k,l)$, the function $f(\mathbf{X})_{ij} - f(\mathbf{X})_{kl}$ is not identically zero on $\mathcal{U}$.

(ii) For each index i and $(K-1)$-subset $\mathcal{S} \neq \mathcal{S}_i^*$, the function $\sum_{(k,l)\in\mathcal{S}} f(\mathbf{X})_{kl} + f(\mathbf{X})_{ii}$ is not identically zero on $\mathcal{U}$.

Proof. Consider diagonal $\mathbf{X} = \operatorname{diag}(\lambda_1, \ldots, \lambda_K)$ with distinct λ_j and $\sum_j \lambda_j \neq 0$. Then $[\mathbf{MXM}]_{ii} = \lambda_i(1 - 2/K) + \bar{\lambda}/K$ and $[\mathbf{MXM}]_{ij} = (\bar{\lambda} - \lambda_i - \lambda_j)/K$ for $i \neq j$. For generic λ_j, all entries of $f(\mathbf{X})$ are distinct, proving (i). For (ii), the linear combination for $\mathcal{S} \neq \mathcal{S}_i^*$ has nonzero coefficient on λ_i, so it is nontrivial. □

Proposition 6 (Properties of $\mathcal{E}$)**.** *The set $\mathcal{E}$ is closed, G_K-invariant, and has Lebesgue measure zero.*

Proof. **Closed:** Conditions (a) and (b) are open (strict inequalities and unique subset-sums are preserved under small perturbations), so $\mathcal{E}^C$ is open.

G_K-invariant: By Proposition 5, scaling and null-space translation leave $f(\mathbf{X})$ unchanged. Permutation conjugates $f(\mathbf{X})$, preserving distinctness and subset-sum structure.

Measure zero: Each entry $f(\mathbf{X})_{ij}$ is a rational (hence analytic) function on $\mathcal{U}$ by Proposition 3. By Lemma 3, the constraints defining violations of (a) and (b) are nontrivial analytic functions. Their zero sets have measure zero by Mityagin (2020), and $\mathcal{E}$ is a finite union of such sets. □

A.4 Basic Result for the Full Space

Theorem 7 (Main Theorem)**.** *For any $\mathbf{X}, \mathbf{Y} \in \mathbb{S}_K \setminus \mathcal{E}$: $g\{f(\mathbf{X})\} = g\{f(\mathbf{Y})\}$ implies $\mathbf{X}$ and $\mathbf{Y}$ are in the same G_K-orbit.*

Proof. By Lemma 2, $g\{f(\mathbf{X})\} = g\{f(\mathbf{Y})\}$ implies $f(\mathbf{Y}) = \mathbf{P}_\pi f(\mathbf{X})\mathbf{P}_\pi^\top$ for some $\pi \in S_K$. By Proposition 5(i), $f(\mathbf{P}_\pi \mathbf{X} \mathbf{P}_\pi^\top) = f(\mathbf{Y})$. By Lemma 1, $\mathbf{Y} = c\,\mathbf{P}_\pi \mathbf{X} \mathbf{P}_\pi^\top + \boldsymbol{w}\mathbf{1}_K^\top + \mathbf{1}_K \boldsymbol{w}^\top$ for some $c \in \mathbb{R}^*$, $\boldsymbol{w} \in \mathbb{R}^K$. □

A.5 Restriction to Centered, Trace-Normalized Matrices

Proposition 7 (Properties of $\mathbb{S}_K^*$)**.** *The subspace $\mathbb{S}_K^*$ is a nonempty, closed affine subspace of dimension $(K+1)(K-2)/2$, satisfies $\mathbb{S}_K^* = \mathrm{Im}(f)$, and is S_K-invariant.*

Proof. The dimension follows from $\dim(\mathbb{S}_K) - (K+1) = K(K+1)/2 - K - 1 = (K+1)(K-2)/2$. For $\mathbb{S}_K^* = \mathrm{Im}(f)$: if $\mathbf{X} \in \mathcal{U}$, then $f(\mathbf{X})\mathbf{1}_K = \mathbf{0}$ and $\mathrm{tr}(f(\mathbf{X})) = K$, so $\mathrm{Im}(f) \subseteq \mathbb{S}_K^*$; conversely, if $\mathbf{Y} \in \mathbb{S}_K^*$, then $\mathbf{MYM} = \mathbf{Y}$ and $f(\mathbf{Y}) = \mathbf{Y}$. For S_K-invariance: $(\mathbf{P}_\pi \mathbf{Y} \mathbf{P}_\pi^\top)\mathbf{1}_K = \mathbf{P}_\pi \mathbf{Y} \mathbf{1}_K = \mathbf{0}$ and the trace is preserved. □

Definition 9 (Restricted Exceptional Set)**.** Define $\mathcal{E}^* = \{\mathbf{Y} \in \mathbb{S}_K^* : \mathbf{Y} \text{ violates (a) or (b)}\}$.

Theorem 8 (Generic Separation on $\mathbb{S}_K^*$)**.** *The set $\mathcal{E}^*$ is closed, S_K-invariant, and has measure zero in $\mathbb{S}_K^*$. For any $\mathbf{Y}_1, \mathbf{Y}_2 \in \mathbb{S}_K^* \setminus \mathcal{E}^*$: $g(\mathbf{Y}_1) = g(\mathbf{Y}_2)$ implies $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are in the same S_K-orbit.*

Proof. The properties of $\mathcal{E}^*$ follow from Proposition 6 restricted to $\mathbb{S}_K^*$, using Lemma 3 (which provides nontrivial constraints via $\mathbb{S}_K^* = \mathrm{Im}(f)$). Separation follows from Lemma 2. □

Corollary 2 (Universal Approximation on $\mathbb{S}_K^*$)**.** *Any continuous, S_K-invariant function on $\mathbb{S}_K^* \setminus \mathcal{E}^*$ can be uniformly approximated on compact subsets by an MLP taking the sorted diagonal and off-diagonal entries as input.*

Proof. Following Blum-Smith et al. (2025), encode g via continuous functions: $g_k^d(\mathbf{Y}) =$ the kth largest diagonal entry, and $g_\ell^o(\mathbf{Y}) =$ the ℓth largest off-diagonal entry. These are S_K-invariant and separate points in $(\mathbb{S}_K^* \setminus \mathcal{E}^*)/S_K$ by Theorem 8. Since S_K is finite (hence compact), the quotient is Hausdorff (Bredon, 1972). By Stone–Weierstrass, the algebra generated by these functions is dense. By the universal approximation theorem (Cybenko, 1989; Hornik et al., 1989), MLPs can approximate any continuous function of these inputs. □

A.6 Restriction to Submatrices

For a distinguished index j, we develop an invariant separating $S_{K-1}^{(j)}$-orbits, where $S_{K-1}^{(j)} = \{\pi \in S_K : \pi(j) = j\}$.

Definition 10 (Submatrix Invariant g_j)**.** For $\mathbf{Y} \in \mathbb{S}_K^*$ and index j, define $g_j(\mathbf{Y}) = (\mathcal{P}_j(\mathbf{Y}), \mathcal{O}_j(\mathbf{Y}))$ where:

$$\mathcal{P}_j(\mathbf{Y}) = \{\{(Y_{kk}, Y_{jk}) : k \neq j\}\},$$
$$\mathcal{O}_j(\mathbf{Y}) = \{\{Y_{kl} : 1 \leq k < l \leq K,\, k \neq j,\, l \neq j\}\}.$$

The multiset $\mathcal{P}_j$ couples each diagonal entry Y_{kk} (for $k \neq j$) with the covariance Y_{jk}. The multiset $\mathcal{O}_j$ contains off-diagonal entries of $\mathbf{Y}_{-j,-j}$. Note that $Y_{jj} = K - \sum_{k \neq j} Y_{kk}$ is implicitly recoverable.

Definition 11 (Genericity Conditions for Submatrices)**.** We say $\mathbf{Y} \in \mathbb{S}_K^*$ satisfies (a_j) if all entries except possibly Y_{jj} are distinct, and (b_j) if for each $k \neq j$, there is a unique $(K-2)$-subset $\mathcal{S}$ of off-diagonal positions in $\mathbf{Y}_{-j,-j}$ with $\sum_{(k,m) \in \mathcal{S}} Y_{km} = -Y_{kk} - Y_{jk}$.

The target sum $-Y_{kk} - Y_{jk}$ arises from the row-sum constraint: $Y_{kk} + Y_{kj} + \sum_{l \neq k, l \neq j} Y_{kl} = 0$.

Theorem 9 (Generic Separation via g_j)**.** *Define $\mathcal{E}_j = \{\mathbf{Y} \in \mathbb{S}_K^* : \mathbf{Y}$ violates (a_j) or $(b_j)\}$. Then $\mathcal{E}_j$ is closed, $S_{K-1}^{(j)}$-invariant, and has measure zero. For $\mathbf{Y}_1, \mathbf{Y}_2 \in \mathbb{S}_K^* \setminus \mathcal{E}_j$: $g_j(\mathbf{Y}_1) = g_j(\mathbf{Y}_2)$ implies $\mathbf{Y}_1, \mathbf{Y}_2$ are in the same $S_{K-1}^{(j)}$-orbit.*

Proof. The proof follows Theorem 8. For separation, the recovery procedure parallels Lemma 2: order pairs in $\mathcal{P}_i$ lexicographically, identify each row's off-diagonal entries via the unique subset-sum, and recover shared entries as singleton intersections. □

Corollary 3 (Universal Approximation via g_j)**.** *Any continuous, $S_{K-1}^{(j)}$-invariant function on $\mathbb{S}_K^* \setminus \mathcal{E}_j$ can be uniformly approximated on compact subsets by an MLP taking the sorted pairs from $\mathcal{P}_j$ and sorted entries from $\mathcal{O}_j$ as input.*

Proof. The proof is analogous to Corollary 2, with pairs replacing diagonal entries. □

A.7 Extension to Utility-Covariance Pairs: Global Invariants

We extend the theory to pairs $(\boldsymbol{v}, \mathbf{Y})$ consisting of a utility vector and covariance matrix.

Definition 12 (Centered Utility-Covariance Space)**.** Define $\mathcal{X}_K = \{(\boldsymbol{v}, \mathbf{Y}) \in \mathbb{R}^K \times \mathbb{S}_K : \sum_{i=1}^K v_i = 0,\, \mathbf{Y}\mathbf{1}_K = \mathbf{0},\, \mathrm{tr}(\mathbf{Y}) = K\}$. Compare to (5) in the main text.[15]

[15] The text imposes $\boldsymbol{\Sigma}^* \succeq \mathbf{0}$ which is required for a valid covariance matrix. But the restriction $\boldsymbol{\Sigma}^* \succeq \mathbf{0}$ is not needed here as it does not provide a useful architectural simplification. It restricts the admissible input domain, but it does not create algebraic equalities that allow the network to ignore any particular part of $\boldsymbol{\Sigma}^*$. The reason is that positive semidefiniteness is an inequality restriction ($\boldsymbol{a}^\top \boldsymbol{\Sigma}^* \boldsymbol{a} \geq 0$ for all $\boldsymbol{a}$)

This space has dimension $(K-1)+(K+1)(K-2)/2 = (K^2+K-4)/2$. It is closed, Hausdorff, and S_K-invariant under $\pi \cdot (\boldsymbol{v}, \mathbf{Y}) = (\mathbf{P}_\pi \boldsymbol{v}, \mathbf{P}_\pi \mathbf{Y} \mathbf{P}_\pi^\top)$.

Definition 13 (Global Joint Invariant)**.** For $(\boldsymbol{v}, \mathbf{Y}) \in \mathcal{X}_K$, define $\tilde{g}(\boldsymbol{v}, \mathbf{Y}) = (\mathcal{P}(\boldsymbol{v}, \mathbf{Y}), \mathcal{O}(\mathbf{Y}))$ where $\mathcal{P}(\boldsymbol{v}, \mathbf{Y}) = \{\{(v_i, Y_{ii}) : i = 1, \ldots, K\}\}$ and $\mathcal{O}(\mathbf{Y}) = \{\{Y_{ij} : 1 \le i < j \le K\}\}$.

The key extension is that $\mathcal{P}$ contains *pairs* (v_i, Y_{ii}) coupling utilities with variances.

Definition 14 (Genericity Conditions)**.** We say $(\boldsymbol{v}, \mathbf{Y}) \in \mathcal{X}_K$ satisfies (ã) if the pairs (v_i, Y_{ii}) are pairwise distinct and the off-diagonal entries are pairwise distinct, and ($\tilde{\text{b}}$) if condition (b) holds for $\mathbf{Y}$.

Theorem 10 (Generic Separation on $\mathcal{X}_K$)**.** *Define* $\tilde{\mathcal{B}} = \{(\boldsymbol{v}, \mathbf{Y}) \in \mathcal{X}_K : (\tilde{a})$ *or* $(\tilde{b})$ *fails*$\}$. *Then* $\tilde{\mathcal{B}}$ *is closed,* S_K*-invariant, and has measure zero. For* $(\boldsymbol{v}_1, \mathbf{Y}_1), (\boldsymbol{v}_2, \mathbf{Y}_2) \in \mathcal{X}_K \setminus \tilde{\mathcal{B}}$*:* $\tilde{g}(\boldsymbol{v}_1, \mathbf{Y}_1) = \tilde{g}(\boldsymbol{v}_2, \mathbf{Y}_2)$ *implies the pairs are in the same* S_K*-orbit.*

Proof. The proof follows Theorem 8. The recovery procedure orders pairs lexicographically to assign labels, then recovers off-diagonal entries via subset-sum intersections. The measure-zero argument extends since $v_i - v_j$ is nontrivial on $\mathcal{X}_K$. □

A.8 Extension to Utility-Covariance Pairs: Per-Alternative Invariants

This section establishes Theorem 1 (Generic Separation) and Theorem 2 (MLP Universal Approximation) from the main text.

Definition 15 (Per-Alternative Invariant g_j)**.** For $(\boldsymbol{v}, \mathbf{Y}) \in \mathcal{X}_K$ and index j, define $g_j(\boldsymbol{v}, \mathbf{Y}) = (\mathcal{T}_j(\boldsymbol{v}, \mathbf{Y}), \mathcal{O}_j(\mathbf{Y}))$ where:

$$\mathcal{T}_j(\boldsymbol{v}, \mathbf{Y}) = \{\{(v_k, Y_{kk}, Y_{jk}) : k \neq j\}\},$$
$$\mathcal{O}_j(\mathbf{Y}) = \{\{Y_{kl} : 1 \le k < l \le K,\, k \neq j,\, l \neq j\}\}.$$

This combines extensions from Sections A.6 and A.7: $\mathcal{T}_j$ contains *triples* (v_k, Y_{kk}, Y_{jk}) coupling each alternative k's utility and variance with its covariance with j, while $\mathcal{O}_j$ contains off-diagonal entries of $\mathbf{Y}_{-j,-j}$. The values $v_j = -\sum_{k \neq j} v_k$ and $Y_{jj} = K - \sum_{k \neq j} Y_{kk}$ are implicitly recoverable.

and does not impose additional algebraic equalities among the entries of $\mathbf{\Sigma}^*$ beyond the centering and trace normalizations. Positive semidefiniteness does not imply, for example, that one row is determined by another row, or that some off-diagonal entries can be omitted from the architecture.

Definition 16 (Genericity Conditions)**.** We say $(\boldsymbol{v}, \mathbf{Y}) \in \mathcal{X}_K$ satisfies $(\tilde{\mathrm{a}}_j)$ if the triples (v_k, Y_{kk}, Y_{jk}) for $k \neq j$ are pairwise distinct and the off-diagonal entries of $\mathbf{Y}_{-j,-j}$ are pairwise distinct, and $(\tilde{\mathrm{b}}_j)$ if condition (b_j) holds.

Definition 17 (Exceptional Set)**.** Define $\mathcal{B}_j = \{(\boldsymbol{v}, \mathbf{Y}) \in \mathcal{X}_K : (\tilde{\mathrm{a}}_j) \text{ or } (\tilde{\mathrm{b}}_j) \text{ fails}\}$.

Proposition 8 (Properties of $\mathcal{B}_j$)**.** *The set $\mathcal{B}_j$ is closed, $S_{K-1}^{(j)}$-invariant, and has measure zero in $\mathcal{X}_K$.*

Proof. The proof follows Proposition 6. Closedness holds because $(\tilde{\mathrm{a}}_j)$ and $(\tilde{\mathrm{b}}_j)$ are open conditions. Invariance under $S_{K-1}^{(j)}$ holds because permutations fixing j preserve distinctness and subset-sum structure. For measure zero: coincidence of two triples defines the intersection of three hyperplanes $\{v_k = v_l\} \cap \{Y_{kk} = Y_{ll}\} \cap \{Y_{jk} = Y_{jl}\}$, each nontrivial; the subset-sum constraints are similarly nontrivial by arguments analogous to Lemma 3(ii). □

Lemma 4 (Recovery Lemma)**.** *If $(\boldsymbol{v}, \mathbf{Y}) \in \mathcal{X}_K$ satisfies $(\tilde{a}_j)$ and $(\tilde{b}_j)$, then $g_j(\boldsymbol{v}, \mathbf{Y})$ determines $(\boldsymbol{v}, \mathbf{Y})$ up to $S_{K-1}^{(j)}$-orbit.*

Proof. The proof extends Lemma 2 to triples. Order triples in $\mathcal{T}_j$ lexicographically to assign labels $(1), \ldots, (K-1)$. For each (k), identify $\mathcal{R}_{(k)} \subset \mathcal{O}_j$ as the unique $(K-2)$-subset summing to $-Y_{(k)(k)} - Y_{j(k)}$. Recover $Y_{(k)(l)}$ as the singleton $\mathcal{R}_{(k)} \cap \mathcal{R}_{(l)}$. Recover v_j and Y_{jj} from the constraints. □

We now prove the main results from Section 5.

Proof of Theorem 1. By Proposition 8, $\mathcal{B}_j$ is closed, $S_{K-1}^{(j)}$-invariant, and has measure zero. Note that these properties continue to hold when we add the restriction $\boldsymbol{\Sigma}^* \succeq \mathbf{0}$ as in (5). For $(\boldsymbol{v}_1, \mathbf{Y}_1), (\boldsymbol{v}_2, \mathbf{Y}_2) \in \mathcal{X}_K \setminus \mathcal{B}_j$ with $g_j(\boldsymbol{v}_1, \mathbf{Y}_1) = g_j(\boldsymbol{v}_2, \mathbf{Y}_2)$, Lemma 4 implies both pairs are determined up to $S_{K-1}^{(j)}$-orbit by this common value. □

Proof of Theorem 2. If $j = 1$, then recover v_1 and $Y_{11} = Y_{j1}$ from the constraints. Otherwise, there is exactly one element of $\overline{\mathcal{T}}_j(\boldsymbol{v}, \mathbf{Y})$ with $1_{\{k=1\}} = 1$. In this case, identify this element to recover v_1, Y_{11}, and Y_{j1}. Encode this information via continuous functions:

$$g_{0,1}^{\text{diag}} = v_1, \qquad g_{0,2}^{\text{diag}} = Y_{11}, \qquad g_{0,3}^{\text{diag}} = Y_{j1}.$$

Next, encode the remaining information in $\overline{\mathcal{T}}_j(\boldsymbol{v}, \mathbf{Y})$ by sorting the remaining tuples lexico-

graphically and setting

$$g_{k,1}^{\text{diag}} = v_{(k)}, \qquad g_{k,2}^{\text{diag}} = Y_{(k)(k)}, \qquad g_{k,3}^{\text{diag}} = Y_{j(k)}$$

for $k = 1, \ldots, K-1-1_{\{j\neq 1\}}$. Next, sort the values in $\mathcal{O}_j(\mathcal{Y})$ and encode them such that $g_m^{\text{off}} = Y_{k[m],l[m]}$, where $k[m]$ and $l[m]$ are the indices k and l of the m-th element in this ordering. This produces $(K-1)(K-2)/2$ such functions. Finally, set $g_C = C^*$. These functions are continuous and $S_{K-2}^{(1,j)}$-invariant, and they are easily seen to separate points on $(\overline{\mathcal{X}}_K \setminus \overline{\mathcal{B}}_j)/S_{K-2}^{(1,j)}$ by Theorem 1.

Since $S_{K-2}^{(1,j)}$ is finite, $\overline{\mathcal{X}}_K/S_{K-2}^{(1,j)}$ is Hausdorff (Bredon, 1972). By Stone–Weierstrass, the algebra generated by these functions is dense on compact subsets. By the universal approximation theorem (Cybenko, 1989; Hornik et al., 1989), MLPs can approximate any continuous function of these inputs. Because the choice probabilities are defined according to the integral given in (2) under Assumption 1, they are continuous functions; thus, the result holds. □

B Estimation Theory

This appendix contains additional information on the estimation theory results in Sections 5.2 and 5.3. Section B.1 provides additional discussion of the approximation conditions: Assumptions 2 and 4. Section B.2 provides a proof of Theorem 4. Section B.3 contains the regularity conditions for Section 5.3. Section B.4 provides a proof of Theorem 6.

B.1 Emulator Approximation Assumptions

This section provides additional discussion of Assumptions 2 and 4, showing how they can be derived from more primitive conditions on the emulator and justifying these conditions through related theoretical results.

Primitive assumptions. We consider a sieve regime in which a sequence of emulators $\hat{P}_n$ is trained with increasing precision as the sample size n grows. This is achieved by progressively (a) increasing the number of simulated training examples $s = s_n$ and (b) increasing the neural network complexity $c = c_n$ (e.g., the number of layers or hidden units). We impose the following primitive conditions:

(P1) *Compact covariate support*: The covariate space $\mathcal{C}$ is compact.

(P2) *Smooth utility and covariance mappings*: The mappings $\boldsymbol{\theta} \mapsto C_i^*(\mathbf{X}_i, \boldsymbol{\theta})$, $\boldsymbol{\theta} \mapsto \boldsymbol{v}_i^*(\mathbf{X}_i, \boldsymbol{\theta})$,

and $\boldsymbol{\theta} \mapsto \boldsymbol{\Sigma}^*(\mathbf{X}_i, \boldsymbol{\theta})$ are continuously differentiable for all $\mathbf{X}_i \in \mathcal{C}$.

(P3) *Interior probability condition*: There exist a compact set $\overline{\mathcal{X}}_K^* \subset \overline{\mathcal{X}}_K$ and $\underline{p} > 0$ such that $\inf_{(C^*, \boldsymbol{v}^*, \boldsymbol{\Sigma}^*) \in \overline{\mathcal{X}}_K^*} \min_j P_j(C^*, \boldsymbol{v}^*, \boldsymbol{\Sigma}^*) \geq \underline{p}$ and the image of $\mathcal{C} \times \Theta$ under the mapping $(\mathbf{X}, \boldsymbol{\theta}) \mapsto (C^*(\mathbf{X}, \boldsymbol{\theta}), \boldsymbol{v}^*(\mathbf{X}, \boldsymbol{\theta}), \boldsymbol{\Sigma}^*(\mathbf{X}, \boldsymbol{\theta}))$ is contained in $\overline{\mathcal{X}}_K^*$.

(P4) *Probability-scale approximation*: For some $\alpha \geq 0$, the emulator satisfies

$$\sup_{(C^*, \boldsymbol{v}^*, \boldsymbol{\Sigma}^*) \in \overline{\mathcal{X}}_K^*} \max_{j=1,\ldots,K} |\hat{P}_{n,j}(C^*, \boldsymbol{v}^*, \boldsymbol{\Sigma}^*) - P_j(C^*, \boldsymbol{v}^*, \boldsymbol{\Sigma}^*)| = o_p(n^{-\alpha/2}).$$

(P5) *Gradient approximation*: The emulator gradients satisfy

$$\sup_{(C^*, \boldsymbol{v}^*, \boldsymbol{\Sigma}^*) \in \overline{\mathcal{X}}_K^*} \max_{j=1,\ldots,K} \left\| \nabla_{C^*, \boldsymbol{v}^*, \boldsymbol{\Sigma}^*} \log \hat{P}_{n,j} - \nabla_{C^*, \boldsymbol{v}^*, \boldsymbol{\Sigma}^*} \log P_j \right\| = o_p(1).$$

We next show how these assumptions imply Assumptions 2 and 4. We then provide justification of assumptions (P4) and (P5) using related results in approximation theory.

From probability-scale errors to log-likelihood errors. We now show that conditions (P1)–(P4) imply Assumption 2. Write the log-likelihood error as

$$\hat{\ell}_n(\boldsymbol{\theta}) - \ell_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \xi_i, \quad \text{where} \quad \xi_i = \log \hat{P}_{n,y_i}(\boldsymbol{\theta}) - \log P_{y_i}(\boldsymbol{\theta}),$$

and $y_i \in \{1, \ldots, K\}$ denotes the observed choice. Conditional on covariates and the trained emulator, the ξ_i are independent. We decompose into bias and centered components:

$$\hat{\ell}_n(\boldsymbol{\theta}) - \ell_n(\boldsymbol{\theta}) = \underbrace{\frac{1}{n} \sum_{i=1}^n \mathbb{E}(\xi_i \mid \mathbf{X}_i)}_{\text{bias}} + \underbrace{\frac{1}{n} \sum_{i=1}^n \{\xi_i - \mathbb{E}(\xi_i \mid \mathbf{X}_i)\}}_{\text{centered term}}.$$

Let $\delta_n = \sup_{(C^*, \boldsymbol{v}^*, \boldsymbol{\Sigma}^*) \in \overline{\mathcal{X}}_K^*} \max_j |\hat{P}_{n,j} - P_j| = o_p(n^{-\alpha/2})$ denote the probability-scale error from (P4). Taylor expanding $\log \hat{P}_{n,j}$ around P_j gives

$$\log \hat{P}_{n,j} - \log P_j = \frac{\hat{P}_{n,j} - P_j}{P_j} - \frac{(\hat{P}_{n,j} - P_j)^2}{2P_j^2} + O\left(\frac{\delta_n^3}{\underline{p}^3}\right).$$

The conditional expectation of ξ_i is

$$\mathbb{E}(\xi_i \mid \mathbf{X}_i) = \sum_{j=1}^{K} P_j(\log \hat{P}_{n,j} - \log P_j).$$

The first-order contribution is $\sum_j P_j \cdot (\hat{P}_{n,j} - P_j)/P_j = \sum_j (\hat{P}_{n,j} - P_j) = 0$, since both probability vectors sum to one. Therefore,

$$\mathbb{E}(\xi_i \mid \mathbf{X}_i) = -\sum_{j=1}^{K} \frac{(\hat{P}_{n,j} - P_j)^2}{2P_j} + O\left(\frac{\delta_n^3}{\underline{p}^3}\right) = O\left(\frac{\delta_n^2}{\underline{p}}\right).$$

For the centered term, the mean value theorem gives $|\log \hat{P}_{n,j} - \log P_j| \le |\hat{P}_{n,j} - P_j|/\underline{p} \le \delta_n/\underline{p}$, and hence $|\xi_i| \le \delta_n/\underline{p}$. The conditional variance satisfies $\mathrm{Var}(\xi_i \mid \mathbf{X}_i) \le \mathbb{E}(\xi_i^2 \mid \mathbf{X}_i) \le \delta_n^2/\underline{p}^2$. Since the centered terms $\xi_i - \mathbb{E}(\xi_i \mid \mathbf{X}_i)$ are conditionally independent with mean zero, the variance of their sample average is

$$\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n}\{\xi_i - \mathbb{E}(\xi_i \mid \mathbf{X}_i)\} \;\middle|\; \mathbf{X}_1, \ldots, \mathbf{X}_n\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(\xi_i \mid \mathbf{X}_i) \le \frac{\delta_n^2}{n\underline{p}^2}.$$

By Chebyshev's inequality, the centered term is $O_p\left(\frac{\delta_n}{\underline{p}\sqrt{n}}\right)$.

Combining via the triangle inequality:

$$|\hat{\ell}_n(\boldsymbol{\theta}) - \ell_n(\boldsymbol{\theta})| = O\left(\frac{\delta_n^2}{\underline{p}}\right) + O_p\left(\frac{\delta_n}{\sqrt{n}\underline{p}}\right).$$

For $\alpha \in [0, 1]$, the bias term $O(\delta_n^2/\underline{p}) = o(n^{-\alpha})$ dominates the centered term $O_p\left(\frac{\delta_n}{\underline{p}\sqrt{n}}\right) = o_p\left(n^{-(\alpha+1)/2}\right)$. Since conditions (P1) and (P2) ensure that the mapping from $(\mathbf{X}, \boldsymbol{\theta})$ to $(C^*, \boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$ is continuous on the compact set $\mathcal{C} \times \Theta$, the bound holds uniformly over $\boldsymbol{\theta} \in \Theta$:

$$\sup_{\boldsymbol{\theta}\in\Theta} |\hat{\ell}_n(\boldsymbol{\theta}) - \ell_n(\boldsymbol{\theta})| = o_p(n^{-\alpha}),$$

which is Assumption 2.

From gradient approximation to Fisher information consistency. We now show that conditions (P1), (P2), (P3), and (P5) imply Assumption 4. By the chain rule,

$$\nabla_{\boldsymbol{\theta}} \log \hat{P}_{n,j}(\boldsymbol{\theta}) = \{\nabla_{\boldsymbol{\theta}}(C^*, \boldsymbol{v}^*, \boldsymbol{\Sigma}^*)\}^\top \nabla_{(C^*, \boldsymbol{v}^*, \boldsymbol{\Sigma}^*)} \log \hat{P}_{n,j},$$

and similarly for $\nabla_{\boldsymbol{\theta}} \log P_j(\boldsymbol{\theta})$. By condition (P2), the Jacobian $\nabla_{\boldsymbol{\theta}}(C^*, \boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$ is continuous on the compact set $\mathcal{C} \times \Theta$ and hence uniformly bounded: $\sup_{(\mathbf{X},\boldsymbol{\theta})\in\mathcal{C}\times\Theta} \|\nabla_{\boldsymbol{\theta}}(C^*, \boldsymbol{v}^*, \boldsymbol{\Sigma}^*)\| \leq D$ for some constant $D < \infty$. Let

$$\varepsilon_n = \sup_{(C^*,\boldsymbol{v}^*,\boldsymbol{\Sigma}^*)\in\overline{\mathcal{X}}_K^*} \max_j \left\| \nabla_{(C^*,\boldsymbol{v}^*,\boldsymbol{\Sigma}^*)} \log \hat{P}_{n,j} - \nabla_{(C^*,\boldsymbol{v}^*,\boldsymbol{\Sigma}^*)} \log P_j \right\| = o_p(1)$$

denote the gradient approximation error from (P5). Then

$$\sup_{\boldsymbol{\theta}\in\Theta} \left\| \nabla_{\boldsymbol{\theta}} \log \hat{P}_{n,j}(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \log P_j(\boldsymbol{\theta}) \right\| \leq D\varepsilon_n = o_p(1).$$

Recall the emulator and true outer-product-of-scores estimators are

$$\hat{\mathbf{J}}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \nabla_{\boldsymbol{\theta}} \log \hat{P}_{n,y_i}(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log \hat{P}_{n,y_i}(\boldsymbol{\theta})^\top,$$
$$\mathbf{J}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \nabla_{\boldsymbol{\theta}} \log P_{y_i}(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log P_{y_i}(\boldsymbol{\theta})^\top.$$

Let $\hat{\boldsymbol{s}}_i = \nabla_{\boldsymbol{\theta}} \log \hat{P}_{n,y_i}(\boldsymbol{\theta})$ and $\boldsymbol{s}_i = \nabla_{\boldsymbol{\theta}} \log P_{y_i}(\boldsymbol{\theta})$. The identity

$$\hat{\boldsymbol{s}}_i\hat{\boldsymbol{s}}_i^\top - \boldsymbol{s}_i\boldsymbol{s}_i^\top = (\hat{\boldsymbol{s}}_i - \boldsymbol{s}_i)\hat{\boldsymbol{s}}_i^\top + \boldsymbol{s}_i(\hat{\boldsymbol{s}}_i - \boldsymbol{s}_i)^\top$$

implies

$$\|\hat{\boldsymbol{s}}_i\hat{\boldsymbol{s}}_i^\top - \boldsymbol{s}_i\boldsymbol{s}_i^\top\| \leq \|\hat{\boldsymbol{s}}_i - \boldsymbol{s}_i\| \cdot (\|\hat{\boldsymbol{s}}_i\| + \|\boldsymbol{s}_i\|).$$

The true scores $\boldsymbol{s}_i$ are uniformly bounded: (P3) ensures that all relevant inputs $(C^*, \boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$ lie in $\overline{\mathcal{X}}_K^*$ where $P_j \geq \underline{p} > 0$, guaranteeing bounded gradients $\nabla_{(C^*,\boldsymbol{v}^*,\boldsymbol{\Sigma}^*)} \log P_j$, and (P2) ensures the Jacobian $\nabla_{\boldsymbol{\theta}}(C^*, \boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$ is continuous and hence bounded on $\mathcal{C} \times \Theta$. Let $M = \sup_{\boldsymbol{\theta}\in\Theta} \max_j \|\nabla_{\boldsymbol{\theta}} \log P_j(\boldsymbol{\theta})\| < \infty$, so that $\|\boldsymbol{s}_i\| \leq M$ for all i. Since $\|\hat{\boldsymbol{s}}_i - \boldsymbol{s}_i\| \leq D\varepsilon_n = o_p(1)$, the emulator scores satisfy $\|\hat{\boldsymbol{s}}_i\| \leq \|\boldsymbol{s}_i\| + D\varepsilon_n \leq M + o_p(1)$. Therefore,

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta} \|\hat{\mathbf{J}}_n(\boldsymbol{\theta}) - \mathbf{J}_n(\boldsymbol{\theta})\| &\leq \frac{1}{n}\sum_{i=1}^n \|\hat{\boldsymbol{s}}_i - \boldsymbol{s}_i\| \cdot (\|\hat{\boldsymbol{s}}_i\| + \|\boldsymbol{s}_i\|) \\ &\leq \frac{1}{n}\sum_{i=1}^n D\varepsilon_n \cdot \{2M + o_p(1)\} = o_p(1),\end{aligned}$$

which is Assumption 4.

Justification from approximation theory. The primitive conditions (P4) and (P5) are supported by related theoretical results on neural network approximation and estimation, though a complete theory tailored to our specific setting remains an area for future work.

The total emulator error can be decomposed into approximation error and estimation error:

$$\sup_{(C^*,\boldsymbol{v}^*,\boldsymbol{\Sigma}^*)\in\overline{\mathcal{X}}_K^*} \max_{j=1,\ldots,K} |\hat{P}_{n,j}(C^*,\boldsymbol{v}^*,\boldsymbol{\Sigma}^*) - P_j(C^*,\boldsymbol{v}^*,\boldsymbol{\Sigma}^*)| \leq \underbrace{a(c_n)}_{\text{approximation}} + \underbrace{e(s_n,c_n)}_{\text{estimation}},$$

where $a(c)$ is the distance from the true choice probability function to the best approximator in the network class of complexity c, and $e(s,c)$ is the deviation of the trained network from this best-in-class approximator due to training on s simulated examples.

For the approximation error, provided the choice probabilities are real-analytic functions of $(C^*,\boldsymbol{v}^*,\boldsymbol{\Sigma}^*)$ on the interior of the parameter space, neural networks with smooth activations can approximate them with exponentially small error. De Ryck et al. (2021) establish that tanh networks with two hidden layers achieve approximation error decaying as $O\left\{N^{k/(d+1)}\exp(-C\cdot N^{1/(d+1)}\log N)\right\}$ in Sobolev norms $W^{k,\infty}$, where N is the network width, d is the input dimension, and $C>0$ depends on the analyticity parameters of the target function. Earlier work established similar exponential rates for deep ReLU networks (E and Wang, 2018), though under the stronger assumption that the target function admits an absolutely convergent power series expansion. Given these exponential convergence results, achieving approximation error ε requires network complexity scaling only polylogarithmically in $1/\varepsilon$.

For the estimation error, standard results from empirical process theory yield bounds of order $O\left\{\sqrt{c\log(sc)/s}\right\}$ in L^2 norms, where c is a complexity parameter (Schmidt-Hieber, 2020; Farrell et al., 2021; Shen et al., 2023). For our setting, however, we require an L^∞ bound. The results below assume minimally that $e(s,c)=O\left\{c^k\ \text{polylog}(s)/\sqrt{s}\right\}$ for some $k<\infty$. Then, with network complexity growing polylogarithmically in the target accuracy, the dependence on c is negligible, and the estimation error is effectively of order $O\left(1/\sqrt{s}\right)$ up to logarithmic factors.

Achieving the rate given in (P4) requires that both $a(c_n)=o(n^{-\alpha/2})$ and $e(s_n,c_n)=o_p(n^{-\alpha/2})$; note that the approximation error is nonstochastic because it concerns pure function approximation. The exponential approximation rates for analytic functions imply that the first condition is satisfied with polylogarithmic growth in c_n. The second condition requires s_n to grow faster than n^α up to logarithmic factors. In practice, since simulation from

the MNP model is computationally inexpensive, generating a training set with $s_n \gg n$ is feasible.

The cases $\alpha = 0$ and $\alpha = 1$ correspond to different inferential goals. When $\alpha = 0$, Assumption 2 requires only that the log-likelihood error vanish in probability, which suffices for consistency of the emulator-based estimator (Theorem 3). When $\alpha = 1$, the stronger requirement that the error vanish faster than n^{-1} ensures that the emulator-based estimator is asymptotically equivalent to the true MLE, inheriting its $\sqrt{n}$-consistency and asymptotic normality (Theorem 4). The latter requires a larger training set: s_n growing faster than n versus s_n growing faster than $n^0 = 1$.

For the gradient approximation in (P5), Czarnecki et al. (2017) show that neural networks are universal approximators in Sobolev norms, and the bounds of De Ryck et al. (2021) hold in $W^{k,\infty}$ norms, establishing that smooth networks can simultaneously approximate both function values and derivatives. Additionally, Cocola and Hand (2020) prove global convergence of gradient flow for Sobolev training with overparameterized two-layer networks, ensuring that the trained network achieves small Sobolev loss under appropriate conditions. These results support the plausibility of (P5), though the precise rates and conditions for our setting merit further investigation.

B.2 Proof of Theorem 4

Proof. We prove the result for approximate maximizers satisfying $\hat{\ell}_n(\hat{\boldsymbol{\theta}}_n) \geq \sup_{\boldsymbol{\theta}\in\Theta} \hat{\ell}_n(\boldsymbol{\theta}) - \eta_n$ where $\eta_n = o_p(n^{-1})$. Exact maximizers correspond to $\eta_n = 0$. We proceed in three steps: first bounding the difference in log likelihoods, then showing that the estimators are asymptotically equivalent, and finally appealing to the asymptotic normality of the exact MLE.

Step 1: Bounding the difference in log likelihoods.

Define $\delta_n = \sup_{\boldsymbol{\theta}\in\Theta} |\hat{\ell}_n(\boldsymbol{\theta}) - \ell_n(\boldsymbol{\theta})| = o_p(n^{-1})$ by Assumption 2 with $\alpha \geq 1$. By the triangle inequality,

$$|\ell_n(\hat{\boldsymbol{\theta}}_n) - \ell_n(\tilde{\boldsymbol{\theta}}_n)| \leq |\ell_n(\hat{\boldsymbol{\theta}}_n) - \hat{\ell}_n(\hat{\boldsymbol{\theta}}_n)| + |\hat{\ell}_n(\hat{\boldsymbol{\theta}}_n) - \ell_n(\tilde{\boldsymbol{\theta}}_n)|.$$

By Assumption 2, the first term is bounded by δ_n. For the second term, we establish upper and lower bounds on $\hat{\ell}_n(\hat{\boldsymbol{\theta}}_n)$. Since $\tilde{\boldsymbol{\theta}}_n$ maximizes ℓ_n:

$$\hat{\ell}_n(\hat{\boldsymbol{\theta}}_n) \leq \ell_n(\hat{\boldsymbol{\theta}}_n) + \delta_n \leq \ell_n(\tilde{\boldsymbol{\theta}}_n) + \delta_n.$$

For the lower bound, by the approximate maximizer property and the fact that $\tilde{\boldsymbol{\theta}}_n \in \Theta$:

$$\hat{\ell}_n(\hat{\boldsymbol{\theta}}_n) \geq \sup_{\boldsymbol{\theta}\in\Theta} \hat{\ell}_n(\boldsymbol{\theta}) - \eta_n \geq \hat{\ell}_n(\tilde{\boldsymbol{\theta}}_n) - \eta_n \geq \ell_n(\tilde{\boldsymbol{\theta}}_n) - \delta_n - \eta_n.$$

Combining these bounds, $|\hat{\ell}_n(\hat{\boldsymbol{\theta}}_n) - \ell_n(\tilde{\boldsymbol{\theta}}_n)| \leq \delta_n + \eta_n$. Therefore,

$$|\ell_n(\hat{\boldsymbol{\theta}}_n) - \ell_n(\tilde{\boldsymbol{\theta}}_n)| \leq 2\delta_n + \eta_n = o_p(n^{-1}),$$

which implies $n|\ell_n(\hat{\boldsymbol{\theta}}_n) - \ell_n(\tilde{\boldsymbol{\theta}}_n)| = o_p(1)$.

Step 2: Asymptotic equivalence of the estimators.

We show that $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n) \xrightarrow{p} \mathbf{0}$. Fix $\epsilon > 0$ and define the neighborhood

$$B_{n,\epsilon} = \left\{ \boldsymbol{\theta} \in \Theta : \sqrt{n}\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_n\| \leq \epsilon \right\}.$$

We show that $\hat{\boldsymbol{\theta}}_n \in B_{n,\epsilon}$ with probability approaching one.

For any $\boldsymbol{\theta} \notin B_{n,\epsilon}$, we have $\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_n\| > \epsilon/\sqrt{n}$. Since $\tilde{\boldsymbol{\theta}}_n$ maximizes ℓ_n, the score vanishes: $\frac{1}{n}\sum_{i=1}^n \psi(y_i, \mathbf{X}_i; \tilde{\boldsymbol{\theta}}_n) = \mathbf{0}$. By Taylor expansion around $\tilde{\boldsymbol{\theta}}_n$,

$$\ell_n(\boldsymbol{\theta}) = \ell_n(\tilde{\boldsymbol{\theta}}_n) + \frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_n)^\top \nabla^2_{\boldsymbol{\theta}} \ell_n(\bar{\boldsymbol{\theta}})(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_n),$$

where $\bar{\boldsymbol{\theta}}$ lies on the segment between $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}_n$. By consistency, both $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$ and $\tilde{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$, so $\bar{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$. By the uniform law of large numbers and Assumption 3 (ii) and (iv),

$$\nabla^2_{\boldsymbol{\theta}} \ell_n(\bar{\boldsymbol{\theta}}) \xrightarrow{p} -\mathbf{J}(\boldsymbol{\theta}_0).$$

Let $\lambda_{\min} > 0$ denote the minimum eigenvalue of $\mathbf{J}(\boldsymbol{\theta}_0)$, which is positive by Assumption 3(v). For sufficiently large n, with probability approaching one, the Hessian $\nabla^2_{\boldsymbol{\theta}} \ell_n(\bar{\boldsymbol{\theta}})$ has all eigenvalues bounded above by $-\lambda_{\min}/2$. Thus, for $\boldsymbol{\theta} \notin B_{n,\epsilon}$,

$$n\left\{\ell_n(\tilde{\boldsymbol{\theta}}_n) - \ell_n(\boldsymbol{\theta})\right\} \geq \frac{\lambda_{\min}}{4} n\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_n\|^2 > \frac{\lambda_{\min}}{4} \cdot \epsilon^2 = \kappa > 0$$

with probability approaching one as $n \to \infty$.

Now suppose $\Pr(\hat{\boldsymbol{\theta}}_n \notin B_{n,\epsilon}) \not\to 0$. On the event that both $\hat{\boldsymbol{\theta}}_n \notin B_{n,\epsilon}$ and the Hessian bound holds, we have $n(\ell_n(\tilde{\boldsymbol{\theta}}_n) - \ell_n(\hat{\boldsymbol{\theta}}_n)) > \kappa$. Since the Hessian bound holds with probability

approaching one, this event has probability bounded away from zero, which implies that $\Pr\left[\left\{n(\ell_n(\tilde{\boldsymbol{\theta}}_n) - \ell_n(\hat{\boldsymbol{\theta}}_n)\right\} > \kappa\right] \not\to 0$. But this contradicts $n|\ell_n(\hat{\boldsymbol{\theta}}_n) - \ell_n(\tilde{\boldsymbol{\theta}}_n)| = o_p(1)$ from Step 1. Therefore, $\Pr(\hat{\boldsymbol{\theta}}_n \in B_{n,\epsilon}) \to 1$.

Since $\epsilon > 0$ was arbitrary, we conclude that $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n) \xrightarrow{p} \mathbf{0}$.

Step 3: Asymptotic normality.

By standard maximum likelihood theory (Theorem 5.39 and Lemma 7.6 of van der Vaart (1998)), the regularity conditions in Assumption 3 ensure that the model is differentiable in quadratic mean at $\boldsymbol{\theta}_0$ with nonsingular Fisher information. Since $\tilde{\boldsymbol{\theta}}_n$ is consistent, $\sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left\{\mathbf{0}, \mathbf{J}(\boldsymbol{\theta}_0)^{-1}\right\}$. Since $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n) \xrightarrow{p} \mathbf{0}$, Slutsky's theorem gives

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n) + \sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left\{\mathbf{0}, \mathbf{J}(\boldsymbol{\theta}_0)^{-1}\right\}. \qquad \square$$

B.3 Regularity Conditions for Inference Under Misspecification

This appendix states the regularity conditions required for the results in Section 5.3. These conditions parallel standard assumptions for M-estimators (van der Vaart, 1998, Chapter 5) and are mild for neural network emulators with smooth activation functions. Recall that in Section 5.3, we treat the neural network as fixed (not varying with n), so we use $\hat{m}$ and $\hat{\psi}$ without n subscripts.

Assumption 5 (Identification and Uniform Convergence)**.** The following conditions hold:

(i) The parameter space Θ is compact.

(ii) The pseudo-true parameter $\boldsymbol{\theta}_0^\dagger = \arg\max_{\boldsymbol{\theta}\in\Theta} \mathbb{E}\left\{\hat{m}(y, \mathbf{X}; \boldsymbol{\theta})\right\}$ exists uniquely and lies in the interior of Θ.

(iii) The pseudo-true parameter is well-separated: for every $\epsilon > 0$,

$$\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0^\dagger\|\geq\epsilon} \mathbb{E}\left\{\hat{m}(y, \mathbf{X}; \boldsymbol{\theta})\right\} < \mathbb{E}\left\{\hat{m}(y, \mathbf{X}; \boldsymbol{\theta}_0^\dagger)\right\}.$$

(iv) The sample criterion function converges uniformly to its expectation:

$$\sup_{\boldsymbol{\theta}\in\Theta} \left| \frac{1}{n}\sum_{i=1}^{n} \hat{m}(y_i, \mathbf{X}_i; \boldsymbol{\theta}) - \mathbb{E}[\hat{m}(y, \mathbf{X}; \boldsymbol{\theta})] \right| \xrightarrow{p} 0.$$

(v) The estimator $\hat{\boldsymbol{\theta}}_n$ is an approximate maximizer of the sample criterion function:

$$\frac{1}{n}\sum_{i=1}^{n}\hat{m}(y_i,\mathbf{X}_i;\hat{\boldsymbol{\theta}}_n) \geq \sup_{\boldsymbol{\theta}\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\hat{m}(y_i,\mathbf{X}_i;\boldsymbol{\theta}) - o_p(n^{-1}).$$

Assumption 6 (Regularity for Asymptotic Normality)**.** The following conditions hold:

(i) (Differentiability) For each $(y,\mathbf{X})$, the map $\boldsymbol{\theta}\mapsto\hat{m}(y,\mathbf{X};\boldsymbol{\theta})$ is differentiable on Θ with gradient $\hat{\psi}(y,\mathbf{X};\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\hat{m}(y,\mathbf{X};\boldsymbol{\theta})$.

(ii) (Local dominance) There exists $\delta > 0$ and a measurable function $\dot{m}(y,\mathbf{X})$ having $\mathbb{E}[\dot{m}(y,\mathbf{X})^2] < \infty$ such that

$$\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0^\dagger\|\leq\delta}\|\hat{\psi}(y,\mathbf{X};\boldsymbol{\theta})\| \leq \dot{m}(y,\mathbf{X}).$$

(iii) (Second-order expansion) The map $\boldsymbol{\theta}\mapsto\mathbb{E}\{\hat{m}(y,\mathbf{X};\boldsymbol{\theta})\}$ admits a second-order Taylor expansion at $\boldsymbol{\theta}_0^\dagger$:

$$\mathbb{E}\{\hat{m}(y,\mathbf{X};\boldsymbol{\theta})\} = \mathbb{E}\left\{\hat{m}(y,\mathbf{X};\boldsymbol{\theta}_0^\dagger)\right\} - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0^\dagger)^\top\mathbf{A}(\boldsymbol{\theta}_0^\dagger)(\boldsymbol{\theta}-\boldsymbol{\theta}_0^\dagger) + o(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0^\dagger\|^2),$$

where $\mathbf{A}(\boldsymbol{\theta}_0^\dagger) = -\mathbb{E}\left\{\nabla^2_{\boldsymbol{\theta}}\hat{m}(y,\mathbf{X};\boldsymbol{\theta}_0^\dagger)\right\}$ is positive definite.

B.4 Proof of Theorems 5 and 6

Proof of Theorem 5. Define the sample and population criterion functions

$$M_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\hat{m}(y_i,\mathbf{X}_i;\boldsymbol{\theta}), \qquad M(\boldsymbol{\theta}) = \mathbb{E}\{\hat{m}(y,\mathbf{X};\boldsymbol{\theta})\}.$$

By Assumption 5(iv), $\sup_{\boldsymbol{\theta}\in\Theta}|M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})| \xrightarrow{p} 0$. By Assumption 5(ii)–(iii), $\boldsymbol{\theta}_0^\dagger$ is the unique maximizer of $M(\boldsymbol{\theta})$ and is well-separated in the sense that $\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0^\dagger\|\geq\epsilon} M(\boldsymbol{\theta}) < M(\boldsymbol{\theta}_0^\dagger)$ for every $\epsilon > 0$. By Assumption 5(v), $M_n(\hat{\boldsymbol{\theta}}_n) \geq \sup_{\boldsymbol{\theta}\in\Theta} M_n(\boldsymbol{\theta}) - o_p(n^{-1})$, which implies $M_n(\hat{\boldsymbol{\theta}}_n) \geq \sup_{\boldsymbol{\theta}\in\Theta} M_n(\boldsymbol{\theta}) - o_p(1)$. The result $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0^\dagger$ now follows from Theorem 5.7 of van der Vaart (1998). □

Proof of Theorem 6. We verify the conditions of Theorem 5.23 of van der Vaart (1998) and apply it to obtain the result.

By Assumption 6(i), the map $\boldsymbol{\theta} \mapsto \hat{m}(y, \mathbf{X}; \boldsymbol{\theta})$ is differentiable at $\boldsymbol{\theta}_0^\dagger$ for every $(y, \mathbf{X})$, with derivative $\hat{\psi}(y, \mathbf{X}; \boldsymbol{\theta}_0^\dagger)$. By Assumption 6(ii), the score is locally dominated by a square-integrable function, which implies the Lipschitz condition

$$|\hat{m}(y, \mathbf{X}; \boldsymbol{\theta}_1) - \hat{m}(y, \mathbf{X}; \boldsymbol{\theta}_2)| \leq \dot{m}(y, \mathbf{X}) \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$$

for $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ in a neighborhood of $\boldsymbol{\theta}_0^\dagger$ via the mean value theorem. Assumption 6(iii) provides the required second-order Taylor expansion of $\boldsymbol{\theta} \mapsto \mathbb{E}\{\hat{m}(y, \mathbf{X}; \boldsymbol{\theta})\}$ at the point of maximum $\boldsymbol{\theta}_0^\dagger$, with nonsingular second derivative matrix $-\mathbf{A}(\boldsymbol{\theta}_0^\dagger)$. By Theorem 5, $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0^\dagger$. By Assumption 5(v), the near-maximization condition $M_n(\hat{\boldsymbol{\theta}}_n) \geq \sup_{\boldsymbol{\theta} \in \Theta} M_n(\boldsymbol{\theta}) - o_p(n^{-1})$ is satisfied.

Applying Theorem 5.23 of van der Vaart (1998):

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0^\dagger) = \mathbf{A}(\boldsymbol{\theta}_0^\dagger)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\psi}(y_i, \mathbf{X}_i; \boldsymbol{\theta}_0^\dagger) + o_p(1).$$

By Assumption 6(ii), $\mathbb{E}\left\{\|\hat{\psi}(y, \mathbf{X}; \boldsymbol{\theta}_0^\dagger)\|^2\right\} \leq \mathbb{E}\{\dot{m}(y, \mathbf{X})^2\} < \infty$. The central limit theorem and Slutsky's lemma yield

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0^\dagger) \xrightarrow{d} \mathcal{N}\left\{\mathbf{0}, \mathbf{A}(\boldsymbol{\theta}_0^\dagger)^{-1} \mathbf{B}(\boldsymbol{\theta}_0^\dagger) \mathbf{A}(\boldsymbol{\theta}_0^\dagger)^{-\top}\right\},$$

where $\mathbf{B}(\boldsymbol{\theta}_0^\dagger) = \mathbb{E}\left\{\hat{\psi}(y, \mathbf{X}; \boldsymbol{\theta}_0^\dagger) \hat{\psi}(y, \mathbf{X}; \boldsymbol{\theta}_0^\dagger)^\top\right\}$. □

C Simulation Study Details

This appendix provides additional details on the factor model parameterization, emulator training procedure, and the scaled Wishart distribution used in the simulation study.

C.1 Input Feature Details

This appendix documents the input features for each component of the per-alternative encoder described in Section 4.2. Table 8 summarizes which features are passed to each component. We also describe transformations to reduce sensitivity to outliers.

Feature	Diagonal	Off-diagonal	Pass-through
v_j^*	✓		✓
v_k^*	✓		
Σ_{jj}^*	✓		✓
Σ_{kk}^*	✓		
σ_j	✓		✓
σ_k	✓		
Σ_{jk}^*	✓		
ρ_{jk}	✓		
$\text{ssign}(z_{jk})$	✓		
Σ_{kl}^*		✓	
ρ_{kl}		✓	
$(v_k^* - v_l^*)^2$		✓	
$\text{ssign}(z_{kl})^2$		✓	
$\Sigma_{kk}^* + \Sigma_{ll}^*$		✓	
$\sigma_k + \sigma_l$		✓	
$\overline{\rho_j}$			✓
$\overline{\rho^2}_j$			✓
$\overline{\Sigma}_j$			✓
$\overline{\Sigma^2}_j$			✓
$\text{ssign}(v_j^*/\sigma_j)$			✓
Optional features			
$1_{\{k=1\}}$	✓		
$\text{ssign}(\log C^*)$			✓
$1_{\{j=1\}}$			✓
K			✓

Table 8: Input features for each component of the per-alternative encoder. Diagonal features are indexed by rival $k \neq j$; off-diagonal features are indexed by pairs (k, l) with $k < l$ and $k, l \neq j$; pass-through features are indexed by the focal alternative j. Optional features are included only for non-factorization-invariant models or when training across variable K.

Smooth Soft-Sign Function

Several features are unbounded (e.g., z-scores, signal-to-noise ratios). We apply a smooth soft-sign function to map these to $(-1, 1)$:

$$\text{ssign}(x) = \frac{0.1x}{\sqrt{1 + (0.1x)^2}}. \tag{9}$$

This function is a C^∞ bijection from $\mathbb{R}$ to $(-1, 1)$, so applying it preserves the universal approximation property (Theorem 2).

Feature Definitions

Let j denote the focal alternative and $k, l \neq j$ denote rival alternatives. We define:

$$\sigma_j = \sqrt{\Sigma^*_{jj}}, \qquad \rho_{jk} = \frac{\Sigma^*_{jk}}{\sigma_j \sigma_k}, \qquad z_{jk} = \frac{v^*_j - v^*_k}{\sqrt{\Sigma^*_{jj} + \Sigma^*_{kk} - 2\Sigma^*_{jk}}}.$$

In the implementation, we add a small constant (10^{-8}) to denominators before division and to arguments before taking square roots to ensure numerical stability.

Diagonal DeepSet. For each pair (j, k) with $k \neq j$, the feature vector $\boldsymbol{d}_{jk}$ contains:

$$v^*_k, \quad \Sigma^*_{kk}, \quad \sigma_k, \quad \Sigma^*_{jk}, \quad \rho_{jk}, \quad v^*_j, \quad \Sigma^*_{jj}, \quad \sigma_j, \quad \text{ssign}(z_{jk}).$$

Off-diagonal DeepSet. For each pair (k, l) with $k < l$ and $k, l \neq j$, the feature vector $\boldsymbol{o}_{kl}$ contains:

$$\Sigma^*_{kl}, \quad \rho_{kl}, \quad (v^*_k - v^*_l)^2, \quad \text{ssign}(z_{kl})^2, \quad \Sigma^*_{kk} + \Sigma^*_{ll}, \quad \sigma_k + \sigma_l.$$

These features are symmetric in k and l, as required since the pairs have no natural ordering.

Pass-through features. For each alternative j, the feature vector $\boldsymbol{s}_j$ contains:

$$v^*_j, \quad \Sigma^*_{jj}, \quad \sigma_j, \quad \bar{\rho}_j, \quad \overline{\rho^2}_j, \quad \bar{\Sigma}_j, \quad \overline{\Sigma^2}_j, \quad \text{ssign}(v^*_j/\sigma_j),$$

where the summary statistics are

$$\bar{\rho}_j = \frac{1}{K-1}\sum_{k\neq j}\rho_{jk}, \qquad \overline{\rho^2}_j = \frac{1}{K-1}\sum_{k\neq j}\rho_{jk}^2,$$
$$\bar{\Sigma}_j = \frac{1}{K-1}\sum_{k\neq j}\Sigma^*_{jk}, \qquad \overline{\Sigma^2}_j = \frac{1}{K-1}\sum_{k\neq j}(\Sigma^*_{jk})^2.$$

Optional features. For models that are not factorization-invariant (Definition 1), we include additional features to prevent information loss from the preprocessing transformation (Section 4.1): the normalizing constant $\text{ssign}(\log C^*)$ and index indicators $1_{\{j=1\}}$ and $1_{\{k=1\}}$. When training a single emulator across different numbers of alternatives, we also include K as a feature.

C.2 Emulator Training

We train the emulator using the Adam optimizer (Kingma and Ba, 2015) with the AdamW weight decay variant (Loshchilov and Hutter, 2019). The base learning rate is 0.003 with momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, numerical stability constant $\epsilon = 10^{-8}$, and weight decay 3×10^{-6}.

Learning rate schedule. We use a learning rate schedule with linear warmup followed by inverse-square-root decay. During the first 5,000 steps, the learning rate increases linearly from zero to the base rate. After warmup, the learning rate decays as $\sqrt{t_{\text{warmup}}/t}$, where t is the current step and $t_{\text{warmup}} = 5{,}000$.

Dropout schedule. We apply dropout during training to encourage diversity among learned representations. We set the dropout rate to $1/(10t)$, where t indexes the training episodes from 1 to 400,000.

Training procedure. We pregenerate 400,000 training examples, each consisting of latent inputs $(\boldsymbol{v}, \mathbf{L})$, simulated choice frequencies based on 10^6 Monte Carlo draws, and target gradients computed using the soft relaxation described in Section 4. During training, we randomly generate directions in this latent space and compute directional derivatives as described in Section 4.3. Gradients are clipped to a maximum norm of 0.01 and non-finite gradient penalties are filtered from the loss function. We train for 400,000 optimizer steps with a batch size of 10,000.

Network architecture. Table 9 summarizes the neural network architecture for each value

of K. All networks use the Swish activation function. Larger values of K require larger networks to maintain approximation accuracy.

Component	$K = 3$	$K = 5$	$K = 10$
Diagonal DeepSet			
ϕ output dim	8	24	32
ϕ hidden layers	0	1	1
ϕ hidden dim	N/A	24	32
ρ output dim	8	24	32
ρ hidden layers	0	0	0
Parameters	152	1,440	2,432
Off-diagonal DeepSet			
ϕ output dim	8	24	32
ϕ hidden layers	0	1	1
ϕ hidden dim	N/A	24	32
ρ output dim	8	24	32
ρ hidden layers	0	0	0
Parameters	128	1,368	2,336
Combining MLP			
Output dim	8	24	32
Hidden dim	12	48	64
Hidden layers	2	2	2
Parameters	572	6,312	10,976
Equivariant layers			
Hidden dim	4	12	16
Hidden layers	1	1	1
Parameters	76	612	1,072
Total parameters	928	9,732	16,816

Table 9: Neural network architecture by number of alternatives K. All configurations use an initial dropout rate of 0.1 and Swish activations. When the number of hidden layers is 0, the hidden dimension is not applicable (N/A).

C.3 Derivation of Scaled Wishart PDF

This section derives the PDF of a scaled Wishart random matrix as described in Section 6.1.

Let $\mathbf{W} \sim W_d(\mathbf{V}, n)$ denote a $d \times d$ Wishart random matrix with scale matrix $\mathbf{V}$ and $n \geq d$ degrees of freedom. We derive the distribution of $\mathbf{\Sigma} = \mathbf{W}/W_{jj}$, where W_{jj} is the (j, j) element of $\mathbf{W}$ for some $j \in \{1, \ldots, d\}$.

The Wishart density is

$$f_{\mathbf{W}}(\mathbf{W}) = \frac{|\mathbf{W}|^{(n-d-1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{W})\right\}}{2^{nd/2}|\mathbf{V}|^{n/2}\Gamma_d(n/2)}$$

where $\Gamma_d(\cdot)$ is the multivariate gamma function.

The inverse transformation is $\mathbf{W} = W_{jj}\boldsymbol{\Sigma}$. Note that $\boldsymbol{\Sigma}$ is positive definite with $\Sigma_{jj} = 1$, so it has $d(d+1)/2-1$ free parameters. Together with W_{jj}, this matches the $d(d+1)/2$ unique elements of $\mathbf{W}$. The Jacobian determinant of this transformation is $w_{jj}^{d(d+1)/2-1}$, where w_{jj} denotes the value of W_{jj}.

Substituting $\mathbf{W} = W_{jj}\boldsymbol{\Sigma}$ and applying the Jacobian yields the joint density

$$f_{W_{jj},\boldsymbol{\Sigma}}(w_{jj}, \boldsymbol{\Sigma}) = \frac{|\boldsymbol{\Sigma}|^{(n-d-1)/2}}{2^{nd/2}|\mathbf{V}|^{n/2}\Gamma_d(n/2)} \cdot w_{jj}^{nd/2-1} \exp\left\{-\frac{w_{jj}}{2}\mathrm{tr}(\mathbf{V}^{-1}\boldsymbol{\Sigma})\right\}.$$

Integrating over $w_{jj} > 0$ using the gamma integral gives the marginal density of $\boldsymbol{\Sigma}$:

$$f_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}) = \frac{\Gamma(nd/2)}{2^{nd/2}|\mathbf{V}|^{n/2}\Gamma_d(n/2)} \cdot \frac{|\boldsymbol{\Sigma}|^{(n-d-1)/2}}{\left\{\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\boldsymbol{\Sigma})\right\}^{nd/2}}$$

for symmetric, positive definite $\boldsymbol{\Sigma}$ with $\Sigma_{jj} = 1$ for some fixed $j \in \{1, \ldots, d\}$.

In the simulation study, we use $d = K-1$, $n = K+10$, $\mathbf{V} = \mathbf{I}_{K-1}$, and $j = 1$. Substituting these values and simplifying yields the density given in Section 6.1.