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

T. Haering*[1], C. Bongiovanni[2], and M. Bierlaire[3]

[1]Transport and mobility laboratory, École Polytechnique Fédérale de Lausanne, Switzerland

[2]Ph.D., Transport and mobility laboratory, École Polytechnique Fédérale de Lausanne, Switzerland

[3]Prof., Transport and mobility laboratory, École Polytechnique Fédérale de Lausanne, Switzerland

SHORT SUMMARY

In this paper, we formulate a mixed integer linear program (MILP) for the simulated maximum likelihood estimation (MLSE) problem and devise a Benders decomposition approach to speed up the solution process. This framework can be applied to any advanced discrete choice model and exploits total unimodularity to keep the master problem linear in the decomposition. The proposed decomposition approach is benchmarked against the original MILP formulation and PandasBiogeme. Computational experiments are performed on a binary logit mode choice model with up to 200 respondents. Results show that the Benders decomposition approach solves instances on average 35 and up to 100 times faster than the MILP while maintaining high quality solutions.

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

1. INTRODUCTION

Maximum likelihood estimation (MLE) is a broadly used method to estimate the parameters of a probability distribution, given observed data (Myung, 2003). It finds its use in many areas of mathematical statistics (Sur & Candès, 2019), physics (Hauschild & Jentschel, 2001), machine learning (Goodfellow, Bengio, & Courville, 2016) and discrete choice modeling (Bierlaire, 2003). The latter specifically relies on the use of MLE to estimate the optimal parameters of convex and non-convex discrete choice models (Bierlaire, 1998). This estimation process is challenging, especially for more advanced discrete choice models, e.g. latent class or probit models, because of nonconvex and nonlinear mathematical properties. Given that the choice probabilities resulting from such models do not have a closed-form expression, optimization approaches have typically relied on simulation techniques, i.e. maximum simulated likelihood estimation (MLSE, see Train, 2009). A general implementation approach for MSLE has been proposed in Fernández Antolín (2018), where the problem is formulated as a mixed integer linear program (MILP). The approach relaxes any assumption on the specific shape of the error term distribution and instead only assumes that it is possible to take draws. This allows it to be flexibly applied to any advanced discrete choice model. With a sufficiently large number of draws, the MILP formulation guarantees convergence to global optimal solutions. However, since the complexity of 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 (Paneque, Bierlaire, Gendron, & Azadeh, 2021).

In this work, we extend the MILP approach in Fernández Antolín (2018) by means of a Benders' decomposition approach (Rahmaniani, Crainic, Gendreau, & Rei, 2017), which speeds-up the MILP solution process for the MLSE and enables to scale-up the tackled instances. Our designed Benders' decomposition approach 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 (Bierlaire, 2020). The decomposition method is demonstrated on a binary logit discrete choice model, together with an analysis of the results.

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. Without loss of generality, the formulation is presented in the context of a multinomial logit formulation, with examples on how to extend it to other model classes, such as probit and latent class models.

MILP formulation

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

s.t.

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

Formulation 1 – MSLE as an MILP

Consider a set of $n = \{1,\ldots,N\}$ individuals choosing exactly one alternative among a set of $i = \{1,\ldots,I\}$ 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 depends on k parameters β which are to be estimated. The objective is to maximize the likelihood function, given by $\prod_n \prod_i P_n(i)^{y_{in}}$, where $P_n(i)$ represents the probability of individual n choosing alternative i. In order to linearize the likelihood function, a set of $r = \{1,\ldots,R\}$ independent scenarios are created by sampling the error term distribution, i.e. ε_{inr}. Denote by ω_{inr} the binary decision variable that indicates whether individual n chooses alternative i in scenario r. In this case, the choice probabilities are approximated by $P_n(i) \approx \frac{1}{R}\sum_r \omega_{inr}$ and are guaranteed to converge to the real probabilities with a sufficiently large number of scenarios R (Paneque et al., 2021). Taking the log of the likelihood and replacing $P_n(i)$ by its estimator yields an objective that still contains the nonlinear term $\ln(\sum_r \omega_{inr})$. This issue is tackled by introducing the auxiliary decision variable $s_{in} = \sum_r \omega_{inr}$, which is defined in constraints (θ_{in}). Similarly, an auxiliary variable z_{in} is introduced to represent the piece-wise linearization of the logarithm. The latter is defined in Constraints (ξ_{inr}), where $L_r = (1+r)\ln(r) - r\ln(1+r)$ and $K_r = \ln(r) - \ln(1+r)$ are constants representing intercepts and slopes used for the linearization. With such pre-processing steps and by ignoring the constant term $-N\ln(R)$, the objective of the problem can be rewritten as stated in Formulation 1. The rest of the constraints model individual choices. Constraints (μ_{nr}) guarantee that only one alternative can be chosen per individual and scenario. Constraints (κ_{inr}) model the utility of each alternative i for individual n in scenario r, i.e. U_{inr}. Constraints (ζ_{nr}), which can be easily linearized using a standard big-M approach, and constraints (α_{inr}) ensure that the choice being made corresponds to the one with the highest utility. Note that Formulation 1 is characterized by the complicating binary decision variables ω.

The MILP formulation can be easily adapted to other model speficiations. For example, in order to tackle a probit model, it is sufficient to add the cholesky factor of the covariance matrix (Dow & Endersby, 2004) in the right-hand side of constraints (κ_{inr}). Note that this transformation increases the number of parameters to be estimated by $\frac{I(I+1)}{2}$. Similarly, in order to tackle a latent class model, it is sufficient to add a class membership indicator γ_{cn} where c is the class index. For each class c, the constraints corresponding to making the best choice ω_{cinr} for that class are duplicated, and finally a global choice variable is defined as $\omega_{inr} = \sum_c \omega_{cinr}\gamma_{cn}$.

Benders decomposition approach

Combinatorial optimization problems that are characterized by complicating variables are typically tackled by a Benders decomposition approach (Benders, 1962). It can be summarized as an iterative solution procedure in which the complicating variables are isolated into a master problem and their solution values are fed to a subproblem, whose dual information is used to create cuts for the master problem in order to create new solutions. For a comprehensive review of the method, see for example (Rahmaniani et al., 2017). In most applications, the complicating variables are integral, resulting in the need to solve an integral master problem at each iteration. This makes Benders 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. This mathematical property allows us to drop the integrality constraints on the choice variables. Formulation 2 and Formulation 3 give the respective definitions of the primal and dual of the subproblem, while Formulation 4 describes the master problem.

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

s.t.

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

Formulation 2 – Primal subproblem

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

s.t.

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

Formulation 3 – Dual subproblem

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 is slighlty modified: The product $\eta_{inrk} = \omega_{inr}\beta_k$ is modeled directly using Constraints (π_{inr}), (λ_{inrk}) and (φ_{nrk}^{β}). 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 its coefficient matrix, which implies it also being contained in the matrix of the dual, i.e. 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.

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

s.t.

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

Formulation 4 – Master problem

Finally, 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. The parameters of the Benders cuts are determined by the achieved objective $\mathscr{L}^*$ and $\phi^*_{nk} = \sum_{ir} \lambda^*_{inrk}$. Each optimal objective value of the master problem serves as a new lower bound on the objective, enforced in Constraint (2).

3. RESULTS AND DISCUSSION

Our approach is tested 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, in terms of objective values and runtimes. Biogeme's objective function is the Log-Likelihood ($LL = \ln(\prod_n \prod_i P_n(i)^{y_{in}})$) , which is approximated by the simulated Log-Likelihood (sLL), the MILP objective. For the purpose of comparison, the LL is also evaluated for the decomposition and the MILP using the estimated parameters. We take random subsets of individuals from the population to get instances that are manageable for the MILP.

Table 1 shows the comparison between the decomposition and the full MILP in terms of sLL and computation times, while Table 2 shows the results in terms of LL. We highlight the following: 1. the decomposition solves the problem on average 35 and up to 100 times faster, 2. comparing the optimal solution values for the full MILP and our decomposition reveals small gaps in optimality, and 3. increasing the number of draws reduces the optimality gap between the exact solution (PandasBiogeme) and the approximation (MILP and decomposition).

Although Benders decomposition is an exact approach, our formulation contains mathematical 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 coefficient 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.

Table 1 – Comparing our decomposition method with the full MILP in terms of *sLL* and runtime (N = population size, R = number of draws, sLL = simulated Log-Likelihood, M = MILP, D = decomposition, T = time in sec.)

N	R	sLL-M	sLL-D	Gap [%]	T-M	T-D
20	50	-12.607	-12.658	-0.40	64.942	10.061
20	100	-12.212	-12.258	-0.38	403.694	9.902
20	200	-12.283	-12.648	-2.97	1117.064	16.939
50	50	-30.848	-31.030	-0.59	286.679	29.780
50	100	-30.461	-31.040	-1.90	1558.604	65.006
50	200	-30.566	-30.692	-0.41	5375.655	98.206
100	50	-65.204	-65.801	-0.92	2820.229	28.781
100	100	-65.784	-67.419	-2.49	4346.067	274.163
100	200	-65.699	-66.018	-0.49	10800+	295.741
200	50	-123.551	-124.027	-0.39	1476.185	120.579
200	100	-124.000	-124.243	-0.20	10800+	327.253
200	200	-124.707	-124.106	0.48	10800+	1262.755

Table 2 – Comparing our decomposition method with the full MILP and Pandas-Biogeme in terms of *LL* (N = population size, R = number of draws, LL = Log-Likelihood, Biog = PandasBiogeme, M = MILP, D = decomposition)

N	R	LL-Biog	LL-M	Gap [%]	LL-D	Gap [%]
20	50	-12.303	-12.444	-1.15	-12.493	-1.55
20	100	-12.303	-12.395	-0.75	-12.411	-0.88
20	200	-12.303	-12.378	-0.61	-12.463	-1.30
50	50	-30.265	-30.326	-0.20	-30.683	-1.38
50	100	-30.265	-30.326	-0.20	-30.481	-0.72
50	200	-30.265	-30.325	-0.20	-30.283	-0.06
100	50	-64.883	-64.898	-0.02	-65.396	-0.79
100	100	-64.883	-64.883	0.00	-66.031	-1.77
100	200	-64.883	-64.893	-0.02	-64.925	-0.06
200	50	-122.689	-122.735	-0.04	-122.690	0.00
200	100	-122.689	-122.920	-0.19	-122.739	-0.04
200	200	-122.689	-123.342	-0.53	-122.721	-0.03

4. CONCLUSIONS

In this paper, we develop a mixed integer linear program (MILP) for the simulated maximum likelihood estimation (MLSE) problem and construct a Benders decomposition approach to speed up the solution process. The methodology can be applied to any advanced discrete choice model and makes use of total unimodularity to keep the master problem linear in the decomposition, avoiding the typical bottleneck in efficiency for a Benders decomposition. The results on a binary logit discrete choice model show an average speed up of factor 35, with instances being solved up to 100 times faster. Small deviations in the optimal solution values between decomposition and full MILP are present. This is currently under investigation, together with applications to more advanced discrete choice models.

ACKNOWLEDGMENT

We would especially like to thank Dr. Stefano Bortolomiol for many enlightening discussions on Benders decompositions, as well as Prof. Virginie Lurkin and Dr. Rico Krueger for their highly constructive inputs and general perspective.

REFERENCES

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

Bierlaire, M. (1998). Discrete choice models. In *Operations research and decision aid methodologies in traffic and transportation management* (pp. 203–227). Springer.

Bierlaire, M. (2003). Biogeme: A free package for the estimation of discrete choice models. In *Swiss transport research conference.*

Bierlaire, M. (2020). A short introduction to pandasbiogeme. *A short introduction to PandasBiogeme*.

CASE, N. M. C. (2017). Data collection.

Dow, J. K., & Endersby, J. W. (2004). Multinomial probit and multinomial logit: a comparison of choice models for voting research. *Electoral studies*, *23*(1), 107–122.

Fernández Antolín, A. (2018). *Dealing with correlations in discrete choice models* (Tech. Rep.). EPFL.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Machine learning basics. *Deep learning*, *1*(7), 98–164.

Gurobi Optimization, LLC. (2021). *Gurobi Optimizer Reference Manual.* Retrieved from https://www.gurobi.com

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.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, *47*(1), 90–100.

Paneque, M. P., Bierlaire, M., Gendron, B., & Azadeh, S. S. (2021). Integrating advanced discrete choice models in mixed integer linear optimization. *Transportation Research Part B: Methodological*, *146*, 26–49.

Rahmaniani, R., Crainic, T. G., Gendreau, M., & Rei, W. (2017). The benders decomposition algorithm: A literature review. *European Journal of Operational Research*, *259*(3), 801–817.

Sur, P., & Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, *116*(29), 14516–14525.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.