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

hEART 2022
Tom Häring, Claudia Bongiovanni, Michel Bierlaire
TRANSP-OR Laboratory, EPFL
Contents

1. Why maximum likelihood estimation (MLE)?
2. Why simulated MLE?
3. Why a mixed integer linear program (MILP)?
4. Simulated MLE as an MILP
5. Why decomposition?
6. The Benders decomposition
7. Results
8. Ideas for future work
Why maximum likelihood estimation (MLE)?

• **MLE** is for example used to estimate the parameters of **discrete choice models**
Why maximum likelihood estimation (MLE)?

• For each individual $n$, every alternative $i$ has an associated utility:

$$U_{in} = U_{in}(\beta, x, \epsilon_{in})$$

  - **parameters to be estimated**
  - **random error term**
  - **exogenous attributes**

• Assumptions:
  1.) **linear** in parameters
  2.) we can **draw** from error terms
Why maximum likelihood estimation (MLE)?

• For each individual \( n \), every alternative \( i \) has an associated utility:

\[
U_{in} = \sum_k \beta_k x_{in} + \epsilon_{in} = V_{in} + \epsilon_{in}
\]

• Behavioral assumption: the individual chooses the alternative with the highest utility
Why maximum likelihood estimation (MLE)?

• Data: **observed choices** \( y_{in} \) (= 1 if ind. \( n \) chose alternative \( i \), else = 0)
• Find parameters \( \beta_k \) such that the **likelihood** of this outcome is **maximized**
• Log-Likelihood function:

\[
\ln \left( \prod_{n} \prod_{i} P_{n}(i)^{y_{in}} \right) = \sum_{n} \sum_{i} y_{in} \ln P_{n}(i)
\]

where

\[
P_{n}(i) = \mathbb{P}(V_{in} + \epsilon_{in} \geq V_{jn} + \epsilon_{jn} \ \forall \ j \in J)
\]
Why simulated MLE?

• DCMs model choices realistically [1], but in general lead to non-convex probabilities [2]
  ▶️ No global optimality certificates, danger of local optima
  ▶️ Non-convex solver ≈ Blackbox

• Simulation mitigates this by giving a linear approximation [3] and allows DCMs to be easily integrated in optimization models [2]

Why simulated MLE?

• How:

  - **Simulate** $R$ scenarios, utilities become **deterministic**:
    
    $$U_{inr} = V_{in} + \epsilon_{inr}$$

  - Let $\omega_{inr}$ be the **choice variables**

  - **Approximated** probabilities:
    $$\hat{P}_n(i) = \frac{1}{R} \sum_{r=0}^{R-1} \omega_{inr}$$

*Meritxell Pacheco: A general framework for the integration of complex choice models into mixed integer optimization (2020)*
Why a mixed integer linear program (MILP)?

• Allow inclusion of **integer variables** in estimation procedure
  ➢ Model **advanced** DCMs, e. g. **latent variables / classes**
  ➢ Additional features, e. g. **automatic / assisted specification**

• Vast literature on efficient **modeling & performance**

• Gives **control** over **optimization process**: information on **bounds**, **optimality gaps**, **user-generated cuts**, etc.
Simulated MLE as an MILP

- **Objective**: \( \text{max } \log\text{-Likelihood} \)

\[
\sum_{n} \sum_{i} y_{in} \ln P_n(i)
\]

- **max sim. Log-Likelihood**

\[
\sum_{in} y_{in} \ln \left( \sum_{r=0}^{R-1} \omega_{inr} - y_{in} \ln R \right)
\]

\[
S_{in} = \sum_{r} \omega_{inr}
\]

\[
z_{in} \leq L_r - K_r S_{in}
\]

\[
\text{max } \sum_{n} \sum_{i} Y_{in} z_{in}
\]

Simulated MLE as an MILP

• Constraints:

\[ \sum_{i} \omega_{inr} = 1 \quad \forall n, r \]

\[ U_{inr} = \sum_{k} \beta_{k} x_{ink} + \epsilon_{inr} \quad \forall i, n, r \]

\[ U_{nr} \geq U_{inr} \quad \forall i, n, r \]

\[ U_{nr} = \sum_{i} U_{inr} \omega_{inr} \quad \forall n, r \]

\[ s_{in} = \sum_{r} \omega_{inr} \quad \forall i, n \]

\[ z_{in} \leq L_{r} - K_{r} s_{in} \quad \forall i, n \]

\[ \omega_{inr} \in \{0, 1\} \]

\[ \beta, s, z, U, U \in \mathbb{R} \]
Why decomposition?

- Problem: Simulation increases problem size by solving many scenarios. Only small instances can be solved in reasonable time [1].
- To solve large MILPs efficiently we consider decomposition methods.

The Benders decomposition

Master Problem (LP)
Compute candidate solution for parameters $\beta$

Sub-Problem (LP)
Totally unimodular when $\beta$ is fixed.
$\Rightarrow$ Solve dual

candidate solution $\beta$

optimality cuts
The Benders decomposition

• For a fixed $\beta_k$ the rest of the MILP becomes a **Knapsack-problem**

$=>$ totally unimodular:

- Utilities become fixed
  $$U_{inr} = \sum_k \beta_{k}^{\text{fixed}} x_{nk} + \epsilon_{inr}$$

- Now:
  $$U_{nr} = \sum_i U_{inr} \omega_{inr}$$
  $$U_{nr} \geq U_{inr}$$
  $$\sum_i \omega_{inr} = 1$$
  $$\omega_{i*nr} = 1$$
  for the alternative $i*$
  with highest utility
The Benders decomposition

• Typically:
  ▪ The variable to be fixed is \textit{integer}, so that the subproblems are linear
  ▪ Thus \textit{MP is an integer program (bottleneck!)}

• But in our case:
  ▪ The variable to be fixed is \textit{continuous}, but thanks to TU-ness the subproblems are (\textit{technically}) still linear!
  ▪ Thus \textit{SP is a linear program}

From solving an MILP to iteratively solving LP’s!
The Benders decomposition

• Difficulty:

Simply adding the constraint \( \beta_k = \beta^\text{fixed}_k \) does not work in our case because of the non-linearity of the problem.
The Benders decomposition

• **Constraints:**

\[
\sum_i \omega_{inr} = 1 \quad \forall n, r
\]

\[
U_{inr} = \sum_k \beta_k x_{ink} + \epsilon_{inr} \quad \forall i, n, r
\]

\[
U_{nr} \geq U_{inr} \quad \forall i, n, r
\]

\[
U_{nr} = \sum_i U_{inr} \omega_{inr} \quad \forall n, r
\]

\[
s_{in} = \sum_r \omega_{inr} \quad \forall i, n
\]

\[
z_{in} \leq L_r - K_r s_{in} \quad \forall i, n
\]

\[
\omega_{inr} \in [0, 1]
\]

\[
\beta, s, z, U, U \in \mathbb{R}
\]

Goal: linear in $\beta_k$
The Benders decomposition

- We design a *quasi*-linearization:

\[
\eta_{inr} = \beta^\text{fixed}_k \omega_{inr} \quad \Rightarrow \quad \eta_{inr} + \beta^\text{fixed}_k \chi_{inr} = 1
\]

\[
\sum_i \eta_{inr} = \beta_k
\]
Application to a mode choice problem

• Dataset: RP data on mode choice, Netherlands, 1987
• Simple binary logit model:
  choice between two modes – car and rail

\[
U_{\text{car},n} = \beta_{\text{time}} \times \text{traveltime}_{\text{car}}
\]

\[
U_{\text{rail},n} = \beta_{\text{time}} \times \text{traveltime}_{\text{rail}}
\]

• Compare decomposition vs. undecomposed MILP
<table>
<thead>
<tr>
<th>N</th>
<th>R</th>
<th>sLL-M</th>
<th>sLL-D</th>
<th>Gap [%]</th>
<th>T-M</th>
<th>T-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>50</td>
<td>-12.607</td>
<td>-12.658</td>
<td>-0.40</td>
<td>64.942</td>
<td>10.061</td>
</tr>
<tr>
<td>20</td>
<td>100</td>
<td>-12.212</td>
<td>-12.258</td>
<td>-0.38</td>
<td>403.694</td>
<td>9.902</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>-30.848</td>
<td>-31.030</td>
<td>-0.59</td>
<td>286.679</td>
<td>29.780</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>-30.461</td>
<td>-31.040</td>
<td>-1.90</td>
<td>1558.604</td>
<td>65.006</td>
</tr>
<tr>
<td>50</td>
<td>200</td>
<td>-30.566</td>
<td>-30.692</td>
<td>-0.41</td>
<td>5375.655</td>
<td>98.206</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>-65.204</td>
<td>-65.801</td>
<td>-0.92</td>
<td>2820.229</td>
<td>28.781</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>-65.784</td>
<td>-67.419</td>
<td>-2.49</td>
<td>4346.067</td>
<td>274.163</td>
</tr>
<tr>
<td>100</td>
<td>200</td>
<td>-65.699</td>
<td>-66.018</td>
<td>-0.49</td>
<td>10800+</td>
<td>295.741</td>
</tr>
<tr>
<td>200</td>
<td>50</td>
<td>-123.551</td>
<td>-124.027</td>
<td>-0.39</td>
<td>1476.185</td>
<td>120.579</td>
</tr>
<tr>
<td>200</td>
<td>100</td>
<td>-124.000</td>
<td>-124.243</td>
<td>-0.20</td>
<td>10800+</td>
<td>327.253</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>-124.707</td>
<td>-124.106</td>
<td>0.48</td>
<td>10800+</td>
<td>1262.755</td>
</tr>
</tbody>
</table>
Application to a mode choice problem

• First **conjecture**: gaps are caused by **log-linearization** in MSLE
• **Remedy**: apply decomposition to *continuous pricing problem (CPP)*

   Almost **equivalent** problem structure, **no log-linearization**
Application to a continuous pricing problem

- Continuous pricing problem:

\[
\begin{align*}
\max_{p, \omega, U, H} & \quad \sum_n \sum_r \sum_i \frac{1}{R} \theta_{in} p_i \omega_{inr} \\
\text{s.t.} & \quad \sum_i \omega_{inr} = 1 \quad \forall n, r \\
& \quad H_{nr} = \sum_i U_{inr} \omega_{inr} \quad \forall n, r \\
& \quad H_{nr} \geq U_{inr} \quad \forall i, n, r \\
& \quad U_{inr} = \sum_{k \neq l} \beta_k x_{ink} + \beta_l p_i + \varepsilon_{inr} \quad \forall i, n, r \\
& \quad \omega \in \{0, 1\} \\
p, U, H & \in \mathbb{R}
\end{align*}
\]
### Application to a Continuous Pricing Problem

<table>
<thead>
<tr>
<th>N</th>
<th>R</th>
<th>obj-MILP</th>
<th>obj-D</th>
<th>Gap [%]</th>
<th>P-MILP</th>
<th>P-D</th>
<th>Gap [%]</th>
<th>T-MILP</th>
<th>T-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>100</td>
<td>202.642</td>
<td>201.712</td>
<td>0.46</td>
<td>28.302</td>
<td>26.576</td>
<td>6.1</td>
<td>37</td>
<td>21</td>
</tr>
<tr>
<td>20</td>
<td>200</td>
<td>200.901</td>
<td>200.185</td>
<td>0.36</td>
<td>30.03</td>
<td>28.721</td>
<td>4.36</td>
<td>205</td>
<td>49</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>440.686</td>
<td>437.243</td>
<td>0.78</td>
<td>28.579</td>
<td>29.989</td>
<td>-4.94</td>
<td>55</td>
<td>27</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>431.088</td>
<td>426.669</td>
<td>1.03</td>
<td>28.99</td>
<td>27.778</td>
<td>4.18</td>
<td>241</td>
<td>62</td>
</tr>
<tr>
<td>50</td>
<td>200</td>
<td>429.605</td>
<td>429.108</td>
<td>0.12</td>
<td>28.574</td>
<td>28.655</td>
<td>-0.28</td>
<td>1022</td>
<td>163</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>990.026</td>
<td>988.732</td>
<td>0.13</td>
<td>29.118</td>
<td>28.944</td>
<td>0.6</td>
<td>252</td>
<td>31</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>977.606</td>
<td>976.149</td>
<td>0.15</td>
<td>30.099</td>
<td>29.925</td>
<td>0.58</td>
<td>1224</td>
<td>69</td>
</tr>
<tr>
<td>100</td>
<td>200</td>
<td>978.589</td>
<td>976.932</td>
<td>0.17</td>
<td>30.106</td>
<td>30.185</td>
<td>-0.26</td>
<td>3039</td>
<td>304</td>
</tr>
<tr>
<td>200</td>
<td>50</td>
<td>1906.696</td>
<td>1904.189</td>
<td>0.13</td>
<td>28.977</td>
<td>28.678</td>
<td>1.03</td>
<td>1144</td>
<td>65</td>
</tr>
<tr>
<td>200</td>
<td>100</td>
<td>1882.793</td>
<td>1877.641</td>
<td>0.27</td>
<td>29.277</td>
<td>30.052</td>
<td>-2.65</td>
<td>4104</td>
<td>359</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>1873.964</td>
<td>1871.614</td>
<td>0.13</td>
<td>29.276</td>
<td>29.343</td>
<td>-0.23</td>
<td>10811</td>
<td>690</td>
</tr>
</tbody>
</table>
Large number of draws (MSLE)

<table>
<thead>
<tr>
<th>N</th>
<th>R</th>
<th>sLL-M</th>
<th>sLL-D</th>
<th>Gap [%]</th>
<th>T-M</th>
<th>T-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>20</td>
<td>-29.417</td>
<td>-29.908</td>
<td>1.67</td>
<td>22</td>
<td>6</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>-29.294</td>
<td>-31.173</td>
<td>6.41</td>
<td>279</td>
<td>26</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>-28.885</td>
<td>-29.42</td>
<td>1.85</td>
<td>1375</td>
<td>42</td>
</tr>
<tr>
<td>50</td>
<td>150</td>
<td>-29.973</td>
<td>-30.092</td>
<td>0.4</td>
<td>2852</td>
<td>70</td>
</tr>
<tr>
<td>50</td>
<td>200</td>
<td>-30.091</td>
<td>-30.101</td>
<td>0.03</td>
<td>10800</td>
<td>131</td>
</tr>
<tr>
<td>50</td>
<td>250</td>
<td>-30.741</td>
<td>-30.775</td>
<td>0.11</td>
<td>10800</td>
<td>156</td>
</tr>
<tr>
<td>50</td>
<td>300</td>
<td>-30.837</td>
<td>-30.843</td>
<td>0.02</td>
<td>10800</td>
<td>133</td>
</tr>
<tr>
<td>50</td>
<td>400</td>
<td>-30.632</td>
<td>-30.638</td>
<td>0.02</td>
<td>10800</td>
<td>130</td>
</tr>
<tr>
<td>50</td>
<td>600</td>
<td>-30.479</td>
<td>-30.51</td>
<td>0.1</td>
<td>10800</td>
<td>289</td>
</tr>
<tr>
<td>50</td>
<td>800</td>
<td>-32.035</td>
<td></td>
<td></td>
<td>10800</td>
<td>319</td>
</tr>
<tr>
<td>50</td>
<td>1000</td>
<td></td>
<td>-30.523</td>
<td></td>
<td>10800</td>
<td>349</td>
</tr>
</tbody>
</table>
Ideas for future work

• Improving Benders:
  ➢ Piece-wise linearization
  ➢ Convex-quadratic formulation

• Column generation methods

• Combined column generation + Benders approach
Thanks!