A conditional trust-region algorithm for the estimation of discrete choice models

Nicola Ortelli

Matthieu de Lapparent

Michel Bierlaire

STRC conference paper 2024 **June 30, 2024**

STRC | **24th Swiss Transport Research Conference**
Monte Verità / Ascona, May 15-17, 2024

A conditional trust-region algorithm for the estimation of discrete choice models

Nicola Ortelli, Matthieu de Lapparent
School of Management and Engineering Vaud
HES-SO
Yverdon-les-Bains, Switzerland
nicola.ortelli@heig-vd.ch

Nicola Ortelli, Michel Bierlaire
Transport and Mobility Laboratory
EPFL
Lausanne, Switzerland

June 30, 2024

Abstract

In the field of choice modeling, the availability of ever-larger datasets has the potential to significantly expand our understanding of human behavior, but this prospect is limited by the poor scalability of discrete choice models (DCMs): as sample sizes increase, the computational cost of maximum likelihood estimation quickly becomes intractable for anything but trivial model structures. To tackle this issue, this study builds upon the idea of using stochastic optimization algorithms for the estimation of DCMs. Specifically, we investigate the use of a dataset reduction technique to generate weighted batches that better represent the whole dataset and, as a result, lead the optimization algorithm to faster convergence. We use a real-world dataset and models of different sizes and complexity to compare the performance of our approach with existing methods used in practice.

Keywords

discrete choice models, maximum likelihood estimation, trust-region methods, stochastic optimization

Suggested Citation

Ortelli, N., de Lapparent, M., Bierlaire, M. (2024). A conditional trust-region algorithm for the estimation of discrete choice models, 24th Swiss Transport Research Conference, Ascona, Switzerland.

1 Introduction

Big data has caused a surge in the amount of data collected on practically any object of study. In the field of discrete choice analysis, the availability of these ever-larger datasets could improve our understanding of human decision-making, but that prospect is limited by the poor scalability of estimation methods for discrete choice models (DCMs).

DCMs are usually estimated via maximum likelihood estimation, which most often relies on optimization algorithms such as Newton's method, BFGS (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970), or one of their variations. These algorithms are extremely effective when estimating simple models on small datasets, but they quickly become computationally expensive as model complexity and dataset sizes grow. To circumvent this limitation, Lederrey *et al.* (2021) recently proposed using stochastic approximations of these methods to estimate DCMs. Similar to the stochastic gradient descent method used to train neural networks, a crucial feature of the algorithms developed by Lederrey *et al.* (2021) is the use of subsets of data — or batches — of increasing size throughout the optimization process: at each iteration, a new batch is randomly drawn whose size is determined according to the advancement of the process, until the full dataset is eventually reached and the algorithm converges to the maximum likelihood estimates of the model parameters. Lederrey *et al.* (2021) empirically demonstrate that the use of batches in the earlier stages of the optimization process significantly contributes to reducing the total computational time of model estimation; similar results are also obtained in some of our previous studies (Ortelli *et al.*, 2023, 2024).

In this study, we further investigate the use of stochastic algorithms for the estimation of DCMs. We propose integrating the resampling technique proposed by Ortelli *et al.* (2024) within a trust-region framework; the resulting *conditional* trust-region (CTR) algorithm builds quadratic approximations of the likelihood function obtained on dynamically adapted weighted subsamples, which, by construction, are less computationally demanding than the full dataset. In doing so, the goal is to better guide the optimization algorithm during its earlier stages, while maintaining a low computational cost per iteration.

The remainder of this paper is organized as follows: Section 2 reviews the main ideas of trust-region methods and then proceeds to describe our proposed algorithm; Section 3 discusses some preliminary results obtained by comparing the performance of our algorithm with a basic trust-region algorithm; finally, Section 4 summarizes the findings of this study and identifies directions for future research.

2 Methodology

Consider a choice dataset of N observations (x_n, i_n), each consisting of a vector x_n of explanatory variables associated with individual n, together with the observed choice i_n of that same individual among J alternatives. In its simplest form, a discrete choice model $P(i \,|\, x_n;\, \theta)$ calculates the probability that individual n chooses any alternative i as a function of x_n and θ, where $\theta \in \mathbb{R}^L$ is a vector of L parameters to be estimated.

The values of the model parameters are typically determined through maximum likelihood estimation (MLE), which consists in finding the values of θ that maximize the joint probability of replicating all observed choices in the dataset. In practice, however, the *log likelihood* is used instead, for numerical reasons:

$$\mathcal{L}(\theta) = \log \prod_{n=1}^{N} P\left(i_n \,|\, x_n; \theta\right) = \sum_{n=1}^{N} \log P\left(i_n \,|\, x_n; \theta\right). \tag{1}$$

The problem of maximizing (1) is typically solved using iterative algorithms such as Newton's method, BFGS, or one of their variations. Among such approaches, trust-region methods are of particular interest (Conn *et al.*, 2000). Those work by defining a region

$$\mathcal{B}_k = \left\{\theta \in \mathbb{R}^L \mid \|\theta - \theta_k\| \leq d_k\right\}$$

of radius d_k around the current iterate θ_k, in which a model function $m_k(\theta)$ serves as a local approximation of the objective function — *i.e.*, $\mathcal{L}(\theta)$, in our case. A trial step s_k to a trial point $\theta_k + s_k$ is chosen such that the model function within the trust region $\mathcal{B}_k$ is maximized and, after each step, the radius d_k of the region is adjusted for the next iteration, based on the agreement between the model and the true objective function: if the actual improvement $[\mathcal{L}(\theta_k + s_k) - \mathcal{L}(\theta_k)]$ observed in the objective function is close to or larger than the expected improvement $[m_k(\theta_k + s_k) - m_k(\theta_k)]$ from the model function, the trust region is expanded; conversely, if the reduction in the model function turns out to be a poor predictor of the actual behavior of the log likelihood, the region is contracted.

While many options exist for the model of the objective function, a common choice is a quadratic formulation, which, in the context of MLE, is given as

$$m_k(\theta_k + s_k) = \mathcal{L}\left(\theta_k\right) + \left\langle \nabla \mathcal{L}\left(\theta_k\right), s_k \right\rangle + \tfrac{1}{2}\left\langle s_k, \nabla^2 \mathcal{L}\left(\theta_k\right) s_k \right\rangle, \tag{2}$$

where $\nabla \mathcal{L}\left(\theta_k\right)$ and $\nabla^2 \mathcal{L}\left(\theta_k\right)$ are the gradient and the Hessian of $\mathcal{L}\left(\theta_k\right)$, respectively.

While the quadratic formulation is effective, the computational cost of evaluating the gradient and Hessian at each iteration is high, in particular with large datasets or with complex model formulations that include many parameters. To mitigate this limitation, we propose replacing the log likelihood $\mathcal{L}(\theta)$ and its derivatives $\nabla\mathcal{L}(\theta)$ and $\nabla^2\mathcal{L}(\theta)$ in (2) by some approximations $\tilde{\mathcal{L}}_k(\theta)$, $\nabla\tilde{\mathcal{L}}_k(\theta)$ and $\nabla^2\tilde{\mathcal{L}}_k(\theta)$ that are faster to compute. For this purpose, we suggest using the resampling technique proposed by Ortelli *et al.* (2024), called LSH-DR, to generate weighted subsamples that closely resemble the full dataset.

In LSH-DR, the number of sampled observations solely depends on a parameter w called the bucket width, and, by construction, so do the quality of the obtained approximations and the computational burden associated with their evaluation. We therefore recommend starting with small subsamples so as to speed up the earlier stages of the optimization process; then, each time the expected improvement $[m_k(\theta_k+s_k) - m_k(\theta_k)]$ turns out to be a poor predictor of the actual behavior $[\mathcal{L}(\theta_k+s_k) - \mathcal{L}(\theta_k)]$, we need to decide between reducing the trust-region radius d_k or reducing the bucket width w_k so as to obtain better approximations $\tilde{\mathcal{L}}_k(\theta)$, $\nabla\tilde{\mathcal{L}}_k(\theta)$ and $\nabla^2\tilde{\mathcal{L}}_k(\theta)$. We base this choice on the ratio between $[\tilde{\mathcal{L}}_k(\theta_k+s_k) - \tilde{\mathcal{L}}_k(\theta_k)]$ and $[m_k(\theta_k+s_k) - m_k(\theta_k)]$.

Our conditional trust-region (CTR) algorithm is organized as follows.

Input An initial point θ_0, an initial bucket width w_0 and an initial trust-region radius d_0 are given, as are the constants $0 < \tilde{\eta} \leq \eta \leq \eta^+ < 1$, $\gamma_d > 1$ and $0 < \gamma_w < 1$.

Initialization Use w_0 to create $\tilde{\mathcal{L}}_0$ and set $k = 0$.

Iteration

1. Define a model m_k in $\mathcal{B}_k$.
2. Compute a trial step s_k such that $\theta_k+s_k \in \mathcal{B}_k$.[1]
3. Compute $m_k(\theta_k + s_k) = \tilde{\mathcal{L}}_k(\theta_k) + \langle \nabla\tilde{\mathcal{L}}_k(\theta_k), s_k\rangle + \frac{1}{2}\langle s_k, \nabla^2\tilde{\mathcal{L}}_k(\theta_k)\, s_k\rangle$ and

   $$\rho_k = \frac{\mathcal{L}(\theta_k+s_k) - \mathcal{L}(\theta_k)}{m_k(\theta_k+s_k) - m_k(\theta_k)}.$$

   If $\rho_k \geq \eta$, set $\theta_{k+1} = \theta_k+s_k$; otherwise, set $\theta_{k+1} = \theta_k$.
4. Compute

   $$\tilde{\rho}_k = \frac{\tilde{\mathcal{L}}_k(\theta_k+s_k) - \tilde{\mathcal{L}}_k(\theta_k)}{m_k(\theta_k+s_k) - m_k(\theta_k)}.$$

   If $\rho_k < \eta$ and $\tilde{\rho}_k \geq \tilde{\eta}$, set $w_{k+1} = \gamma_w w_k$; otherwise, set $w_{k+1} = w_k$.

[1] For instance, using a truncated conjugate-gradient method.

5. Set

$$d_{k+1} = \begin{cases} \gamma_d \, d_k & \text{if } \rho_k \geq \eta^+, \\ d_k & \text{if } \rho_k \in [\eta, \eta^+), \\ d_k & \text{if } \rho_k < \eta \text{ and } \tilde{\rho}_k \geq \tilde{\eta}, \\ d_k/2 & \text{if } \rho_k < \eta \text{ and } \tilde{\rho}_k < \tilde{\eta}. \end{cases}$$

6. If

$$\max_j \frac{[\nabla \tilde{\mathcal{L}}_k(\theta_k)]_j \theta_{k,j}}{\tilde{\mathcal{L}}_k(\theta_k)} < \epsilon,$$

stop the algorithm; otherwise, use w_{k+1} to create $\tilde{\mathcal{L}}_{k+1}$ and increment k by one.

3 Experiments

We evaluate the performance of our algorithm on a series of nine increasingly complex logit, nested logit and cross-nested logit models. Those are based on the London passenger mode choice (LPMC) dataset (Hillel *et al.*, 2018), which consists of over 81'000 trip records collected over three years. Four modes are distinguished: walk, cycle, ride public transport and drive.[2] Table 1 reports the number of explanatory variables and parameters considered in each model.

Table 1: Complexity of the six considered models.

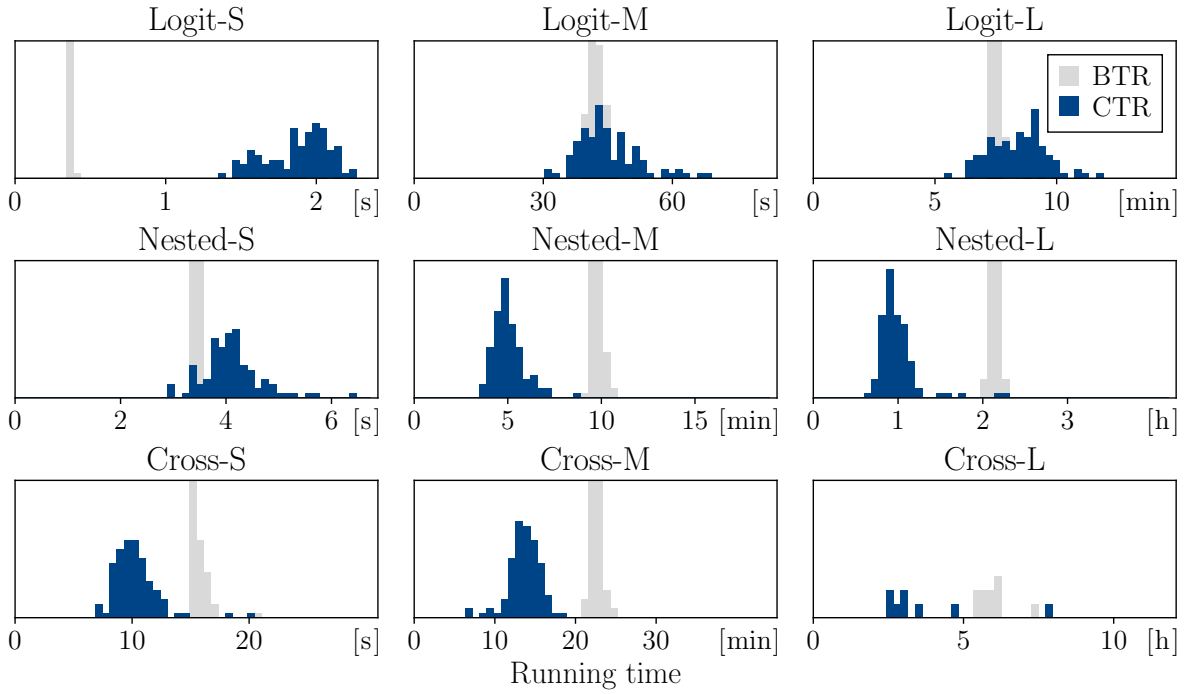
	Logit			Nested			Cross		
	S	M	L	S	M	L	S	M	L
Continuous variables	10	11	13	10	11	13	10	11	13
Binary variables	0	15	18	0	15	18	0	15	18
Parameters	13	53	100	14	54	101	15	55	102

Our CTR algorithm is compared to the basic trust region (BTR) algorithm implemented in Biogeme (Bierlaire, 2023). The two algorithms are used to estimate each model a 100 times;[3] the results are then compared in terms of overall running time and number of epochs. All estimations are performed on two Intel Xeon Platinum 8360Y processors running at 2.4 GHz, for a total of 72 cores and 512 GB of RAM.

[2]The "drive" alternative also includes car passenger, taxi, van and motorbike.
[3]Due to its cost, the estimation of the Cross-L model is repeated only 10 times.

Figure 1 compares the obtained overall running times for the two algorithms. For the more complex models, our CTR algorithm is shown to outperform BTR by a significant margin. In particular, the Nested-M and Nested-L models are shown to be estimated twice as fast as with the basic algorithm, whereas an average of 2.3 hours are saved during the estimation of the Cross-L model. As regards the simpler models, it appears that the performance of the two algorithms is comparable, but this is due to the additional time that CTR requires to subsample the dataset; indeed, as shown in Figure 2, CTR is actually more efficient in the way it uses the data. Logit-S is the model for which this phenomenon is the most striking: CTR is almost 5 times slower than BTR in terms of running time, but it actually takes half the epochs to reach convergence.

Figure 1: Running times of the BTR and CTR algorithms.



4 Conclusion

In this study, we propose a stochastic version of the trust-region algorithm for the estimation of discrete choice models. Our approach relies on a simple rule to decide when the batch size needs to be increased, and the batches are obtained from a dataset reduction technique that is specifically designed to preserve the diversity of observations in the data, so as to better guide the optimization algorithm. The presented preliminary results highlight the potential of this approach in the estimation of logit, nested logit and cross-nested logit models of medium to large sizes.

Intended future work focuses on extending our algorithm to mixed logit models and maximum simulated likelihood estimation, which requires to carefully examine the interaction between the subsampling method and Monte Carlo integration. Following the work of Bastin *et al.* (2006), a promising approach could consist in considered the number of draws using in Monte Carlo integration as an additional parameter that starts from a low value and grows iteratively during the optimization process.

Figure 2: Number of epochs of the BTR and CTR algorithms.

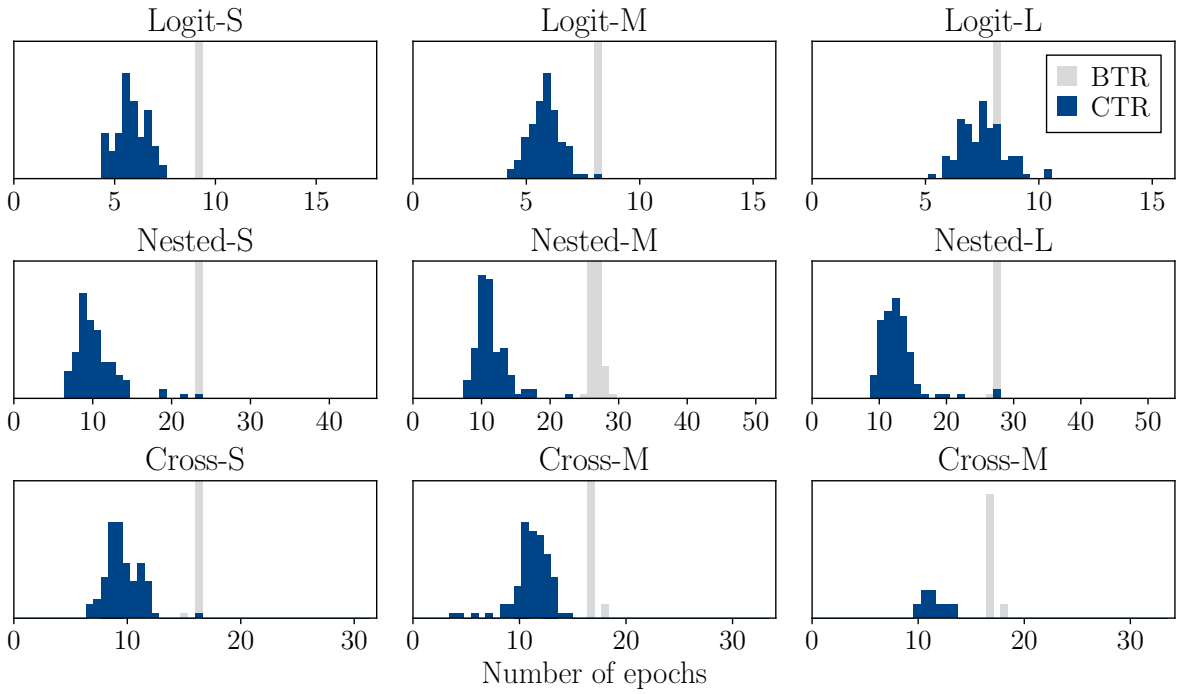


Table 2: Running times and numbers of epochs of the BTR and CTR algorithms.

Model	Running time		Epochs	
	BTR	CTR	BTR	CTR
Logit-S	0.4 ± 0.0	1.9 ± 0.2	9.0 ± 0.0	5.8 ± 0.8
Nested-S	3.4 ± 0.0	4.1 ± 0.6	23.0 ± 0.0	10.4 ± 2.8
Cross-S	15.5 ± 0.7	10.2 ± 1.9	16.0 ± 0.1	9.6 ± 1.5
Logit-M	42 ± 2	45 ± 7	8.0 ± 0.0	5.8 ± 0.7
Nested-M	582 ± 18	303 ± 50	26.5 ± 0.7	11.3 ± 2.3
Cross-M	1'350 ± 42	820 ± 117	17.0 ± 0.2	11.2 ± 1.7
Logit-L	447 ± 13	505 ± 70	8.0 ± 0.0	7.5 ± 1.0
Nested-L	7'708 ± 180	3'578 ± 869	27.0 ± 0.1	12.7 ± 3.0
Cross-L	21'743 ± 1'906	13'435 ± 5'960	26.5 ± 1.0	16.9 ± 6.8

5 References

Bastin, F., C. Cirillo and P. L. Toint (2006) An adaptive monte carlo algorithm for computing mixed logit estimators, *Computational Management Science*, **3**, 55–79.

Bierlaire, M. (2023) A short introduction to Biogeme, *Technical Report*, TRANSP-OR 230620. Transport and Mobility Laboratory, ENAC, EPFL.

Broyden, C. G. (1970) The convergence of a class of double-rank minimization algorithms 1. general considerations, *IMA Journal of Applied Mathematics*, **6** (1) 76–90.

Conn, A. R., N. I. Gould and P. L. Toint (2000) *Trust region methods*, SIAM.

Fletcher, R. (1970) A new approach to variable metric algorithms, *The computer journal*, **13** (3) 317–322.

Goldfarb, D. (1970) A family of variable-metric methods derived by variational means, *Mathematics of computation*, **24** (109) 23–26.

Hillel, T., M. Z. Elshafie and Y. Jin (2018) Recreating passenger mode choice-sets for transport simulation: A case study of London, UK, *Proceedings of the Institution of Civil Engineers-Smart Infrastructure and Construction*, **171** (1) 29–42.

Lederrey, G., V. Lurkin, T. Hillel and M. Bierlaire (2021) Estimation of discrete choice models with hybrid stochastic adaptive batch size algorithms, *Journal of choice modelling*, **38**, 100226.

Ortelli, N., Lapparent, M. (de) and M. Bierlaire (2023) Stochastic adaptive resampling for the estimation of discrete choice models, paper presented at the *Proceedings of the 23rd Swiss Transportation Research Conference*.

Ortelli, N., Lapparent, M. (de) and M. Bierlaire (2024) Resampling estimation of discrete choice models, *Journal of Choice Modelling*, **50**, 100467, ISSN 1755-5345.

Shanno, D. F. (1970) Conditioning of quasi-newton methods for function minimization, *Mathematics of computation*, **24** (111) 647–656.