Bayesian Optimization for Discrete Choice Model Likelihood Estimation

Hongxian Huang [1] **Shih-Lun Huang** [2] **Shreyas Pimpalgaonkar** [3]

1. Abstract

In this study, we examine Bayesian Optimization (BO) as an alternative to Maximum Likelihood Estimation (MLE) for logit choice models. Our findings show that while BO quickly converges to local minima, it can struggle to identify global optima efficiently. However, with extensive iterations, its accuracy approaches that of MLE. We propose a novel hybrid method, combining MLE's precision with the convergence speed of Parallel Bayesian Optimization, achieving comparable accuracy to MLE but more efficiently. This approach is especially pertinent for complex models, highlighting the potential of our hybrid method in computationally challenging scenarios.

2. Introduction and Related Work

A choice model is a sophisticated mathematical framework used to understand and predict the decision-making processes of individuals or groups. The basis of choice modeling lies in the theory of utility, which suggests that individuals make decisions based on the perceived satisfaction or utility they derive from each option (Porell & Adams, 1995; Kim et al., 2014; Lovreglio et al., 2014). In essence, the higher the utility, the more likely an option is to be chosen. There are various types of choice models, each tailored to specific scenarios. Discrete choice models, for instance, are used when the options are distinct and countable, like choosing a brand of cereal (Nevo, 2000; 2001). Continuous choice models, on the other hand, are applicable when choices can be measured on a continuous scale, such as deciding how much money to save (Mansur et al., 2005; Hewitt & Hanemann, 1995; Garen, 1984). Additionally, stochastic models introduce a probability component to account for the randomness and unpredictability inherent in decision-making processes (Pinsky & Karlin, 2010; Mesbah, 2016; Evenson & Kislev, 1976). Key components of choice models include the attributes of each choice, which are the characteristics influencing decision-making, and the characteristics of the decision-makers themselves, such as personal or demographic factors. Constraints, like budget or availability, also play a significant role in shaping choices.

[1] hh1643, hh1643@stern.nyu.edu
[2] sh7008, sh7008@nyu.edu
[3] sgp9467, sgp9467@nyu.edu

In this paper, we are going to restrict our attention on discrete choice models, which form a cornerstone in the quantitative analysis within the fields of economics (Hensher & Johnson, 2018; Greene, 2009; Berry, 1994), transportation (Cirillo & Xu, 2011; Rust, 1987; 1994), and business academia (Chintagunta & Nair, 2011; Ben-Akiva & Boccara, 1995; Anderson et al., 1992). The modeling process involves several crucial steps. It starts with data collection, where information about actual choices made and the factors influencing them is gathered. This is followed by model specification, where the functional form of the model is defined. The next step is estimation, which involves using statistical methods to determine the parameters of the model. Finally, validation is performed by testing the model with new data to ensure its accuracy and reliability (Feng et al., 2022). Among the most popular choice models are the Logit and Probit models. The Logit model is favored for its simplicity and ability to calculate choice probabilities effectively (Demaris, 1992). The Probit model is similar but assumes a different distribution for the error term. The Nested Logit model is also notable for allowing the grouping of similar choices, reflecting their similarities in the decision-making process (Muthén, 1979).

Nevertheless, estimating these discrete choice models presents a significant challenge owing to the non-differentiable nature of the likelihood functions involved. A salient illustration is the likelihood function of the Mixed Logit model, which incorporates an integral within its formulation and lacks an explicit analytical expression (Hensher & Greene, 2003). Historically, researchers have frequently employed the simulated maximum likelihood estimation method to address these complexities, yet this approach does not ensure convergence and global optimum, and is also invariably associated with substantial computational demands (Lee, 1997; Cappellari & Jenkins, 2003; Bhat, 2001).

Our research project tends to utilize the Bayesian Optimization method, which is well-suited for optimization problems where the objective function is (i) non-differentiable, (ii) does not have an explicit form or easy-to-compute gradients, (iii) has multiple local optima, and (iv) expensive to evaluate (Pelikan et al., 1999; Frazier, 2018; Shahriari et al., 2015). To be specific, for those frequently-used discrete choice models in literature such as Simple Logit and Mixed

Logit Model, we try to apply the Bayesian Optimization method to estimate the key parameters in them, in place of the simulated maximum likelihood estimation procedure.

Our comparative analysis of the Bayesian Optimization method and the (simulated) likelihood estimator, within the context of discrete choice models, focuses on two key dimensions: (i) the rapidity of convergence, and (ii) the efficacy in identifying the global optimum. We will apply both methods on a simulated data set for which we know the 'true' parameters and a real-world data set used by previous study. In Particular, we will focus on two prevalent discrete choice models: the Simple Logit model and Mixed Logit model. These methodological frameworks enable us to deduce preferences for the factors influencing decisions (such as price) by examining a sequence of choices made by individuals. Furthermore, the Logit framework assumes that agents choose an option that gives the maximum utility which is done by maximizing the likelihood functions (Hensher & Greene, 2003; Demaris, 1992; Hausman & McFadden, 1984).

As a beginning, the Simple Logit model posits that the probability of making choices follows a multi-nomial distribution, with its likelihood function formulated as

$$\mathcal{L}(\theta) = \prod_{t=1}^{T} \prod_{i=1}^{N} \prod_{j=1}^{J} P(Y_{it} = j \mid X_{jt})^{y_{ijt}} \tag{1}$$

where Y_{it} shows individual i's choice at time t, β denotes the vector of parameters that represents individual i's sensitiveness towards choice j's time-varying characteristics X_{jt}; $y_{ijt} = 1$ if individual i chooses choice j at time t, and equals 0 otherwise; T is the total number of time periods; N is the number of observations; and M is the number of possible choices. We always parameterize the choice probability $P(Y_{it} = j \mid X_{jt})$ using the softmax function:

$$P(Y_{it} = j \mid X_{jt}) = \frac{exp(\alpha_j + \beta X_{jt})}{1 + \sum_{k=1}^{J} exp(\alpha_k + \beta X_{kt})} \tag{2}$$

, which implies $\theta = (\alpha_{k=1,\dots,J}, \beta)$.

We then explores the Mixed Logit model, also known as the Random Coefficient Logit model, which relaxed the Independence of Irrelevant Alternatives (IIA) assumption by allowing taste of parameters to be heterogeneous across individuals (Hensher & Greene, 2003). In other words, the likelihood of choosing one alternative over another can be affected not only by the characteristics and availability of other options in the set, but also by the personal characteristics and tastes of the individuals themselves. For this study, we will focus on the case where each individual i's β_i (i.e. preference towards different factors such as price) follows a continuous distribution such as multi-variate Gaussian distribution, and the likelihood function for each individual i can hence be expressed as

$$\mathcal{L}_i(\theta) = \int_{\beta_i} [\prod_{t=1}^{T} \prod_{j=0}^{J} P(j \mid \beta_i)^{y_{ijt}}] \phi(\beta_i) d\beta_i \tag{3}$$

, where $\phi(\beta_i)$ is the density function of a Gaussian Distribution $N(\bar{\beta}, \sigma_\beta^2)$, and we still parameterize the choice probability $P(j \mid \beta_i)$ in a way similar to equation (2):

$$P(j \mid \beta_i) = \frac{exp(\alpha_j + \beta_i X_{jt})}{1 + \sum_{k=1}^{J} exp(\alpha_k + \beta_i X_{kt})} \tag{4}$$

Based on equation (3), we can also get the full likelihood function of the Mixed Logit model as:

$$L(\theta) = \prod_i L_i(\theta) = \prod_i \int_{\beta_i} [\prod_{t=1}^{T} \prod_{j=0}^{J} P(j \mid \beta_i)^{y_{ijt}}] \phi(\beta_i) d\beta_i \tag{5}$$

, where $\theta = (\alpha_{k=1,\dots,J}, \bar{\beta}, \sigma_\beta^2)$.

It is worth mentioning that obtaining the likelihood function would require computing the integral of a complex term, which is intractable and often impossible to solve analytically. The current approach is to implement the simulated maximum likelihood estimator (Lee, 1997), which however, can be computationally intensive, and has no guarantee of the convergence. Hence, we propose utilizing Bayesian optimization to address these challenges, which will be discussed in detail in the next section.

3. Proposed Method and Innovation

Our methodological innovation mainly lies in the fact that although the discrete choice model has been developed more than 40 years (Chintagunta & Nair, 2011), to the best of our knowledge, there exists no previous paper applying the Bayesian Optimization method for the maximum likelihood optimization evaluation, which is quite surprising. Furthermore, we are going the treat the 'traditional' (simulated) maximum likelihood estimator as the 'baseline', and compare our proposed Bayesian optimization estimator's performance with it by applying both of the methods on one simulated data set, for which we know the 'true' parameters. Following is a brief overview and applicability of the proposed approach.

3.1. Bayesian Optimizaton

Bayesian optimization is a class of machine-learning based optimization methods focused on solving the optimization problem to find a global optimum of the following: $\max_{x \in A} f(x)$, where $f, x,$ and A typically satisfy certain regularity constraints. It is especially effective when evaluations are costly, and the function lacks derivatives, which

prohibits the use of gradient descent (Ruder, 2016) or Newton's methods (Moré & Sorensen, 1982). It is particularly useful in scenarios where collecting new data is time-consuming or costly, such as hyperparameter tuning in machine learning, material design, or drug discovery. This method stands out for its efficiency in finding the best parameters with as few evaluations as possible, which is achieved through a smart balance between exploration of new possibilities and exploitation of known promising areas. Additional conditions are mentioned in detail in Frazier 2018 and Snoek et al. 2012. In our setting, the likelihood function and its gradient is expensive to evaluate, and do not have explicit analytical forms, thereby making the approach a viable choice.

At the heart of Bayesian optimization is the use of a surrogate probabilistic model to estimate the objective function. The most common model used is a Gaussian Process (GP) (MacKay et al., 1998), though others like Bayesian neural networks can also be used (Kononenko, 1989). A Gaussian Process is a collection of random variables, any finite number of which have a joint Gaussian distribution (MacKay et al., 1998). It is favored because of its flexibility, ability to quantify uncertainty, and tractable computational properties. By assuming a GP prior over the function, Bayesian optimization efficiently encodes assumptions about smoothness and can provide uncertainty estimates for predictions. The optimization process involves two main components: the acquisition function and the Bayesian update mechanism Rasmussen & Williams 2005. The acquisition function guides the selection of the next point to evaluate by balancing exploration and exploitation. Several acquisition functions exist, such as Expected Improvement (Zhan & Xing, 2020), Probability of Improvement (Couckuyt et al., 2014), and Upper Confidence Bound (Greene, 2009). Each of these functions has its way of dealing with the trade-off between exploring new, uncertain areas and exploiting areas where the model predicts high values.

Once a new data point is chosen and evaluated, the model is updated. This Bayesian update step incorporates the new observation into the GP, refining the model's understanding of the objective function. The updated model is then used to select the next point for evaluation, and this process is repeated until a stopping criterion is met (Rasmussen & Williams, 2005; Wilson, 2014). This could be a predefined number of iterations, a convergence threshold, or when the improvement becomes negligible. The optimization routine (**Algorithm 1**). consists of two main components, a Bayesian statistical model for modeling the objective function (could be Gaussian processes) and an acquisition function that governs the sampling. We extensively studied Gaussian Processes (GP) in the BML course offered at NYU, an overview can be found in Rasmussen & Williams 2005 and Wilson 2014. In short, Gaussian processes are a non-parametric, Bayesian approach to machine learning that define a prior over functions, where any finite collection of function values is jointly Gaussian, characterized by mean and covariance functions. Acquisition functions are used for efficient sampling. There are multiple commonly used acquisition functions like Expected Improvement, Knowledge Gradient, Entropy Search and Predictive Entropy Search, and Multi-Step Optimal Acquisition Functions. Expected Improvement, the most popular algorithm, is defined as $EI_n(x) := E_n[|f(x) - f_n^*]^+]$ where $E_n[.] = E[.|x_{1:n}, y_{1:n}]$ and the algorithm evaluates at the point with the largest expected improvement $x_{n+1} = \text{argmax} EI_n(x)$. In our setting we test with different covariance kernels for the gaussian process. Also, we compare performance of the algorithm with different acquisition functions.

Algorithm 1 Pseudo-code for Bayesian optimization[a]

Initialize with a Gaussian process prior over f
Sample function f at n_0 points using a predefined experimental design for initial spacing. Assign $n = n_0$.
while $n \leq N$ **do**
  Refresh the posterior distribution of f incorporating all the data observed
  Identify x_n that maximises the acquisition function derived from the updated posterior of f
  Record $y_n = f(x_n)$ as the new observation
  Proceed to the next iteration by incrementing n
end while
Return the solution: the sampled point with either the highest function value of $f(x)$ or the greatest posterior mean.

Bayesian optimization is particularly advantageous in high-dimensional spaces where grid or random search would be prohibitively expensive. Its ability to incorporate prior knowledge and quantify uncertainty makes it robust against overfitting and capable of efficiently navigating complex search spaces.

3.2. Application

The applicability of our proposed 'Bayesian Optimization estimator' is notably extensive. Given that nearly all sophisticated choice models—ranging from the well-known BLP (Berry, Levinsohn, and Pakes) model (Berry, 1994; Berry et al., 1995) to dynamic discrete choice models (Rust, 1987; Aguirregabiria & Mira, 2010) — necessitate the optimization of their respective likelihood functions, our estimator can be applied across a vast spectrum of scenarios.

4. Results and Discussion

4.1. Creating Simulated Datasets

We will use one simulated data set to test the performance of our proposed Bayesian optimization estimator, under which we know the values of the 'true' parameters. For our simulated data set, we consider a brand choice case where consumer $i, (i = 1, ..., I)$ chooses to buy one of the J brands or not to buy anything $(j = 0)$ at time $t(t = 1, ..., T)$. Consumer i's utility from choosing option $j > 0$ at time t is given by: $u_{ijt} = \alpha_j + \beta_i p_{jt} + \epsilon_{ijt}$ where p_{jt} is the price of brand j at time t, α_j is the brand intercept, β_i is consumer i's price sensitivity, and ϵ_{ijt} is an i.i.d. extreme value distributed idiosyncratic error.

For Simple Logit model, we simulated four consumers with $\alpha = [0.2, 0.3, 0.4, 0.2, 0.15]$, respectively with a uniform price sensitivity $\beta = -0.1$. This setup allows us to model consumer choice behavior under fixed individual preferences and a common response to price changes.

In contrast, the Mixed Logit model introduces variability in consumer preferences. We normalize the mean utility level for $j = 0$ to be zero, i.e., $u_{i0t} = \epsilon_{i0t}$. And we also assume that the population distribution of β_i is normal with mean $\bar{\beta}$ and standard deviation σ_β, i.e., $\beta_i \sim N(\bar{\beta}, \sigma_\beta)$. Then we can set the values for the Number of consumers I, brands J, and time periods T, as well as the true parameter values α_j, $\bar{\beta}$, and σ_β to simulate the data set.

To be more specific, we firstly simulate the prices $p_{jt} \sim N(\bar{p}, \sigma_p^2)$, and use the following values as the 'true parameters' to simulate consumers' utilities and choices:

- $I = 100$, $J = 2$,and $T = 10$.
- $(\alpha_1, \alpha_2, \bar{\beta}, \sigma_\beta, \bar{p}, \sigma_p)$ = $(1.0, 1.0, -0.1, 0.5, 2.0, 0.5)$.

4.2. Simple Logit model

The likelihood function of a Simple Logit model is tractable. Hence, here we compared the performance of Maximum Likelihood Estimator, specifically using the Quasi-Newton method for the optimization process, to Bayesian Optimization. We ran the simple logit model for different parameter settings on a simulated dataset and compared their estimates. Quasi-Newton estimates yielded a loss of 3,577 on the dataset and the computation took 4 seconds to complete. On the other hand, BayesOpt took 21 seconds to finish with a loss of 3582. As expected, the gradient-based optimizer outperformed Bayesian Optimization estimator in both computation efficiency and accuracy.

4.3. Mixed Logit model

For Mixed Logit model, we set the coefficients of the utility functions β to follow a Gaussian distribution, which also specifies the reason why the Mixed Logit model is also called the 'Random Coefficient Model'. We thoroughly investigated both the efficiency and accuracy of Bayesian Optimization in this complex scenario through four detailed experiments.

4.3.1. Fixed initial parameters

In the first experiment, we wanted to test how the efficiency and accuracy of Bayesian Optimization compares to the simulated maximum likelihood estimator. We started by giving both methods the same initialization of $\alpha_1 = 1$, $\alpha_2 = 1$, $\bar{\beta} = -1$, and $\sigma_\beta = 0.25$. The simulated maximum likelihood estimator yielded a result of $\alpha_1 = 1.1203$, $\alpha_2 = 1.4972$, $\bar{\beta} = -0.0985$, and $\sigma_\beta = 0.3550$, for which the process took 359 seconds with a loss of 991.5201. In contrast, for the Bayesian Optimization method, we fixed it with RBF kernel and ran for 100 iterations. The hypothesis space of the parameters were shown as follows: $\alpha_1 = 1$, $\alpha_2 = 1$, $\bar{\beta} = -1$, and $\sigma_\beta = 0.25$. In addition, we implemented both Expected Improvement (EI) and Lower Confidence Bound (LCB) as the acquisition function, separately. The Expected Improvement (EI) method resulted in a loss value of 999, with a processing duration of 168 seconds. The estimators derived from this method were $\alpha_1 = 1.9073$, $\alpha_2 = 2.2351$, $\bar{\beta} = -0.4441$, and $\sigma_\beta = 0.5407$. On the other hand, the Lower Confidence Bound (LCB) approach yielded a loss of 1006, with the estimators $\alpha_1 = 0.3825$, $\alpha_2 = 0.8125$, $\bar{\beta} = 0.2851$, and $\sigma_\beta = 0.6694$, over a similar processing time of 166 seconds. We can see that despite converging faster and yielding a similar loss, Bayesian Optimization estimator provided worse estimators than the simulated maximum likelihood estimator.

4.3.2. Accuracy enhancement

We employed a two-step optimization process to thoroughly explore the possibility of achieving better estimators. Initially, we derived estimators using a numerical method. These estimators were then used as the starting point for Bayesian Optimization, which was set up identically to the previous experiment. After 100 iterations, it was observed that the best parameters and the minimum loss value remained at the initial state provided by the numerical method. This outcome suggests that in this specific case, the application of Bayesian Optimization did not enhance the accuracy beyond what was achieved with the initial numerical approach.

4.3.3. Multiple starting point approach

To address the limited exploration observed in previous experiments, we employed a third experiment by initializing Bayesian Optimization at multiple random points within the parameter space. This approach broadened the algorithm's

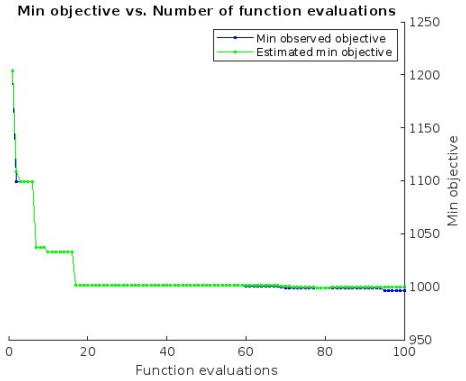

Figure 1. Bayesian Optimization with Expected Improvement. Initialized at $\alpha_1 = 1$, $\alpha_2 = 1$, $\bar{\beta} = -1$, and $\sigma_\beta = 0.25$.

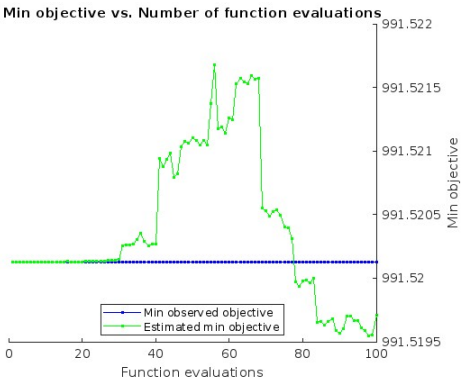

Figure 2. Bayesian Optimization with Expected Improvement. Initialized at the optimized results after numerical optimization.

exploration, enhancing its understanding of the space prior to the optimization phase. However, the multiple starting point approach does entail a higher preparation cost. We conducted the experiment by randomly selecting ten initial sets of parameters while reducing the Expected Improvement function's exploration rate to 0.3. After completing 100 iterations in 267 seconds, the model achieved a loss of 992.1867 which closely approaches that of the MLE method (991.5201), as depicted in Figure 3. The optimized parameters were as follows: $\alpha_1 = 0.9577$, $\alpha_2 = 1.3165$, $\bar{\beta} = -0.0462$, and $\sigma_\beta = 0.3682$.

4.3.4. Hybrid approach

According to the experiments in the previous section, Bayesian Optimization showed ability to quickly converge to a near-optimal solution in complex optimization problems. This characteristic is particularly beneficial in situations where the likelihood function is highly non-linear or exhibits multiple local maxima or minima. By rapidly identifying a region in the parameter space that is close to optimal, BO could effectively bypass the potentially time-consuming process of exploring less promising areas. We saw that while BO is efficient in quickly locating a promising area of the solution space, it may not always be precise in pinpointing the exact optimal solution, especially in highly complex landscapes. MLE, on the other hand, excels in accuracy but can be slow and resource-intensive. Combining BO and MLE thus aims to balance these trade-offs, utilizing the speed of BO for initial convergence and the precision of MLE for final optimization.

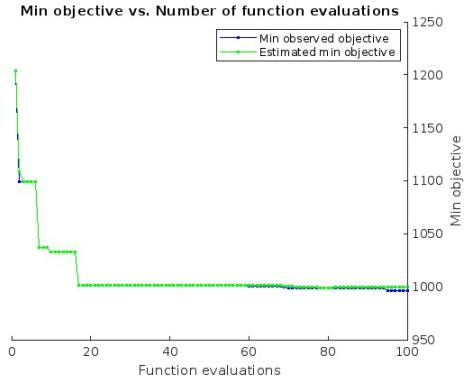

Figure 3. Bayesian Optimization with multiple starting points.

In our empirical study, we conducted a comparative analysis of the convergence rates across three distinct algorithmic scenarios: Maximum Likelihood Estimation (MLE), Bayesian Optimization (BO), and a hybrid approach involving an initial MLE phase followed by BO. This examination, while not rooted in a theoretical framework, provides an initial foundation for more comprehensive algorithmic comparisons.

In each scenario, we tracked the number of iterations required for convergence. Our findings reveal that MLE, though superior in achieving an optimal solution compared to BO, demands a significantly higher iteration count. The hybrid approach—commencing with MLE and subsequently applying BO for 50 iterations—attained the same level of solution quality as MLE alone, but in a reduced number of iterations. An illustration is presented in figure 4

Based on these observations, we posit that a sequential application of MLE followed by BO offers a more efficient convergence pathway compared to employing either method in isolation. This insight could be pivotal for optimizing algorithmic performance in computational tasks where both solution quality and computational efficiency are paramount.

4.4. Performance enhancements

To further advance performance metrics, we explored the implementation of Parallel Bayesian Optimization (Par-

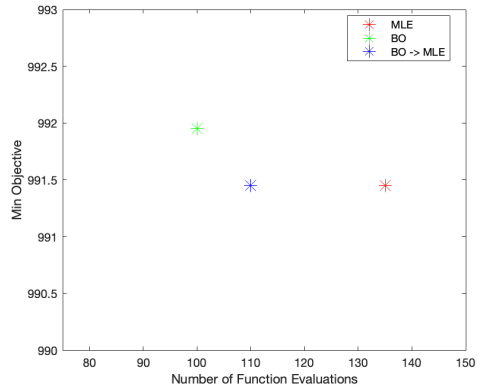

Figure 4. Bayesian Optimization as a starting point for MLE

allel BO) using an 8 core machine. This approach was benchmarked against both Maximum Likelihood Estimation (MLE) and the sequential Bayesian Optimization (BO) method. Our empirical findings indicate that Parallel BO demonstrates superior performance when compared to its MLE and sequential BO counterparts.

The efficacy of Parallel BO is notably pronounced in its acceleration of the optimization process. The empirical data showcases an approximate threefold increase in computational speed (approx 3x speedup) when employing Parallel BO. This significant enhancement in processing efficiency does not come at the cost of a substantial sacrifice in solution quality. The increment in loss, as observed, is minimal, suggesting that Parallel BO maintains a high degree of accuracy while vastly improving the time efficiency. Results can be found in figure 5

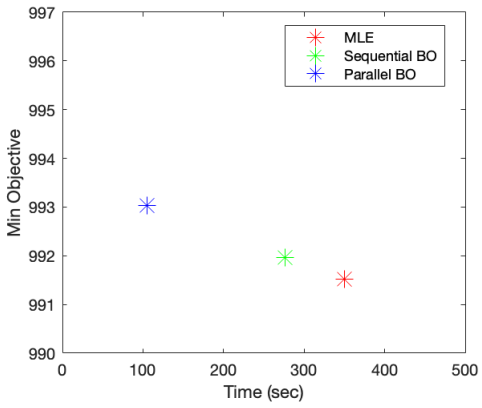

Figure 5. Parallel BO is faster than MLE and Sequential BO

5. Conclusion

In this study, we investigated the viability of implementing Bayesian Optimization as an alternative to the traditional MLE method for logit choice models. We observed that Bayesian Optimization converges more rapidly to a local minima while struggling to escape and locate the global optimum within a reasonable timeframe. However, with sufficient iterations and extensive exploration of the parameter space, its accuracy can still approach that of the MLE method. These observations led us to develop a novel hybrid method. By integrating both the accuracy of the MLE method and the convergence efficiency of Parallel Bayesian Optimization, our approach demonstrated a promising improvement compared to MLE alone with a close enough estimation while spending less time. It is noteworthy that the choice model setting in this study was relatively simple compared to models for real-world data, which often involve more complex priors and a high-dimensional parameter space. In such cases, the objective evaluation becomes computationally more expensive than Gaussian Process modeling, hence, underscoring the necessity for our hybrid approach.

Looking forward, we aim to refine and enhance Bayesian Optimization's performance in the later stages of the process, with the ultimate goal of potentially eliminating the need for employing Maximum Likelihood Estimation (MLE) as a secondary step.

References

Aguirregabiria, V. and Mira, P. Dynamic discrete choice structural models: A survey. *Journal of Econometrics*, 156(1):38–67, 2010.

Anderson, S. P., De Palma, A., and Thisse, J.-F. *Discrete choice theory of product differentiation*. MIT press, 1992.

Ben-Akiva, M. and Boccara, B. Discrete choice models with latent choice sets. *International journal of Research in Marketing*, 12(1):9–24, 1995.

Berry, S., Levinsohn, J., and Pakes, A. to econometrica. *Econometrica*, 63(4):841–890, 1995.

Berry, S. T. Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, pp. 242–262, 1994.

Bhat, C. R. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B: Methodological*, 35(7):677–693, 2001.

Cappellari, L. and Jenkins, S. P. Multivariate probit regression using simulated maximum likelihood. *The STATA journal*, 3(3):278–294, 2003.

Chintagunta, P. K. and Nair, H. S. Structural workshop paper—discrete-choice models of consumer demand in marketing. *Marketing Science*, 30(6):977–996, 2011.

Cirillo, C. and Xu, R. Dynamic discrete choice models for transportation. *Transport Reviews*, 31(4):473–494, 2011.

Couckuyt, I., Deschrijver, D., and Dhaene, T. Fast calculation of multiobjective probability of improvement and expected improvement criteria for pareto optimization. *Journal of Global Optimization*, 60:575–594, 2014.

Demaris, A. *Logit modeling: Practical applications*. Number 86. Sage, 1992.

Evenson, R. E. and Kislev, Y. A stochastic model of applied research. *Journal of Political Economy*, 84(2):265–281, 1976.

Feng, Q., Shanthikumar, J. G., and Xue, M. Consumer choice models and estimation: A review and extension. *Production and Operations Management*, 31(2):847–867, 2022.

Frazier, P. I. A tutorial on bayesian optimization. *arXiv preprint arXiv: 1807.02811*, 2018.

Garen, J. The returns to schooling: A selectivity bias approach with a continuous choice variable. *Econometrica: Journal of the Econometric Society*, pp. 1199–1218, 1984.

Greene, W. Discrete choice modeling. In *Palgrave Handbook of Econometrics: Volume 2: Applied Econometrics*, pp. 473–556. Springer, 2009.

Hausman, J. and McFadden, D. Specification tests for the multinomial logit model. *Econometrica: Journal of the econometric society*, pp. 1219–1240, 1984.

Hensher, D. A. and Greene, W. H. The mixed logit model: the state of practice. *Transportation*, 30:133–176, 2003.

Hensher, D. A. and Johnson, L. W. *Applied discrete-choice modelling*. Routledge, 2018.

Hewitt, J. A. and Hanemann, W. M. A discrete/continuous choice approach to residential water demand under block rate pricing. *Land Economics*, pp. 173–192, 1995.

Kim, J., Rasouli, S., and Timmermans, H. Hybrid choice models: principles and recent progress incorporating social influence and nonlinear utility functions. *Procedia Environmental Sciences*, 22:20–34, 2014.

Kononenko, I. Bayesian neural networks. *Biological Cybernetics*, 61(5):361–370, 1989.

Lee, L.-F. Simulated maximum likelihood estimation of dynamic discrete choice statistical models some monte carlo results. *Journal of Econometrics*, 82(1):1–35, 1997.

Lovreglio, R., Borri, D., dell'Olio, L., and Ibeas, A. A discrete choice model based on random utilities for exit choice in emergency evacuations. *Safety science*, 62: 418–426, 2014.

MacKay, D. J. et al. Introduction to gaussian processes. *NATO ASI series F computer and systems sciences*, 168: 133–166, 1998.

Mansur, E. T., Mendelsohn, R. O., and Morrison, W. A discrete-continuous choice model of climate change impacts on energy. 2005.

Mesbah, A. Stochastic model predictive control: An overview and perspectives for future research. *IEEE Control Systems Magazine*, 36(6):30–44, 2016.

Moré, J. J. and Sorensen, D. C. Newton's method. Technical report, Argonne National Lab., IL (USA), 1982.

Muthén, B. A structural probit model with latent variables. *Journal of the American Statistical Association*, 74(368): 807–811, 1979.

Nevo, A. A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of economics & management strategy*, 9(4):513–548, 2000.

Nevo, A. Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69(2):307–342, 2001.

Pelikan, M., Goldberg, D. E., Cantú-Paz, E., et al. Boa: The bayesian optimization algorithm. In *Proceedings of the genetic and evolutionary computation conference GECCO-99*, volume 1. Citeseer, 1999.

Pinsky, M. and Karlin, S. *An introduction to stochastic modeling*. Academic press, 2010.

Porell, F. W. and Adams, E. K. Hospital choice models: a review and assessment of their utility for policy impact analysis. *Medical Care Research and Review*, 52(2):158–195, 1995.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Mass., 2005. ISBN 9780262182539.

Ruder, S. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

Rust, J. Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica: Journal of the Econometric Society*, pp. 999–1033, 1987.

Rust, J. Structural estimation of markov decision processes. *Handbook of econometrics*, 4:3081–3143, 1994.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and De Freitas, N. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104 (1):148–175, 2015.

Snoek, J., Larochelle, H., and Adams, R. P. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

Wilson, A. G. *Covariance kernels for fast automatic pattern discovery and extrapolation with Gaussian processes*. PhD thesis, University of Cambridge, 2014.

Zhan, D. and Xing, H. Expected improvement for expensive optimization: a review. *Journal of Global Optimization*, 78(3):507–544, 2020.