A New Efficient Monte Carlo Algorithm for Estimating Mixed Logit Models

Fabian Bastin* Cinzia Cirillo† Philippe L. Toint†

*Research Fellow of the National Fund for Scientific Research (FNRS)
Department of Mathematics, University of Namur
5000 Namur, Belgium
fbas@math.fundp.ac.be

†Transportation Research Group
Department of Mathematics, University of Namur
5000 Namur, Belgium
{ccir,pht}@math.fundp.ac.be

1 Introduction

Discrete choice modelling is a powerful technique for describing how individuals perform a selection amongst a finite set of alternatives; in particular, the multinomial logit and its extensions are widely used, but the more powerful mixed logit modelling is gaining popularity among practioners and researchers. However, since the inherent choice probabilities are multidimensional integrals, the numerical cost associated with the evaluation of mixed logit models is significant, even with Monte Carlo approximations. As a consequence, several researchers proposed to use cheaper quasi-Monte Carlo approaches: Bhat (2001) and Train (1999) for instance advocate using Halton sequences for mixed logit models This trend is not without drawbacks: Bhat (2001) has pointed out that the coverage of the integration domain by Halton sequences rapidly deteriorates for high integration dimensions and consequently has proposed a heuristic based on the use of randomized scrambled Halton sequences; Hess *et al.* (Submitted) have also proposed the use of randomly shifted uniform vectors. By contrast, the dimensionality problem is irrelevant in pure Monte Carlo methods, which also benefit from a credible theory for the convergence of the calibration process, as well as of stronger statistical foundations (see for instance Rubinstein and Shapiro (1993), Shapiro (2000, 2003) for application to stochastic programming), in particular concerning statistical inference on the optimal value. This led us to reinvestigate pure Monte Carlo methods for mixed logit estimation, and to propose a new algorithm for stochastic programming using Monte Carlo methods, based on the trust-region technique, that allows the use of small subsets of an initially generated set of random draws, when approximating the objective far from the solution. This technique results in an algorithm that is numerically competitive with existing tools for mixed logit models, while giving more information to the practitioner.

2 The mixed logit problem

Let I be the population size and $\mathcal{A}(i)$ the set of available alternatives for individual i, $i = 1, \ldots, I$. For each individual i, each alternative A_j, $j = 1, \ldots, |\mathcal{A}(i)|$, has an associated utility which is assumed to have the form

$$U_{ij} = V_{ij} + \boldsymbol{\epsilon}_{ij}, \tag{1}$$

where $V_{ij} = V_{ij}(\beta_j, x_{ij})$ is a function of a vector of model parameters β_j and of x_{ij}, the observed attributes of alternative A_j, while $\boldsymbol{\epsilon}_{ij}$ is a random term reflecting the unobserved part of the utility. Without loss of generality, β_j may be assumed constant across alternatives (i.e. $\beta_j = \beta$ for all j). The theory assumes that individual i selects the alternative that maximizes his/her utility. If the random terms $\boldsymbol{\epsilon}_{ij}$ are independently Gumbel distributed with mean 0 and scale factor 1.0, the probability that the individual i chooses alternative j can be expressed with the logit formula

$$L_{ij}(\beta) = \frac{e^{V_{ij}(\beta)}}{\sum_{l=1}^{|\mathcal{A}(i)|} e^{V_{il}(\beta)}}, \tag{2}$$

where we have simplified our notation by dropping the explicit mention of the dependence of L_{ij} and V_{ij} on x_{ij}. Formula (2) characterizes the classical multinomial logit model.

Mixed logit models relax the assumption that the parameters β are the same for all individuals, by assuming instead that individual parameters vectors $\beta(i)$, $i = 1, \ldots, I$, are realizations of a random vector $\boldsymbol{\beta}$, that is itself derived from a random vector $\boldsymbol{\gamma}$ and a parameters vector θ, which we express $\boldsymbol{\beta} = \beta(\boldsymbol{\gamma}, \theta)$. For example, if $\boldsymbol{\beta}$ is a K-dimensional normally distributed random vector, we may choose $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_K)$, with $\gamma_k \sim N(0,1)$[1], and let θ specify the means and standard deviations of the components of $\boldsymbol{\beta}$. The probability choice is then

$$P_{ij}(\theta) = E_P\left[L_{ij}\left(\boldsymbol{\gamma}, \theta\right)\right] = \int L_{ij}(\gamma, \theta) P(d\gamma) = \int L_{ij}(\gamma, \theta) f(\gamma) d\gamma, \tag{3}$$

where P is the probability measure associated with γ and $f(\cdot)$ is its distribution function.

The vector of parameters θ is then estimated by maximizing the log-likelihood function:

$$\max_{\theta} LL(\theta) = \max_{\theta} \frac{1}{I} \sum_{i=1}^{I} \ln P_{ij_i}(\theta), \tag{4}$$

where j_i is the alternative choice made by the individual i. In order to face to integrals evaluation, the value of $P_{ij_i}(\theta)$ is replaced by a Monte Carlo estimate obtained by sampling over γ, and given by

$$SP^R_{ij_i}(\theta) = \frac{1}{R} \sum_{r_i=1}^{R} L_{ij_i}(\gamma_{r_i}, \theta),$$

where R is the number of random draws γ_{r_i}, taken from the distribution function of $\boldsymbol{\gamma}$. As a result, θ is now computed as the solution of the simulated log-likelihood problem

$$\max_{\theta} SLL^R(\theta) = \max_{\theta} \frac{1}{I} \sum_{i=1}^{I} \ln SP^R_{ij_i}(\theta). \tag{5}$$

[1] $N(\mu, \sigma)$ stands for the normal distribution with mean μ and standard deviation σ.

We will denote by θ^* a solution of the true problem (4) and by θ_R^* a solution of this last approximate problem, that we call the Sample Average Approximation, or SAA, by analogy to stochastic programming. This analogy can also be used to prove almost sure convergence of the SAA estimators towards true mixed logit likelihood estimators (Bastin *et al.*, Submitted), as the sampling size R tends to infinity, for a fixed population size. Moreover, from the delta method, $SLL^R(\theta)$ can be shown to be an asymptotically unbiased estimator of $LL(\theta)$, and the asymptotic value of the confidence interval radius is given by

$$\epsilon_\delta = \alpha_\delta \frac{1}{I} \sqrt{\sum_{i=1}^{I} \frac{\sigma_{ij_i}^2(\theta)}{R\left(P_{ij_i}(\theta)\right)^2}}, \tag{6}$$

where α_δ is the quantile of a $N(0,1)$, associated with some level of signification δ (for instance $\alpha_{0.9} \approx 1.64$). Finally, the simulation bias for finite R can be approximated by the quantity

$$E[SLL^R(\theta)] - LL(\theta) = -\frac{I\epsilon_\delta^2}{2\alpha_\delta^2}.$$

Details of these derivations can be again found in Bastin *et al.* (Submitted).

3 A new algorithm for solving the SAA problem

Statistical inference can be used to reduce the cost associated to the solution of the SAA problem (5), by limiting the number of draws needed in the early iterations, away from the solution. The main idea is to generate a sample set prior the optimization process, with $R_{\max}$ i.i.d. random draws per individual. At iteration k, only a (possibly small) subset of this sample set will be used, by selecting R_k of the $R_{\max}$ random draws for each individual (for simplicity, the first R_k). This idea is exploited in a trust-region algorithm (see Conn *et al.* (2000)). The main idea of a trust-region algorithm is, at a current iterate θ_k, to calculate a trial point $\theta_k + s_k$ by maximizing a model m_k of the objective function inside a trust region $\mathcal{B}_k = \{\theta \in \mathbb{R}^m | \|\theta - \theta_k\| \leq \Delta_k\}$, where Δ_k is called the trust-region radius. The predicted and actual increases in objective function values are then compared, leading to different algorithmic decisions. In particular, if the model approximates the SAA objective function well and gives a sufficient increase compared to its accuracy, we surmise that we could work with a less precise approximation and therefore reduce the sample size. On the other hand, if the model adequation or predicted increase is poor compared to the precision of the objective function, we put the sample size to a higher value. More formally, we have the following algorithm, whose details and convergence proof can be found in Bastin (2004).

Algorithm 1 (A trust-region algorithm with dynamic accuracy).

Step 0. Initialization. *An initial point θ_0 and trust-region radius Δ_0 are given. Set a minimum number of draws $R_{\min} = R_{\min}^0$ and a sample size R_0 satisfying $\|\nabla_\theta SLL^{R_0}(\theta_0)\| \neq 0$ if $\epsilon_\delta^{R_0}(\theta_{k+1}) \neq 0$, except if $R_0 = R_{\max}$. Compute $SLL^{R_0}(\theta_0)$ and set $k = 0$.*

Step 1. Stopping test. *Due to the presence of statistical error, classical stopping tests can lead to final iterations that produce insignificant objective increases compared to the approximation accuracy, so we stop the iterative process if*

$$\|\nabla_{\theta_k} SLL^{R_k}(\theta_k)\| \leq \max(0.2\epsilon_{0.9}^{R_k}(\theta_k), 10^{-6}),$$

and either the maximum sample size $R_{\max}$ is used or, in order to consider the multinomial logit case, the estimated log-likelihood accuracy is sufficiently small. We also stop the algorithm if the norm of the computed step falls under a user-defined significativity threshold.

Otherwise go to Step 2.

Step 2. Model definition. *Define a quadratic model $m_k^{R_k}$ of $SLL^{R_k}(\theta)$ in $\mathcal{B}_k$:*

$$m_k^R(\theta_k + s) = SLL^R(\theta_k) + \langle \nabla_\theta SLL^R(\theta_k), s\rangle + \frac{1}{2}\langle s, H_k s\rangle,$$

where H_k is a symmetric approximation to $\nabla^2_{\theta\theta} SLL^R(\theta_k)$. Compute a new adequate sample size R^+ (see Algorithm 2 below). Set $R^- = R_k$.

Step 3. Step calculation. *Compute a step s_k using the Steihaug-Toint method. Set $\Delta m_k^{R_k} = m_k^{R_k}(\theta_k + s_k) - m_k^{R_k}(\theta_k)$.*

Step 4. Comparison of increases. *Compute $SLL^{R^+}(\theta_k + s_k)$ and define*

$$\rho_k = \frac{SLL^{R^+}(\theta_k + s_k) - SLL^{R_k}(\theta_k)}{\Delta m_k^{R_k}}. \tag{7}$$

Step 5. Sample size update. *If $\rho_k < 0.01$ and $R_k \neq R^+$, modify R^- and the candidate sample size R^+ to take into account the approximation error variations. We can indeed observe a decrease in the objective value due to differences in SAA variance and due to the increase of the bias (in absolute value), when the number of draws goes down, while the candidate iterate is a good one. Recompute ρ_k.*

Step 6. Acceptance of the trial point. *If $\rho_k < 0.01$, define $\theta_{k+1} = \theta_k$, $R_{k+1} = R^-$. Otherwise define $\theta_{k+1} = \theta_k + s_k$ and set $R_{k+1} = R^+$.*

In order to ensure convergence to a solution of the original SAA problem, increase the sample size R_{k+1} if we encounter a first-order critical point associated to a number draws less than $R_{\max}$, with an non-null approximation error $\epsilon_\delta^{R_{k+1}}(\theta_{k+1})$. Increase also the minimum sample size $R_{\min}^{k+1}$ if $R_k \neq R_{k+1}$ and no sufficient increase has been observed since the last evaluation of $SLL^{R_{k+1}}$.

Step 7. Trust-region radius update. *Increase the trust-region radius if ρ_k is sufficiently large, otherwise reduce it. Increment k by 1 and go to Step 1.*

Prior to the optimization, the user chooses the maximum sample size $R_{\max}$, and the minimum sample size $R_{\min}^0$ is defined to allow estimation of the accuracy (we used $R_{\min}^0 = 36$). The choice of R^+ in Step 3 of the previous algorithm is described below.

Algorithm 2 (Candidate sample size selection).
Define a constant ν_1 such that $\nu_1 \in (0, 1)$. Use (6) to estimate the size needed to obtain a precision equal to the model increase, that is

$$R^s = \max\left\{ R_{\min}^k, \left\lceil \frac{\alpha_\delta^2}{(I\Delta m_k^{R_k})^2} \sum_{i=1}^{I} \frac{\left(\sigma_{ij_i}^{R_k}(\theta)\right)^2}{\left(P_{ij_i}^{R_k}(\theta)\right)^2} \right\rceil \right\}.$$

Compute the ratio between the model improvement and the estimated accuracy,

$$\tau_1^k = \frac{\Delta m_k^{R_k}}{\epsilon_\delta^{R_k}(\theta_k)},$$

and the ratio between the current sample size and the suggested one for the next iteration:

$$\tau_2^k = \frac{R_k}{\min\{R_{\max}, R^s\}}.$$

Then define

$$R' = \begin{cases} \min\{\lceil 0.5R_{\max}\rceil, \lceil R^s\rceil\} & \text{if } \tau_1^k \geq 1, \\ \min\left\{\lceil 0.5R_{\max}\rceil, \lceil \tau_1^k R^s\rceil\right\} & \text{if } \tau_1^k < 1 \text{ and } \tau_1^k \geq \tau_2^k, \\ \lceil 0.5R_{\max}\rceil & \text{if } \nu_1 \leq \tau_1^k < 1 \text{ and } \tau_1^k < \tau_2^k, \\ R_{\max} & \text{if } \tau_1^k < \nu_1 \text{ and } \tau_1^k < \tau_2^k. \end{cases}$$

Set $R^+ = \max\{R', R_{\min}^k\}$.

If $\tau_1^k \geq 1$, the model increase is greater or equal to the estimated accuracy, so we can reduce the sample size to the minimum between R^s and $\lceil 0.5R_{\max}\rceil$. If $\tau_1^k < 1$ the improvement is smaller than the precision. However, a sufficient improvement during several consecutive iterations may lead to a significant improvement compared to the log-likelihood accuracy, while keeping the computational costs lower than if $R_{\max}$ draws were used. If $\tau_1^k \geq \tau_2^k$, we capitalize on the fact that the ratio between the current sample size and the potential next one is lower than the ratio between the model increase and the estimated error, while if $\tau_1^k < \tau_2^k$, it may be cheaper to continue to work with a smaller sample size, defined again as $\lceil 0.5R_{\max}\rceil$, as long as τ_1^k is superior to some threshold $\nu_1 > 0$ (set to 0.2 in our tests). Below to this threshold, we consider that the increase is too small compared to the log-likelihood accuracy, and we possibly increase the sample size.

We briefly illustrate the numerical behaviour of the proposed method in Figure 1, which shows the number of draws used per individual at each iteration (on the left) and its evolution with the log-likelihood value (on the right). These graphs correspond to the calibration of a model based on the *Mobidrive* data set (Axhausen *et al.*, 2002)

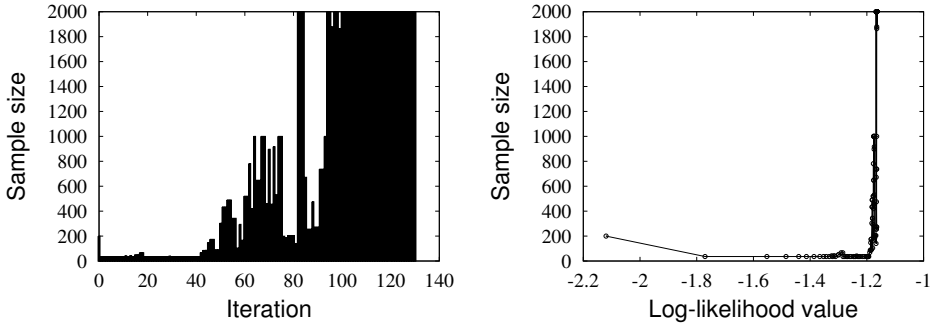

Figure 1: Variation of sample size

4 Conclusion

In this paper, we have introduced a new trust-region algorithm for SAA problems occurring in mixed logit models estimation. This algorithm allows to vary the number of used draws from iteration to iteration, and can be shown to be convergent while numerically efficient (Bastin, 2004). Moreover, the use of statistical inference allows to give information about the quality of the approximation of the optimal value, an interesting indication when we have to decide of the number of draws to use. Extensions to more complicated models, as well as other approximations techniques, are currently under investigation.

References

K. W. Axhausen, A. Zimmerman, S. Schönfelder, G. Rindsfüser, and T. Haupt, "Observing the rythms of daily life: A six week travel diary", *Transportation*, 29 (2), 95–124 (2002).

F. Bastin, Trust-region, Ph.D. thesis, Department of Mathematics, University of Namur, Namur, Belgium (2004).

F. Bastin, C. Cirillo, and P. L. Toint, "Convergence theory for nonconvex stochastic programming with an application to mixed logit", *Mathematical Programming, Series B* (Submitted).

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

A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust-Region Methods*, SIAM, Philadelphia, USA (2000).

S. Hess, K. Train, and J. Polak, "On the use of randomly shifted uniform vectors in the estimation of a Mixed Logit model for vehicle choice", *Transportation Research B* (Submitted).

R. Y. Rubinstein and A. Shapiro, *Discrete Event Systems*, John Wiley & Sons, Chichester, England (1993).

A. Shapiro, "Stochastic Programming by Monte Carlo Simulation Methods", *SPEPS* (2000).

A. Shapiro, "Monte Carlo Sampling Methods", in A. Shapiro and A. Ruszczyński (eds), *Stochastic Programming*, vol. 10 of *Handbooks in Operations Research and Management Science*, 353–425, Elsevier, 2003.

K. Train, "Halton Sequences for Mixed Logit", Working paper No. E00-278, Department of Economics, University of California, Berkeley, USA, 1999.