SNM: Stochastic Newton Method for optimization of Discrete Choice Models

Gael Lederrey
Transport and Mobility Laboratory
École Polytechnique Fédérale de Lausanne
Station 18, CH-1015 Lausanne
gael.lederrey@epfl.ch

Virginie Lurkin
Transport and Mobility Laboratory
École Polytechnique Fédérale de Lausanne
Station 18, CH-1015 Lausanne
virginie.lurkin@epfl.ch

Michel Bierlaire
Transport and Mobility Laboratory
École Polytechnique Fédérale de Lausanne
Station 18, CH-1015 Lausanne
michel.bierlaire@epfl.ch

***Abstract*—BLA BLA BLA**

***Index Terms*—Discrete Choice Models, Optimization**

I. Introduction

☞ **Note:**

- Not a lot of work on optimization of DCMs
- ML are doing this a lot
- Expecting a lot more data. =¿ Standard quasi Newton may have troubles
- Becomes interesting to search for new algorithms exploiting the structure of DCMs

II. Related Work

☞ **Note:**

- ???

III. Methodology

In this section, we present the model used in this paper as well as some algorithms that have been tested.

A. Model

☞ **Note:**

- Show the model.
- Give results from biogeme.
- Describe algorithms.

We use the *Swissmetro* dataset [1] and build a multinomial logit model denoted by $\mathcal{M}$:

$$\begin{aligned}
V_{\text{Car}} &= \text{ASC}_{\text{Car}} + \beta_{\text{TT,Car}}\text{TT}_{\text{Car}} + \beta_{\text{C,Car}}\text{C}_{\text{Car}} + \beta_{\text{Senior}}\mathbb{1}_{\text{Senior}} \\
V_{\text{SM}} &= \text{ASC}_{\text{SM}} + \beta_{\text{TT,SM}}\text{TT}_{\text{SM}} + \beta_{\text{C,SM}}\text{C}_{\text{SM}} \\
&\quad + \beta_{\text{HE}}\text{HE}_{\text{SM}} + \beta_{\text{Senior}}\mathbb{1}_{\text{Senior}} \\
V_{\text{Train}} &= \text{ASC}_{\text{Train}} + \beta_{\text{TT,Train}}\text{TT}_{\text{Train}} + \beta_{\text{C,Train}}\text{C}_{\text{Train}} + \beta_{\text{HE}}\text{HE}_{\text{Train}}
\end{aligned} \tag{1}$$

where $\mathbb{1}_{\text{Senior}}$ is a boolean variable equal to one if the age of the respondent is over 65 years olds, 0 otherwise, C denotes the cost, TT the travel time, and HE the headway for the train and Swissmetro. On this model, we remove all observations with unknown choice, unkown age and non-positive travel time. This gives a total of 9,036 observations.

This model is first estimated with Biogeme [2] to obtain the optimal parameter values and verify that all parameters are significant. However, we do not use the usual log-likelihood. Instead, we are using a nomralized log-likelihood which simply corresponds to the log-likelihood divided by the number of observations. Therefore, the final normalized log-likelihood is -0.7908 and the parameters are given in Table I.

Name	Value	Std err	t-test	p-value
ASC_{Car}	0	-	-	-
ASC_{SM}	$7.86 \cdot 10^{-1}$	$6.93 \cdot 10^{-2}$	11.35	0.00
$\text{ASC}_{\text{Train}}$	$9.83 \cdot 10^{-1}$	$1.31 \cdot 10^{-1}$	7.48	0.00
$\beta_{\text{TT,Car}}$	$-1.05 \cdot 10^{-2}$	$7.89 \cdot 10^{-4}$	-8.32	0.00
$\beta_{\text{TT,SM}}$	$-1.44 \cdot 10^{-2}$	$6.36 \cdot 10^{-4}$	-21.29	0.00
$\beta_{\text{TT,Train}}$	$-1.80 \cdot 10^{-2}$	$8.65 \cdot 10^{-4}$	-20.78	0.00
$\beta_{\text{C,Car}}$	$-6.56 \cdot 10^{-3}$	$7.89 \cdot 10^{-4}$	-8.32	0.00
$\beta_{\text{C,SM}}$	$-8.00 \cdot 10^{-3}$	$3.76 \cdot 10^{-4}$	-21.29	0.00
$\beta_{\text{C,Train}}$	$-1.46 \cdot 10^{-2}$	$9.65 \cdot 10^{-4}$	-15.09	0.00
β_{Senior}	-1.06	$1.16 \cdot 10^{-1}$	-9.11	0.00
β_{HE}	$-6.88 \cdot 10^{-3}$	$1.03 \cdot 10^{-3}$	-6.69	0.00

TABLE I
PARAMETERS OF THE OPTIMIZED MODEL $\mathcal{M}$ BY BIOGEME.

We also provide a normalized model $\mathcal{M}_N$ where the values of travel time, cost and headway have been divided by 100. The parameters for this nomralized model are the same as model $\mathcal{M}$ except that the values of parameters associated to the features normalized are multiplied by 100. This is done such that all the parameters are in only one order of magnitude as opposed to the values in Table I where the parameter values are in four orders of magnitude.

B. Algorithms

To train models $\mathcal{M}$ and $\mathcal{M}_N$, many different algorithms were used. These algorithms fall in three different categories: first-order methods, second-order methods and quasi-newton methods. As first-order methods, we use mini-batch SGD [3] and Adagrad [4]. For the quasi-newton methods, we use BFGS algorithm [5] and RES-BFGS [6], a regularized stochastic version of BFGS. The main second-order algorithm is the Newton method [7]. In addition, we present in this paper a stochastic version of the Newton method simply denoted by Stochastic Newton Method (SNM), see Algorithm **??**. All these algorithms are run with a backtracking Line Search method using the Armijo-Goldstein condition [8].

IV. Results

☞ **NOTE:** In the results, I want to show that:

- First order methods do not work well. Especially when the model is not nomralized.
- Second-order methods works well, even when the model is not normalized.
- Quasi Newton method works better than 1st order methods but worse than second order methods.

V. Discussion

VI. Conclusion

VII. Acknowledgements

☞ **NOTE:** Thanks Tim!

References

[1] M. Bierlaire, K. Axhausen, and G. Abay, "The acceptance of modal innovation: The case of Swissmetro," *Swiss Transport Research Conference 2001*, Mar. 2001. [Online]. Available: https://infoscience.epfl.ch/record/117140

[2] M. Bierlaire, "BIOGEME: a free package for the estimation of discrete choice models," *Swiss Transport Research Conference 2003*, Mar. 2003. [Online]. Available: https://infoscience.epfl.ch/record/117133

[3] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv:1609.04747 [cs]*, Sep. 2016, arXiv: 1609.04747. [Online]. Available: http://arxiv.org/abs/1609.04747

[4] J. Duchi, E. Hazan, and Y. Singer, "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011. [Online]. Available: http://jmlr.org/papers/v12/duchi11a.html

[5] R. Fletcher, *Practical Methods of Optimization; (2Nd Ed.)*. New York, NY, USA: Wiley-Interscience, 1987.

[6] A. Mokhtari and A. Ribeiro, "RES: Regularized Stochastic BFGS Algorithm," *IEEE Transactions on Signal Processing*, vol. 62, no. 23, pp. 6089–6104, Dec. 2014.

[7] J. Caswell, "A treatise of algebra, both historical and practical : with some additional treatises I. of the cono-cuneus; being a body representing in part a conus, an part a cuneus ; II. of angular sections; and other things relating there unto, and to Trigonometry ; III. of the angle of contact; with other things appertaining to the composition of magnitudes, the inceptive of magnitudes, and the composition of motions, with the results thereof ; IV. of combination, alternations, and aliquot parts," Tech. Rep., 1685.

[8] L. Armijo, "Minimization of functions having Lipschitz continuous first partial derivatives," *Pacific Journal of Mathematics*, vol. 16, no. 1, pp. 1–3, Jan. 1966. [Online]. Available: https://msp.org/pjm/1966/16-1/p01.xhtml