A Benders decomposition for maximum simulated likelihood estimation of advanced discrete choice models

T. Haering[a,*], C. Bongiovanni[a] and M. Bierlaire[a]

[a] École Polytechnique Fédérale de Lausanne, Transport and mobility laboratory, Lausanne, Switzerland
tom.haering@epfl.ch, claudia.bongiovanni@epfl.ch, michel.bierlaire@epfl.ch
* Corresponding author

Extended abstract submitted for presentation at the 11th Triennial Symposium on Transportation Analysis conference (TRISTAN XI)
June 19-25, 2022, Mauritius Island

April 11, 2022

Keywords: maximum likelihood estimation, discrete choice, simulation, mixed integer linear programming, benders decomposition

1 INTRODUCTION

Maximum likelihood estimation (MLE) is a broadly used method to estimate the parameters of a previously specified distribution, given observed data. It finds its use in many areas of physics (e.g. Hauschild & Jentschel (2001)), machine learning (e.g. Goodfellow *et al.* (2016)) and discrete choice modelling (e.g. Bierlaire (2003)). Advanced discrete choice specifications like latent class or probit models are challenging to estimate, as the choice probabilities resulting from such models are highly non-convex and do typically not have a closed-form expression. For this reason, optimization approaches have relied on simulation techniques, i.e. maximum simulated likelihood estimation (MSLE), see Train (2009). A general approach for MSLE has been proposed in Fernández Antolín (2018), where the problem is formulated as a mixed integer linear program (MILP). This allows to define the model in terms of its error components, enabling the approach to be flexibly applied to any advanced discrete choice model. It is furthermore independent from the complexity of the error term distributions, the only requirement being that its possible to take draws. If the number of draws is sufficiently large, the MILP formulation is guaranteed to convergence to a globally optimal solution. However, since the complexity of the MILP scales exponentially with the number of draws, the approach can currently only be applied to solving small-scale instances (i.e., with few individuals and alternatives).

In this work, we extend the range of applicability of the MILP approach in Fernández Antolín (2018) by means of a Benders decomposition, which speeds-up the MILP solution process for the MLSE drastically and enables to scale-up the tackled instances. Our designed Benders decomposition exploits total unimodularity to keep the master problem linear, thus eliminating the bottleneck in computational time usually associated with Benders decomposition. The proposed approach is benchmarked against the full MILP and PandasBiogeme. Preliminary results on a simple logit model show that the Benders decomposition approach solves instances up to 60x times faster than the MILP, while retaining high quality solutions.

2 METHODOLOGY

In this section, we formally introduce an MILP formulation for the MSLE problem, based on the work in Fernández Antolín (2018), and a problem-specific Benders decomposition approach.

2.1 MILP formulation

$$\max_{\beta,\omega,s,z,U,H} \sum_n \sum_i y_{in} z_{in}$$

s.t.

$$\begin{aligned}
\sum_i \omega_{inr} &= 1 && (\mu_{nr}) \\
H_{nr} &= \sum_i U_{inr}\omega_{inr} && (\zeta_{nr}) \\
H_{nr} &\geq U_{inr} && (\alpha_{inr}) \\
s_{in} &= \sum_r \omega_{inr} && (\theta_{in}) \\
z_{in} &\leq L_r - K_r s_{in} && (\xi_{inr}) \\
U_{inr} &= \sum_k \beta_k x_{ink} + \epsilon_{inr} && (\kappa_{inr}) \\
\omega_{inr} &\in \{0,1\} \\
\beta, s, z, U, H &\in \mathbb{R}
\end{aligned}$$

Formulation 1 – *MSLE as an MILP*

Consider a set of $n = \{1, \dots, N\}$ individuals choosing exactly one alternative among a set of $i = \{1, \dots, I\}$ alternatives. Such choice is depicted by a binary decision variable y_{in}. Assume that each individual n selects the alternative i corresponding to the maximal utility U_{in}, i.e. $y_{in} = 1 \Leftrightarrow U_{in} = \max_j U_{jn}$. The utility function is defined in constraints (κ_{inr}) and depends on k parameters β which are to be estimated. The simulated choices for each scenario $r = \{1, \dots, R\}$ are captured by the binary variables ω_{inr}, together with constraints (μ_{nr}), limiting the choice to a single alternative. Constraints (ζ_{nr}) and (α_{inr}) guarantee that the chosen alternative corresponds to the one with the highest utility. The objective is to maximize the simulated log-likelihood, given by $\ln(\prod_n \prod_i \hat{P}_n(i)^{y_{in}})$, where $\hat{P}_n(i)$ represents the estimator for the probability of individual n choosing alternative i and is given by $\frac{1}{R}\sum_r \omega_{inr}$. Constraints (θ_{in}) and (ξ_{inr}) model a piece-wise linearization of the log-transformation, utilizing constants $L_r = (1+r)\ln(r) - r\ln(1+r)$ and $K_r = \ln(r) - \ln(1+r)$ for the intercepts and slopes.

2.2 Benders decomposition approach

Combinatorial optimization problems that are characterized by complicating integer decision variables are typically tackled by a Benders decomposition approach (Benders (1962)). As the integral master problem needs to be solved repeatedly, Benders is notorious for its slow convergence. In our case we can use an elegant trick to avoid this issue: by identifying the continuous estimation parameters β as the complicating variables and fixing them in the subproblem, the utilities of all the alternatives become fixed as well. Thus the problem of choosing the highest utility alternative simplifies to a knapsack problem, which is totally unimodular, which allows to drop the integrality constraints on the choice variables. Formulations 2 and 3 give the respective definitions of the primal and dual of the subproblem, while Formulation 4 describes the master problem.

As the linearization of constraint (ζ_{nr}) using a big-M approach no longer works when integrality constraints are relaxed, the formulation in the primal subproblem needs to be modified. The product $\eta_{inrk} = \omega_{inr}\beta_k$ is modeled directly using constraints (π_{inr}), (λ_{inrk}) and (φ^{β}_{nrk}). This formulation is equivalent to Formulation 1. It is important to mention that, in order to preserve the inegrality of the primal, information about β^{fixed} had to be kept in the matrix, which implies it also being contained in the matrix of the dual (constraints (χ_{inr})). This means the feasible region of the dual subproblem is not constant over iterations, which might distort the Bender cuts. Lastly, both the primal and the dual models are fully decomposable on the individuals n, as individuals select alternatives independently from each other and across scenarios.

2.2.1 Subproblem

$$\min_{\beta,\omega,\chi,\eta,s,z,H} -\sum_n \sum_i y_{in} z_{in}$$

s.t.

$$\begin{aligned}
\sum_i \omega_{inr} &= 1 && (\mu_{nr}) \\
\sum_k \beta_k x_{ink} - H_{nr} &\leq -\epsilon_{inr} && (\alpha_{inr}) \\
H_{nr} - \sum_{ik} \eta_{inrk} x_{ink} &\leq \sum_i \omega_{inr}\epsilon_{inr} && (\zeta_{nr}) \\
\chi_{inr} + \omega_{inr} &= 1 && (\pi_{inr}) \\
\eta_{inrk} + \beta_k^{\text{fixed}} \chi_{inr} &= \beta_k^{\text{fixed}} && (\lambda_{inrk}) \\
\beta_k - \sum_i \eta_{inrk} &= 0 && (\varphi^{\beta}_{nrk}) \\
s_{in} - \sum_r \omega_{inr} &= 0 && (\theta_{in}) \\
z_{in} + K_r s_{in} &\leq L_r && (\xi_{inr}) \\
\omega, \chi, s &\in \mathbb{R}_{\geq 0} \\
\beta, \eta, z, H &\in \mathbb{R}
\end{aligned}$$

Formulation 2 – *Primal subproblem*

$$\max_{\mu,\alpha,\zeta,\mu,\lambda,\varphi^{\beta},\theta,\xi} \sum_{nr} \mu_{nr} - \sum_{inr} \epsilon_{inr}\alpha_{inr} + \sum_{inr} \pi_{inr} + \sum_{inrk} \beta_k^{\text{fixed}} \lambda_{inrk} + \sum_{inr} L_r \xi_{inr}$$

s.t.

$$\begin{aligned}
\mu_{nr} - \zeta_{nr}\epsilon_{inr} + \pi_{inr} - \theta_{in} &\leq 0 && (\omega_{inr}) \\
\pi_{inr} + \sum_k \beta_k^{\text{fixed}} \lambda_{inrk} &\leq 0 && (\chi_{inr}) \\
-\sum_i \alpha_{inr} + \zeta_{nr} &= 0 && (H_{nr}) \\
-\zeta_{nr} x_{ink} + \lambda_{inrk} - \varphi^{\beta}_{nrk} &= 0 && (\eta_{inrk}) \\
\theta_{in} + \sum_r K_r \xi_{inr} &\leq 0 && (s_{in}) \\
\sum_r \xi_{inr} &= -y_{in} && (z_{in}) \\
\sum_{inr} \alpha_{inr} x_{ink} + \sum_{nr} \varphi^{\beta}_{nrk} &= 0 && (\beta_k) \\
\mu, \pi, \lambda, \theta, \varphi^{\beta} &\in \mathbb{R} \\
\alpha, \zeta, \xi &\in \mathbb{R}_{\leq 0}
\end{aligned}$$

Formulation 3 – *Dual subproblem*

2.2.2 Master problem

$$\min_{\mathcal{L},\beta} \mathcal{L}$$

s.t.

$$\begin{aligned}
\mathcal{L} &\geq \mathcal{L}^* + \sum_n \sum_k \phi^*_{nk}(\beta_k - \beta_k^{\text{fixed}}) && (1) \\
\mathcal{L} &\geq \mathcal{L}^{\text{best}} && (2) \\
\mathcal{L}, \beta &\in \mathbb{R}
\end{aligned}$$

Formulation 4 – *Master problem*

The master problem reduces to finding optimal values for the estimation parameters β. For each β^{fixed}, after solving the dual subproblem, a Benders cut of the same type as constraint (1) is added. Each optimal objective value serves as a new lower bound on the objective, enforced in constraint (2). The parameters ϕ^*_{nk} of the Benders cuts in our case are determined by $\phi^*_{nk} = \sum_{ir} \lambda_{inrk}$.

3 PRELIMINARY RESULTS

To test our approach, we conduct experiments on a binary logit model. A mode choice problem between two alternatives, public transport (pt) and car, is considered. The systematic utilities of the alternatives are:

$$\begin{aligned}
V_{\text{car}} &= \beta_{\text{time}} \cdot \text{traveltime}_{\text{car}} \\
V_{\text{pt}} &= \beta_{\text{time}} \cdot \text{traveltime}_{\text{pt}}
\end{aligned}$$

The dataset is extracted from revealed preference data on mode choice collected in 1987 for the Netherlands Railways, consisting of 228 respondents (CASE, 2017). Experiments are performed using GUROBI 9.5.0 (Gurobi Optimization, LLC, 2021) on a 2.6 GHz 6-Core Intel Core i7 processor with 16 GB of RAM, with a three hour time limit per instance. Our proposed Benders approach is benchmarked against PandasBiogeme (Bierlaire, 2020) and the full MILP, comparing the objective values, the parameter values, and the runtimes. Biogeme's objective function

is the log-likelihood (LL), which is approximated by the simulated log-likelihood (sLL), the MILP objective. For comparison, the LL is also evaluated using the estimated parameters from decomposition and full MILP. We take random subsets of individuals from the population to get instances that are manageable for the MILP. The results shown in Table 1 highlight the following: 1. On average, the decomposition solves the problem over 30 to 60 times faster, 2. comparing the optimal solution values for the full MILP and our decomposition reveals differences, and 3. these differences are small for the objective function and more significant for the estimated parameters. Although a Benders decomposition is an exact approach, our formulation contains mathematical

N	R	LL-Bio	LL-D	LL-M	sLL-D	sLL-M	B-Bio	B-D	B-M	T-D	T-M
20	50	-12.3029	-12.4934	-12.4443	-12.658	-12.6074	-1.5579	-0.9704	-1.0481	10.0613	64.9419
20	100	-12.3029	-12.4109	-12.3948	-12.2581	-12.2116	-1.5579	-1.1097	-1.1432	9.9023	403.6936
20	200	-12.3029	-12.4633	-12.3778	-12.6478	-12.2834	-1.5579	-2.1604	-1.1823	16.9385	1117.0644
50	50	-30.2645	-30.683	-30.3258	-31.0302	-30.8476	-1.4095	-0.9354	-1.2226	29.7795	286.6792
50	100	-30.2645	-30.481	-30.3258	-31.0396	-30.4611	-1.4095	-1.7825	-1.2226	65.0063	1558.6037
50	200	-30.2645	-30.2825	-30.3254	-30.6917	-30.5655	-1.4095	-1.3072	-1.2232	98.2058	5375.6553
100	50	-64.8827	-65.3962	-64.8978	-65.8014	-65.2044	-0.9481	-0.6117	-0.889	28.7805	2820.2287
100	100	-64.8827	-66.0306	-64.8828	-67.419	-65.7837	-0.9481	-0.4513	-0.9431	274.1626	4346.0671
100	200	-64.8827	-64.9245	-64.8933	-66.0181	-65.6991	-0.9481	-0.8502	-0.8987	295.7408	10800+
200	50	-122.6885	-122.6895	-122.7352	-124.0274	-123.5507	-1.31	-1.322	-1.3901	120.5793	1476.1851
200	100	-122.6885	-122.7386	-122.92	-124.2428	-124.0001	-1.31	-1.3929	-1.4898	327.2528	10800+
200	200	-122.6885	-122.7213	-123.3417	-124.1058	-124.7073	-1.31	-1.377	-1.0213	1262.7548	10800+

Table 1 – *Comparing our decomposition method with the full MILP and PandasBiogeme (N = population size, R = number of draws, LL = log-likelihood , sLL = simulated log-likelihood, $B = \beta_{time}$, T = time (sec.), Bio = Biogeme, D = decomposition, M = MILP)*

aspects that may currently prevent the convergence to the real global optimum. As mentioned in the methodology, a possible explanation for the deviations is the fact that information about the master variables is maintained in the matrix of the dual. Other explanations include numerical issues, stemming for example from the linearization of the logarithm or the way certain solvers handle specific constraints.

4 FUTURE WORK

This is ongoing work. We are currently investigating the cause of the deviations between the optimal values from the decomposition and the full MILP, as well as testing extensions and performance on various advanced discrete choice models. Additional results obtained by the time of the conference will of course be included in the presentation.

References

Benders, Jacques F. 1962. Partitioning procedures for solving mixed-variables programming problems. *Numerische mathematik*, **4**(1), 238–252.

Bierlaire, Michel. 2003. BIOGEME: A free package for the estimation of discrete choice models. *In: Swiss transport research conference*.

Bierlaire, Michel. 2020. A short introduction to PandasBiogeme. *A short introduction to PandasBiogeme*.

CASE, NETHERLANDS MODE CHOICE. 2017. Data collection.

Fernández Antolín, Anna. 2018. *Dealing with Correlations in Discrete Choice Models*. Tech. rept. EPFL.

Goodfellow, Ian, Bengio, Yoshua, & Courville, Aaron. 2016. Machine learning basics. *Deep learning*, **1**(7), 98–164.

Gurobi Optimization, LLC. 2021. *Gurobi Optimizer Reference Manual*.

Hauschild, T, & Jentschel, M. 2001. Comparison of maximum likelihood estimation and chi-square statistics applied to counting experiments. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, **457**(1-2), 384–401.

Train, Kenneth E. 2009. *Discrete choice methods with simulation*. Cambridge university press.