Bayesian Estimation of Large Choice Models with Adaptive Sampling and Likelihood-Free Inference

Anne-Valérie Preto*[1], Xinwei Li*[2], Prateek Bansal[2], and Michel Bierlaire[1]

[1]Transport and Mobility Laboratory (TRANSP-OR), School of Architecture, Civil and Environmental Engineering, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

[2]Behavioural Computational Science (BeCoS) Lab, Department of Civil and Environmental Engineering, National University of Singapore (NUS), Singapore

∗ These authors contributed equally. Corresponding author: anne-valerie.preto@epfl.ch

Short summary

Understanding how individuals make travel decisions is central to transportation research. Discrete choice models provide a rigorous framework to infer preferences from observed behavior and to anticipate responses to policies. However, their practical application fundamentally relies on the ability to enumerate or sample from a predefined choice set. This assumption breaks down in many important contexts, such as route choice on large-scale networks or activity scheduling, where the number of feasible alternatives is combinatorial.

This paper addresses this limitation by introducing estimation methods that do not require explicit enumeration of the choice set. Instead of treating the choice set as a fixed input, we embed alternative generation directly within the estimation process.

We develop two complementary approaches. The first relies on adaptive sampling within a Bayesian framework, using only binary comparisons to recover the full model. The second leverages Bayesian optimization for likelihood-free inference (BOLFI), replacing likelihood evaluation with a learned surrogate of the discrepancy between observed and simulated behavior. Both approaches fundamentally bypass the need to enumerate or approximate the full choice set.

Using controlled synthetic experiments with known ground truth, we show that these methods recover model parameters with the same accuracy as full choice set estimation, while achieving orders-of-magnitude computational gains. Crucially, their performance scales independently of the size of the choice set. These results demonstrate that discrete choice models can be estimated reliably without enumerating alternatives, substantially expanding the range of problems that can be addressed in practice. In particular, they enable the estimation of behaviorally rich models in settings previously considered intractable, including network-wide route choice and combinatorial activity scheduling.

Keywords: Discrete Choice Models, Intractable Choice Sets, Scalable Bayesian Inference, Bayesian Optimization, Likelihood-Free Inference.

1 Introduction

Discrete choice models are central in transportation research. They explain how travelers make decisions, which is essential for demand forecasting, for policy evaluation and for investment appraisal.

However, discrete choice models face a fundamental computational limitation. Choice probabilities require evaluating the full set of alternatives. When the choice set is very large or cannot be enumerated at all, estimation becomes intractable. In route choice, feasible paths grow exponentially with network size. In activity scheduling, combinations of activities, locations, times, and modes are combinatorial. This creates structurally intractable problems.

The standard solution is sampling of alternatives (McFadden (1978)): a fixed subset of the choice set is drawn before estimation, and a bias correction is applied. This approach is well established in the frequentist, maximum likelihood framework. However, the sampling distribution is fixed independently of the model being estimated. Generation and estimation remain disconnected. This works when the choice set is moderate. While the approach has been extended to more general

model structures beyond logit Guevara & Ben-Akiva (2013b,a); Guevara (2010), this paper focuses on multinomial logit as the baseline. In route choice, the exponential growth of paths with network size has motivated sampling approaches Bierlaire & Lucadamo (2007); Frejinger et al. (2009); Lai & Bierlaire (2015), but these still require predefined and fixed sampling distributions. More recent Bayesian and MCMC approaches still require likelihood evaluation, inheriting the same bottleneck. While Approximate Bayesian Computation and Bayesian Optimization have been developed as general inference frameworks (Sisson et al. (2018), Järvenpää et al. (2021), Gutmann & Corander (2016)), their application to discrete choice models with intractable choice sets remains limited. McFadden's correction has been validated recently in Bayesian inference Dekker et al. (2025), but this does not address the core bottleneck: the sampling distribution is still fixed and separate from estimation.

This paper explores an alternative paradigm: the estimation of discrete choice models without enumerating the choice set. Rather than addressing this problem through fixed, exogenously determined sampling of alternatives, we propose two complementary approaches that integrate the generation of alternatives directly into the inference process itself.
The first approach is adaptive sampling, which performs Bayesian inference by comparing each observed choice to only one sampled competitor. The key insight is that this competitor is drawn dynamically, so the generator model evolves as parameters are refined, allowing the procedure to identify behaviorally relevant alternatives.
The second approach is Bayesian Optimization for Likelihood-Free Inference (BOLFI), which avoids computing choice probabilities by learning how well different parameters explain observed choices. It establishes a surrogate model by iteratively exploring parameter space on effectively identified optimal regions to approximate the likelihood of any parameter value being the true parameter of the observed data. Then, it applies MCMC to extract posterior estimates. This works for choice sets that are implicit or combinatorial.
Both methods address the fundamental computational constraint that makes estimation intractable when choice sets are large or implicit.
The remainder of this paper unfolds as follows. We present the mathematical framework underlying both methods and detail their implementation. We then describe a benchmarking study on synthetic restaurant choice data, comparing adaptive sampling and BOLFI against full-set Bayesian inference and maximum likelihood estimation with sampling, across varying choice set sizes. The results demonstrate competitive parameter recovery while avoiding enumeration, with direct implications for route choice and activity-based modeling applications.

2 Methodology

Consider a logit model. For an observation n, the choice probability of alternative i is

$$P(i \mid C_n, \theta) = \frac{\exp(\mu V_i(\theta))}{\sum_{j \in C_n} \exp(\mu V_j(\theta))}.$$

The evaluation of the denominator requires a summation over the full choice set C_n, which is infeasible when $|C_n|$ is large or implicit. We propose two approaches that avoid this bottleneck.

Adaptive sampling

Instead of evaluating the full choice set, we sample a single competitor alternative at each iteration and compare it to the chosen alternative. The key insight is to sample this competitor from the current parameter estimate, so that the estimation procedure learns which alternatives are most relevant as the parameters evolve. We reduce the choice set to its minimum and we avoid enumeration, which reduces the computational cost from $O(|\mathcal{C}|)$ to $O(1)$ per iteration per observation. Over the course of estimation, the chosen alternative is compared to many different competitors.

The procedure has two main elements: the sampling of the competitor alternative, and the sampling of the parameters. Each iteration t can be expressed as follows:

1. **Sample $j^t \sim \mathbb{P}(j \mid \mathcal{C}; \phi^t)$, forming $\mathcal{D}_{ij}^t$.**
   In this step, we form a binary comparison set $\mathcal{D}_{ij}^t = \{i, j^t\}$, where i is the observed choice and j^t is the sampled competitor. We draw j^t from a generation model with parameter ϕ^t:

$$\mathbb{P}(\mathcal{D}_{ij}^t \mid \mathcal{C}; \phi^t) = \mathbb{P}(j^t \mid \mathcal{C}; \phi^t) = \frac{\exp(\mu V_j(\phi^t))}{\sum_{k \in \mathcal{C}} \exp(\mu V_k(\phi^t))}.$$

In practice, the competitor can be refreshed without computing the denominator, using a finite Metropolis–Hastings transition or Multiple-Try Metropolis over a proposal distribution (e.g., uniform over $\mathcal{C}$ or a random walk).

2. **Sample θ_{t+1} from the conditional posterior.**
In this step, we update our parameter estimates by sampling from the posterior distribution conditioned on the binary comparison set $\mathcal{D}_{ij}^t$:

$$\theta_{t+1} \sim \mathbb{P}(\theta \mid \mathcal{D}_{ij}^t, i; \phi^t) \propto \mathbb{P}(i \mid \mathcal{D}_{ij}^t, \theta; \phi^t)\,\mathbb{P}(\theta).$$

If we assume that the competitor is drawn from the target generation distribution, then, conditional on the unordered pair $\mathcal{D}_{ij}^t = \{i, j^t\}$, the probability that i is the observed alternative is:

$$\begin{aligned}\mathbb{P}(i \mid \mathcal{D}_{ij}^t, \theta; \phi^t) &= \frac{\mathbb{P}(j^t \mid \mathcal{C}; \phi^t)\,\mathbb{P}(i \mid \mathcal{C}; \theta)}{\mathbb{P}(j^t \mid \mathcal{C}; \phi^t)\,\mathbb{P}(i \mid \mathcal{C}; \theta) + \mathbb{P}(i \mid \mathcal{C}; \phi^t)\,\mathbb{P}(j^t \mid \mathcal{C}; \theta)} \\ &= \frac{\exp(\mu V_j(\phi^t))\,\exp(\mu V_i(\theta))}{\exp(\mu V_j(\phi^t))\,\exp(\mu V_i(\theta)) + \exp(\mu V_i(\phi^t))\,\exp(\mu V_j(\theta))}.\end{aligned}$$

The first equality follows from the induced binary comparison. The normalizing sums over $\mathcal{C}$ cancel exactly in the second equality. When $jt = i$, the pair collapses to a singleton and contributes zero to the binary log-likelihood. This expression requires no evaluation of the full choice set $\mathcal{C}$, reducing the computational cost per iteration.
The generation parameter is set adaptively to $\phi^t = \theta^t$ at each iteration. This is not a standard MCMC scheme targeting a fixed distribution; rather, it is a stochastic-approximation procedure where the sampling distribution evolves with the parameter estimates. As θ_t stabilizes, $\phi^t = \theta^t$ also stabilizes and the transition mechanism becomes approximately stationary.

Trust-region Bayesian optimization for likelihood-free inference

Trust-region Bayesian Optimization for Likelihood-Free Inference (TuRBOLFI) is a simulation-based inference method that bypasses likelihood evaluation entirely. Rather than computing choice probabilities over the full choice set, TuRBOLFI uses Bayesian Optimization to efficiently fit a Gaussian Process surrogate model on the discrepancy between observed and simulated choices. This approach avoids enumeration and applies to general discrete choice models.
The key insight is to avoid likelihood evaluation by leveraging well-chosen discrepancy measures and a surrogate model that learns the mapping from parameters to discrepancy. We employ a two-stage approach using two complementary discrepancies for multinomial logit, progressing from coarse to fine likelihood approximation.
The procedure has two main elements: the sequential discrepancy specifications, and the iterative optimization and inference.

1. **Specify the discrepancy sequentially and generate simulations.**
For a given parameter candidate ϕ_t at iteration t, we independently generate M synthetic choice alternatives from the model. We compute the discrepancy between observed and simulated choices in two stages:

   Stage 1: Utility-based discrepancy for initial results. We first train the surrogate model using utility-based discrepancy, which compares the utilities of the observed choice against simulated alternatives:

$$\Delta^u(\theta, \phi_t)_n = \frac{\sum_{m=1}^M \exp\left\{\mu\left(V_{j_{nt}^{(m)}}(\phi_t) - V_{i_n}(\phi_t)\right)\right\}}{M},$$

   where i_n is the observed choice, $\{j_{nt}^{(m)}\}_{m=1,\cdots,M}$ are independently simulated choice based on ϕ_t, and μ is the known scale parameter. The total discrepancy is $\Delta^u(\theta, \phi_t) = \sum_{n=1}^N \log \Delta^u(\theta, \phi_t)_n$. The utility-based formulation is concave in the parameter space, allowing Bayesian Optimization to rapidly locate a feasible region.

   Stage 2: Attribute-based discrepancy for refinement. Once the utility-based optimization converges, we switch to attribute-based discrepancy for finer exploration:

$$\Delta^a(\theta, \phi_t) = \frac{\sum_{m=1}^M \left(T_\theta - T_{\phi_t}^{(m)}\right)^\top W \left(T_\theta - T_{\phi_t}^{(m)}\right)}{M},$$

where T_θ, a sufficient summary statistic for MNL, is the sum of attribute vectors for chosen alternatives in the data and simulations (analogous for $T^{(m)}_{\phi_t}$ based on $j^{(m)}_{nt}$). And W is a variance-weighting matrix of T_{ϕ^*} when $\phi^* = \arg\min\{\phi | \Delta^u(\theta, \phi_t), t = 1, \cdots, t_1\}$ is the best acquisition point in stage 1 coarse optimization. This attribute-based specification refines the parameter estimates.

This sequential approach reaches better optimization results more efficiently by first using utility-based discrepancy to identify the promising region, then refining within that region using attribute-based discrepancy.

2. **Fit a Gaussian Process surrogate and optimize.**
A Gaussian Process model approximates the discrepancy as a function of parameters:

$$\Delta(\theta, \phi) = f(\phi) + \eta, \quad f(\phi) \sim GP_\theta(\mu_\zeta(.), k_\zeta(.))$$

where $\eta \sim N(0, \sigma^2)$ is simulation noise and ζ parametrizes the GP.

Bayesian Optimization is used to select parameter values that are most informative for learning the GP in the next round $t + 1$. The procedure initializes with t_0 random samples (typically twice the number of parameters), then iterates:

- Fit the GP to the accumulated discrepancy evaluations $\{\Delta(\theta, \phi_t)\}_{t=1,\cdots,t_0}$.
- Select the next candidate ϕ_{t+1} from $GP_\theta(\zeta_t)$ with predicted low discrepancy.
- Evaluate the discrepancy at the new candidate and update $t = t + 1$.

The transition between stages and final termination rely on the trust-region algorithm when no candidate point improves the current best acquisition value by more than a predefined threshold for several consecutive attempts (Eriksson et al. (2019)).

3. **Sample from the posterior using the learned surrogate.**
Once the GP has converged, MCMC sampling is applied to $GP_\theta(\zeta_t)$ to draw posterior samples. The fitted GP may be multimodal, so MCMC is preferred for global exploration and handling parameter correlations.

Benchmarking framework

The proposed methods are evaluated on a synthetic restaurant choice dataset with known parameters, providing a controlled environment to systematically compare estimation methodologies. An individual n chooses from a universal choice set $\mathcal{C}$ of J restaurants. For each individual, the utility of restaurant j is:

$$V_{jn} = \theta_{\text{rate}} \cdot \text{rating}_j + \theta_{\text{cost}} \cdot \text{cost}_j + \theta_{\log d} \cdot \log(\text{dist}_{jn}) + \sum_{c \in \mathcal{C}_{\text{cuisine}}} \theta_c \, \mathbf{1}\{j \text{ has cuisine } c\},$$

We benchmark four estimators:

1. Full-likelihood Bayesian inference (BI) using the complete choice set.
2. Maximum Likelihood Estimation with uniform sampling of 40% of alternatives (MLE-SA).
3. The proposed Adaptive Sampler (AS).
4. The proposed TurBOLFI.

Parameter recovery is assessed across all parameter estimates and configurations using posterior means and credible intervals.
Figure 1 presents posterior estimates for all 10 parameters across experimental configurations. The choice set size varies as $J \in \{100, 200, 500\}$ and sample size as $N \in \{10^4, 2 \times 10^4, 5 \times 10^4\}$. For each configuration, we report the posterior mean and 94% highest density interval (or Confidence Interval for MLE) for all four estimators. The red line indicates the known true parameter value. The results show that both the Adaptive Sampler and TurBOLFI recover parameters with accuracy comparable to full-likelihood Bayesian inference and MLE-SA. The Adaptive sampler has slightly wider credible intervals, reflecting the additional uncertainty from sampling a single competitor, but it remains robust across all configurations. TurBOLFI achieves similar accuracy with narrower intervals.

For Bayesian methods, we ran all chains until convergence, using four independent MCMC chains per configuration with diagnostics assessed via $\hat{R} < 1.01$. This convergence criterion ensures reliable posterior inference and credible interval coverage.

The adaptive sampler demonstrates competitive computational properties across the configurations. For smaller configurations, AS and MLE-SA exhibit comparable runtimes. However, choice set size does not influence AS performance: memory usage grows only with the number of observations, remaining independent of J (0.4 GB to 2.4 GB across configurations), while MLE-SA memory scales with subset size, reaching 54.54 GB at the largest configuration. At large choice sets, AS runtime becomes substantially more efficient, running 3.2 times faster than MLE-SA at $J = 500, N = 50,000$. Table 1 compares wall-clock time and peak memory usage across all three configurations on the same machine, providing a direct benchmark.

The TurBOLFI estimator achieves parameter recovery comparable to AS and BI, with the advantage of operating as a likelihood-free method. Its total computational cost consists of two phases: approximately 30 minutes for the training process (fitting the surrogate model and Bayesian Optimization), followed by a few minutes for posterior sampling via MCMC. This upfront cost favors scenarios where multiple inference tasks or model variants are needed, since posterior sampling itself remains fast once the surrogate is fitted. The likelihood-free property provides robustness when the choice model structure is uncertain or when explicit likelihood evaluation is infeasible.

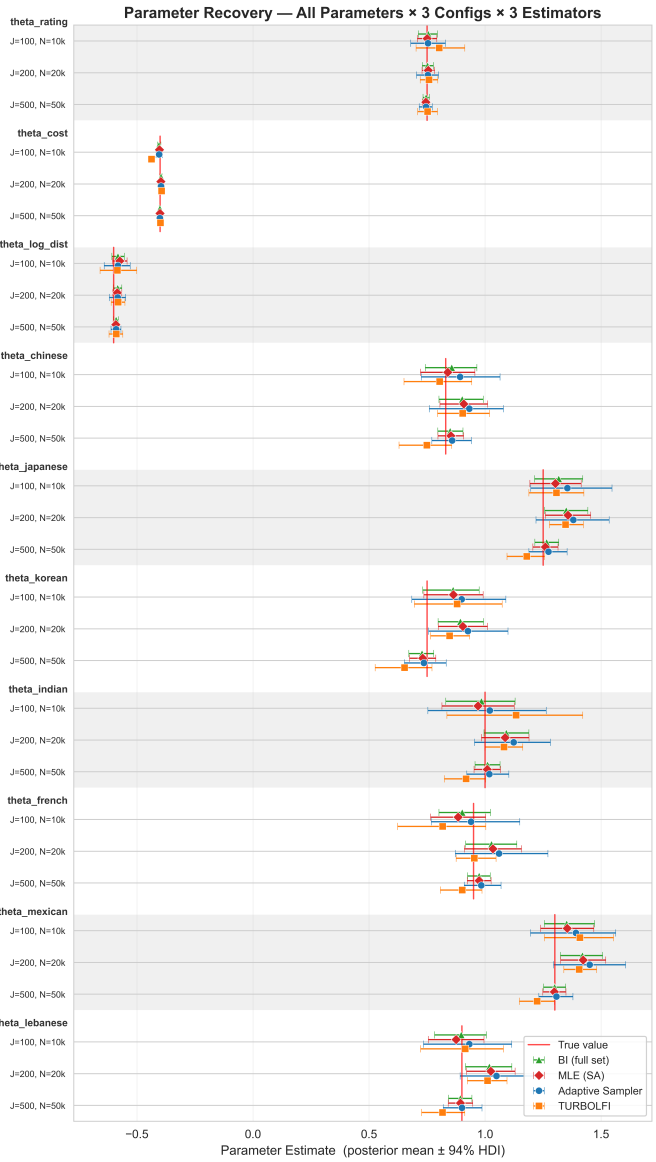


Figure 1: Parameter recovery.

Config.	Wall time (mm:ss)				RAM (GB)			
(J, N)	BI	MLE-SA	AS	TuRBOLFI	BI	MLE-SA	AS	TuRBOLFI
100, 10k	10:49	2:35	3:38	42:28	121.6	2.60	0.42	11.22
200, 20k	38:35	6:30	6:19	63:37	349.3	9.39	0.67	15.23
500, 50k	197:40	48:04	14:50	113:14	~800	54.54	2.43	29.11

Table 1: Computational efficiency comparison.

Results and discussion

All four estimators recover parameters with comparable accuracy across configurations, as shown in Figure 1. The key distinction emerges in computational scalability: the adaptive sampler's memory and runtime scale only with the number of observations, uniformly outperforming alternatives at large choice sets. Notably, while AS exhibits slightly wider credible intervals, this reflects the fact that AS compares only two alternatives (the observed choice and a single sampled competitor), yet still achieves comparable parameter recovery to MLE-SA, which samples 40–200 alternatives depending on choice set size. This efficiency is a testament to the power of adaptive generation.

These results validate the initial idea: by integrating sampling into the inference loop and adapting the generation mechanism to evolving parameters, we achieve a method that scales to large choice sets while maintaining parameter recovery accuracy.

3 Conclusions

Large choice sets represent a fundamental barrier to practical discrete choice estimation. This paper proposes two complementary solutions: adaptive sampling, which integrates alternative generation into a Bayesian inference loop with computational cost independent of choice set size, and BOLFI, which bypasses likelihood evaluation entirely through a likelihood-free framework. Both methods recover parameters with accuracy comparable to full-set inference while removing the bottleneck of explicit likelihood evaluation. The adaptive sampler scales uniquely: memory usage remains independent of choice set size, growing only with the number of observations. BOLFI offers flexibility when explicit model structure is uncertain. These results demonstrate that even in a synthetic setting with known parameters, we can achieve accurate estimation without enumerating the choice set, opening the door to applications previously considered computationally infeasible. The practical implications are immediate. Route choice on realistic networks, where path enumeration is infeasible, becomes estimable. Activity-based models with combinatorial schedule spaces, previously intractable, are now within reach. For practitioners, adaptive sampling is preferred when choice set size is large and the multinomial logit structure holds; BOLFI becomes valuable when choice sets are implicit or model structure is uncertain. Extensions to mixed logit and nested logit are straightforward. These methodologies remove a longstanding barrier between behavioral theory and computational feasibility, enabling choice model estimation at the scale required by modern transportation systems.

Acknowledgements

Anne-Valérie Preto and Michel Bierlaire developed the Adaptive Sampling methodology at EPFL. Xinwei Li and Prateek Bansal developed TurBOLFI at NUS. This work benefited from a fruitful collaboration between EPFL and NUS, where the researchers shared knowledge and ideas to ensure both methods were robust and comparable. The empirical benchmarking framework was designed jointly to provide fair and rigorous performance assessment across both approaches.

References

Bierlaire, M., & Lucadamo, A. (2007). *Sampling of alternatives using multidimensional analysis* (Technical Report No. TRANSP-OR 070522). Lausanne, Switzerland.

Dekker, T., Bansal, P., & Huo, J. (2025). Revisiting McFadden's correction factor for sampling of alternatives in multinomial logit and mixed multinomial logit models. *Transportation Research Part B: Methodological*, *192*, 103129. doi: 10.1016/j.trb.2024.103129

Eriksson, D., Pearce, M., Gardner, J., Turner, R. D., & Poloczek, M. (2019). Scalable global optimization via local bayesian optimization. *Advances in neural information processing systems*, *32*.

Frejinger, E., Bierlaire, M., & Ben-Akiva, M. (2009). Sampling of alternatives for route choice modeling. *Transportation Research Part B: Methodological*, *43*(10), 984-994. (Accepted on Mar 03, 2009) doi: 10.1016/j.trb.2009.03.001

Guevara, C. A. (2010). *Endogeneity and sampling of alternatives in spatial choice models* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Guevara, C. A., & Ben-Akiva, M. (2013b). Sampling of alternatives in multivariate extreme value (MEV) models. *Transportation Research Part B*, *48*, 31– 52. doi: 10.1016/j.trb.2012.11.001

Guevara, C. A., & Ben-Akiva, M. E. (2013a). Sampling of alternatives in logit mixture models. *Transportation Research Part B: Methodological*, *58*, 185–198. doi: 10.1016/j.trb.2013.08.011

Gutmann, M. U., & Corander, J. (2016, January). Bayesian optimization for likelihood-free inference of simulator-based statistical models. *J. Mach. Learn. Res.*, *17*(1), 4256–4302.

Järvenpää, M., Gutmann, M. U., Vehtari, A., & Marttinen, P. (2021, March). Parallel gaussian process surrogate bayesian inference with noisy likelihood evaluations. *Bayesian Analysis*, *16*(1). Retrieved from `http://dx.doi.org/10.1214/20-BA1200` doi: 10.1214/20-ba1200

Lai, X., & Bierlaire, M. (2015). Specification of the cross-nested logit model with sampling of alternatives for route choice models. *Transportation Research Part B: Methodological*, *80*, 220-234. (Accepted on Jul 04, 2015) doi: 10.1016/j.trb.2015.07.005

McFadden, D. (1978). Modelling the choice of residential location. In A. Karlquist *et al.* (Ed.), *Spatial interaction theory and residential location* (p. 75-96). Amsterdam: North-Holland.

Sisson, S. A., Fan, Y. ., & Beaumont, M. A. (Eds.). (2018). *Handbook of approximate bayesian computation*. Chapman and Hall/CRC.