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

16th Workshop on Discrete Choice Models
6–8 June 2024 | EPFL, Switzerland

Nicola Ortelli,[1,2] Matthieu de Lapparent,[1] Michel Bierlaire[2]

[1] IIDE, HEIG-VD, Switzerland
[2] TRANSP-OR, EPFL, Switzerland

Motivation

DCMs in the era of big data

- Developing DCMs is time-consuming.
- Ever-larger datasets:
  - "Wider" data — more variables;
  - "Taller" data — more observations.
- Two distinct problems: specification and estimation.

Motivation

DCMs in the era of big data

- Developing DCMs is time-consuming.
- Ever-larger datasets:
  - "Wider" data — more variables;
  - "Taller" data — more observations.
- Two distinct problems: specification and estimation.

Speeding up model estimation

- Optimization methods scale poorly with dataset size.
- Solution: consider fewer observations!

Intuition

Maximum likelihood estimation (MLE)

- Let $\mathcal{N} = \{(\boldsymbol{x}_n, i_n) : n = 1, \ldots, N\}$ be a choice dataset.
- Log likelihood function:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log P(i_n \mid \boldsymbol{x}_n; \boldsymbol{\theta}).$$

- Computational time is linear in N.

Intuition

Maximum likelihood estimation (MLE)

- Let $\mathcal{N} = \{(\boldsymbol{x}_n, i_n) : n = 1, \ldots, N\}$ be a choice dataset.
- Log likelihood function:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log P(i_n \mid \boldsymbol{x}_n; \boldsymbol{\theta}).$$

- Computational time is linear in N.

Factoring-out redundancy

- Suppose that some observations in $\mathcal{N}$ are identical.
- For $G < N$ unique observations:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{g=1}^{G} N_g \log P(i_g \mid \boldsymbol{x}_g; \boldsymbol{\theta}).$$

- Evaluation takes $\approx \frac{G}{N}$ less time!

Intuition

Maximum likelihood estimation (MLE)

- Let $\mathcal{N} = \{(\boldsymbol{x}_n, i_n) : n = 1, \ldots, N\}$ be a choice dataset.
- Log likelihood function:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log P(i_n \mid \boldsymbol{x}_n; \boldsymbol{\theta}).$$

- Computational time is linear in N.

Factoring-out redundancy

- Suppose that some observations in $\mathcal{N}$ are identical.
- For $G < N$ unique observations:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{g=1}^{G} N_g \log P(i_g \mid \boldsymbol{x}_g; \boldsymbol{\theta}).$$

- Evaluation takes $\approx \frac{G}{N}$ less time!

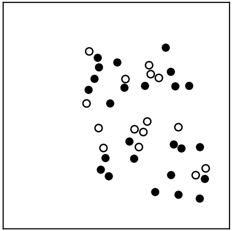
$\Rightarrow$ **Extend factorization to “nearly identical” observations!**

Resampling estimation of DCMs [Ortelli *et al.*, 2024]

Toy dataset

- 2 alternatives.
- 2 expl. variables.

Procedure



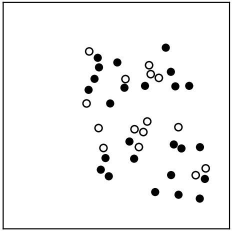
Resampling estimation of DCMs [Ortelli *et al.*, 2024]

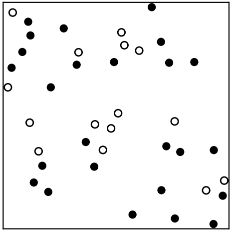
Toy dataset

- 2 alternatives.
- 2 expl. variables.

Procedure

1. Normalization.





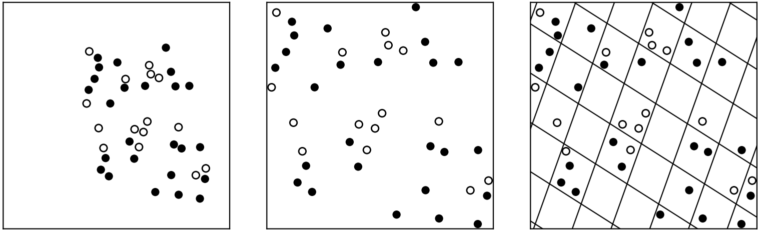
Resampling estimation of DCMs [Ortelli *et al.*, 2024]

Toy dataset

- 2 alternatives.
- 2 expl. variables.

Procedure

1. Normalization.
2. "Bucketing". (width w)



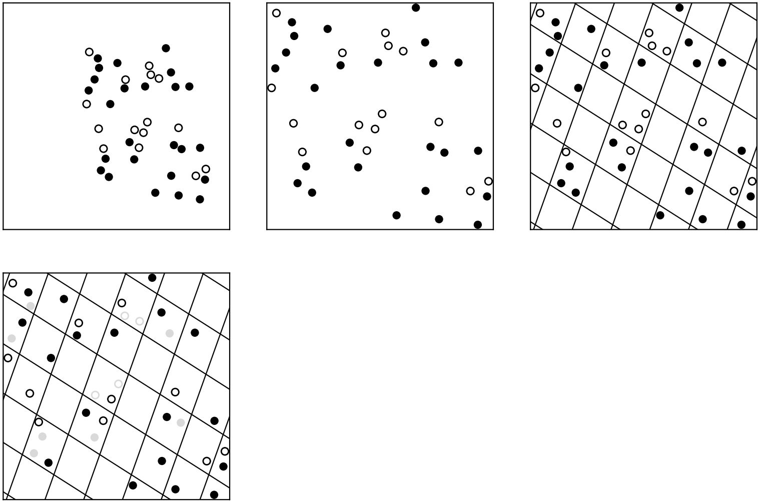
Resampling estimation of DCMs [Ortelli *et al.*, 2024]

Toy dataset

- 2 alternatives.
- 2 expl. variables.

Procedure

1. Normalization.
2. "Bucketing". (width w)
3. Sampling.



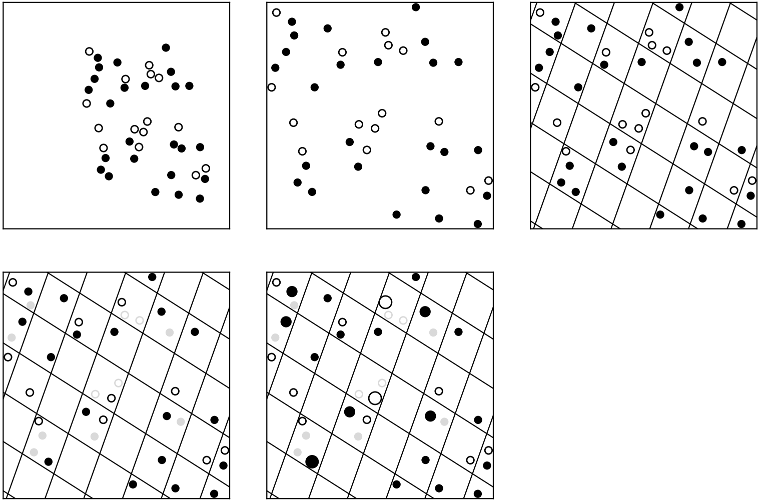
Resampling estimation of DCMs [Ortelli *et al.*, 2024]

Toy dataset

- 2 alternatives.
- 2 expl. variables.

Procedure

1. Normalization.
2. "Bucketing". (width w)
3. Sampling.
4. Weighting.



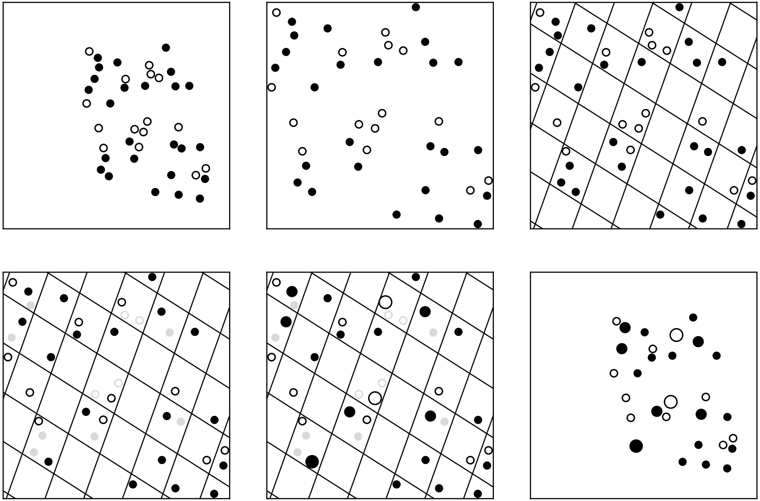
Resampling estimation of DCMs [Ortelli *et al.*, 2024]

Toy dataset

- 2 alternatives.
- 2 expl. variables.

Procedure

1. Normalization.
2. "Bucketing". (width w)
3. Sampling.
4. Weighting.
5. Rescaling.



Resampling estimation of DCMs [Ortelli *et al.*, 2024]

Toy dataset

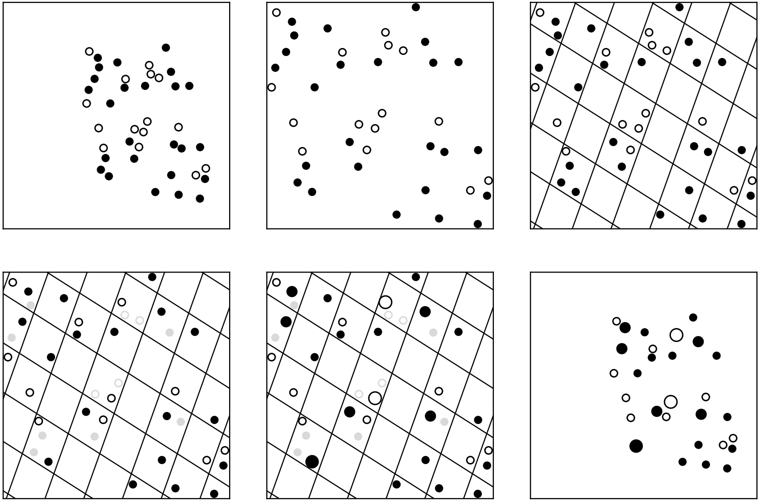
- 2 alternatives.
- 2 expl. variables.

Procedure

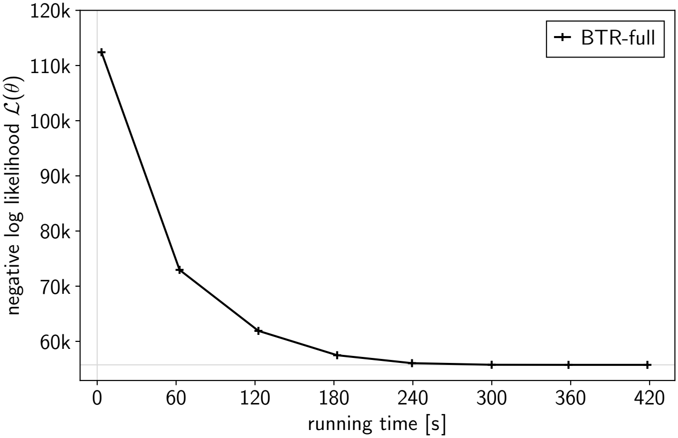
1. Normalization.
2. “Bucketing”. (width w)
3. Sampling.
4. Weighting.
5. Rescaling.

Weighted log likelihood

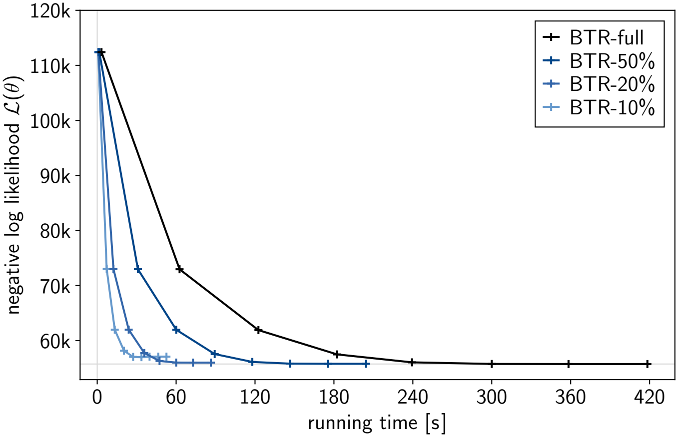
$$\tilde{\mathcal{L}}(\theta) = \sum_{g=1}^{G} N_g \log P(i_g | x_g; \theta)$$



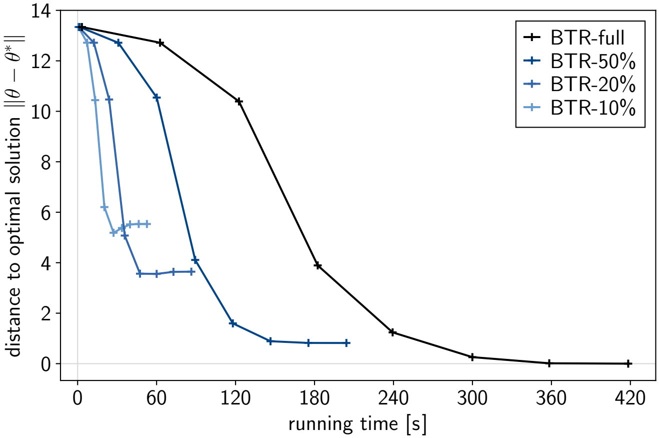
Illustration



Illustration



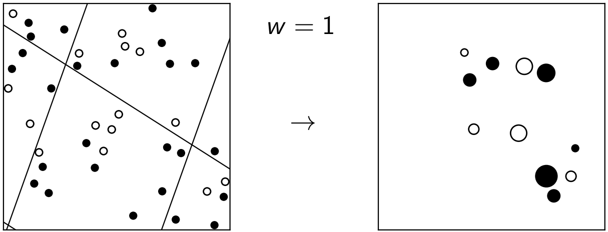
Illustration



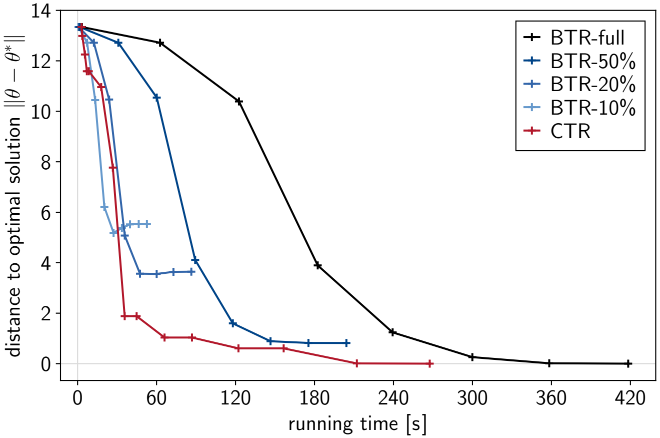
Adaptive resampling

Main idea

- Embed resampling within the model estimation process.
- Generate weighted batches for stochastic optimization.
- Start small and increase batch size dynamically.



Illustration



Basic trust-region (BTR)

Trust region $\mathcal{B}_k$

- Define $\mathcal{B}_k = \left\{\theta \in \mathbb{R}^L \mid \|\theta - \theta_k\| \leq d_k\right\}$ around iterate θ_k.
- Within $\mathcal{B}_k$, use a model function $m_k(\theta)$ as a local approximation of $\mathcal{L}(\theta)$.
- Find a step s_k that maximizes $m_k(\theta_k + s_k)$ in $\mathcal{B}_k$.
- Adjust d_k after each step.

Basic trust-region (BTR)

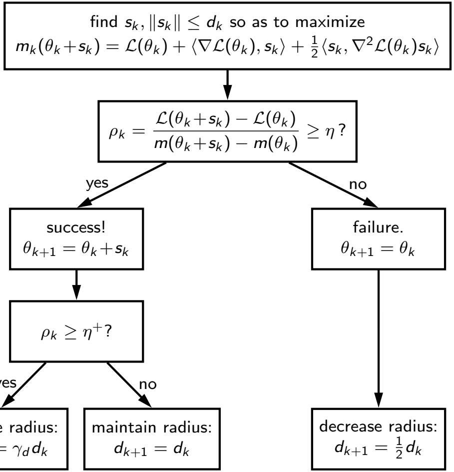
Trust region $\mathcal{B}_k$

- Define $\mathcal{B}_k = \left\{\theta \in \mathbb{R}^L \mid \|\theta - \theta_k\| \leq d_k\right\}$ around iterate θ_k.
- Within $\mathcal{B}_k$, use a model function $m_k(\theta)$ as a local approximation of $\mathcal{L}(\theta)$.
- Find a step s_k that maximizes $m_k(\theta_k + s_k)$ in $\mathcal{B}_k$.
- Adjust d_k after each step.

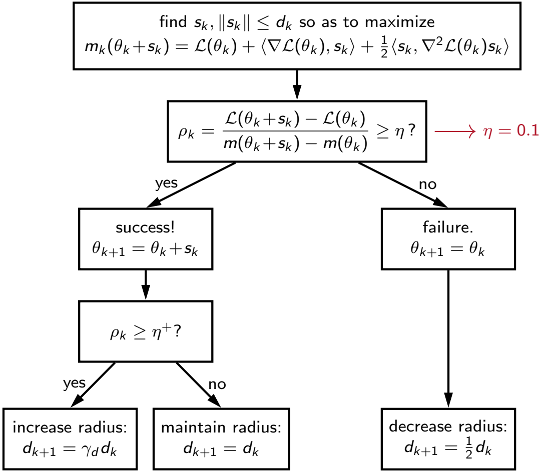
Model function $m_k(\theta)$

- Quadratic formulation.
- $m_k(\theta_k) = \mathcal{L}(\theta_k)$
- $m_k(\theta_k + s_k) = \mathcal{L}(\theta_k) + \langle \nabla \mathcal{L}(\theta_k), s_k \rangle + \frac{1}{2}\langle s_k, \nabla^2 \mathcal{L}(\theta_k) s_k \rangle$

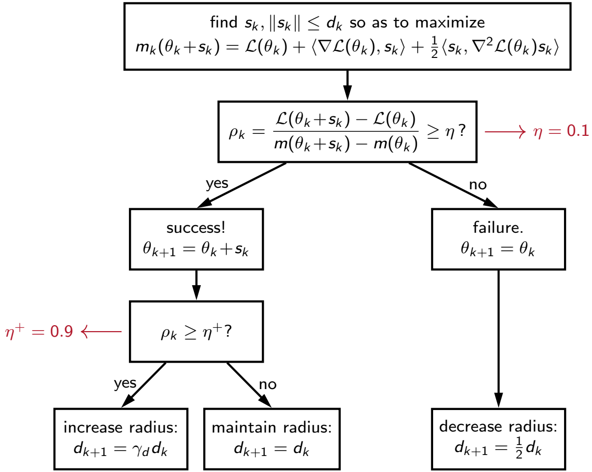
Basic trust-region (BTR)



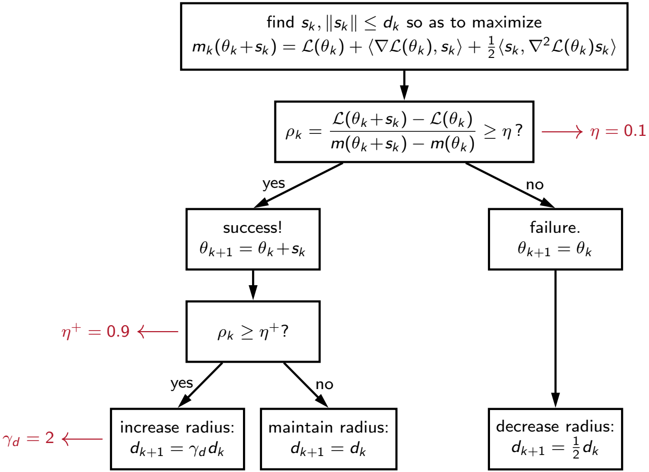
Basic trust-region (BTR)



Basic trust-region (BTR)



Basic trust-region (BTR)



Conditional trust-region (CTR)

Main idea

- $\mathcal{L}(\theta)$, $\nabla\mathcal{L}(\theta)$ and $\nabla^2\mathcal{L}(\theta)$ are computationally expensive.
- Replace them with approximations $\tilde{\mathcal{L}}(\theta)$, $\nabla\tilde{\mathcal{L}}(\theta)$ and $\nabla^2\tilde{\mathcal{L}}(\theta)$!

Conditional trust-region (CTR)

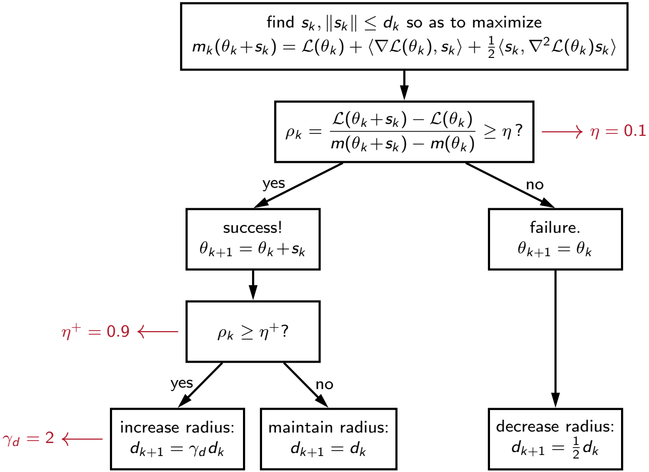
Main idea

- $\mathcal{L}(\theta)$, $\nabla\mathcal{L}(\theta)$ and $\nabla^2\mathcal{L}(\theta)$ are computationally expensive.
- Replace them with approximations $\tilde{\mathcal{L}}(\theta)$, $\nabla\tilde{\mathcal{L}}(\theta)$ and $\nabla^2\tilde{\mathcal{L}}(\theta)$!

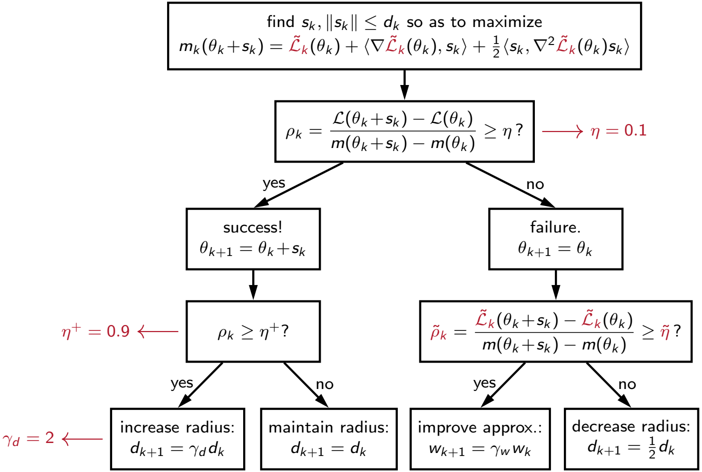
Procedure

- Start with a (very) small sample.
- After a failed iteration:
  - If $\tilde{\mathcal{L}}(\theta_k)$ and $\mathcal{L}(\theta_k)$ agree, decrease trust-region radius d_{k+1}.
  - If $\tilde{\mathcal{L}}(\theta_k)$ and $\mathcal{L}(\theta_k)$ disagree, decrease w_{k+1} to increase sample size;

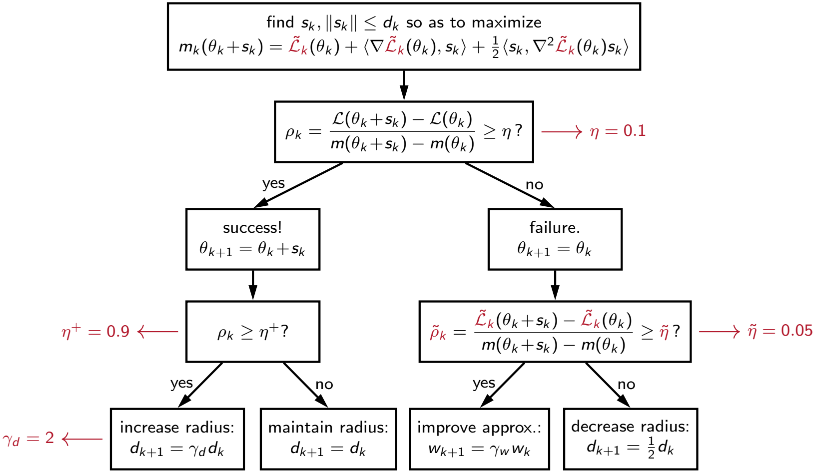
Basic trust-region (BTR)



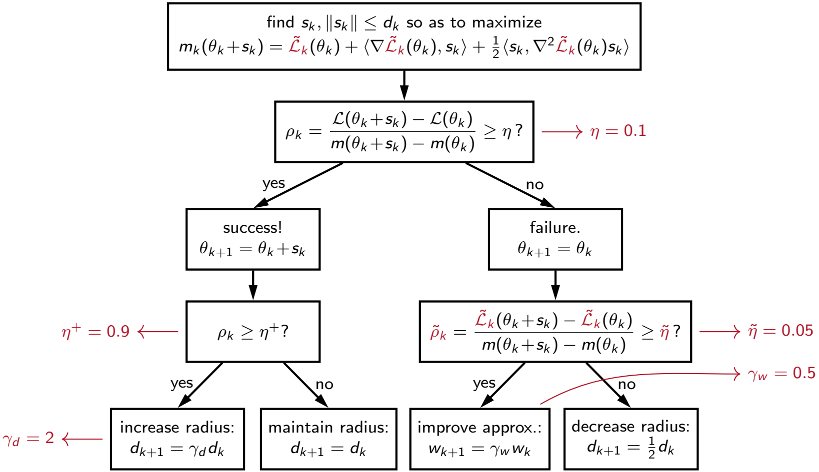
Conditional trust-region (CTR)



Conditional trust-region (CTR)



Conditional trust-region (CTR)



Experimental design

LPMC data [Hillel *et al.*, 2018]

- Mode choice.
- 4 alternatives: walk, cycle, drive, public transport.
- 81k observations.

Experimental design

Models

Logit-S
- 10 expl. variables.
- **13 parameters**.

Logit-M
- 26 expl. variables.
- **53 parameters**.

Logit-L
- 31 expl. variables.
- **100 parameters**.

Nested-S
- 10 expl. variables.
- **14 parameters**.

Nested-M
- 26 expl. variables.
- **54 parameters**.

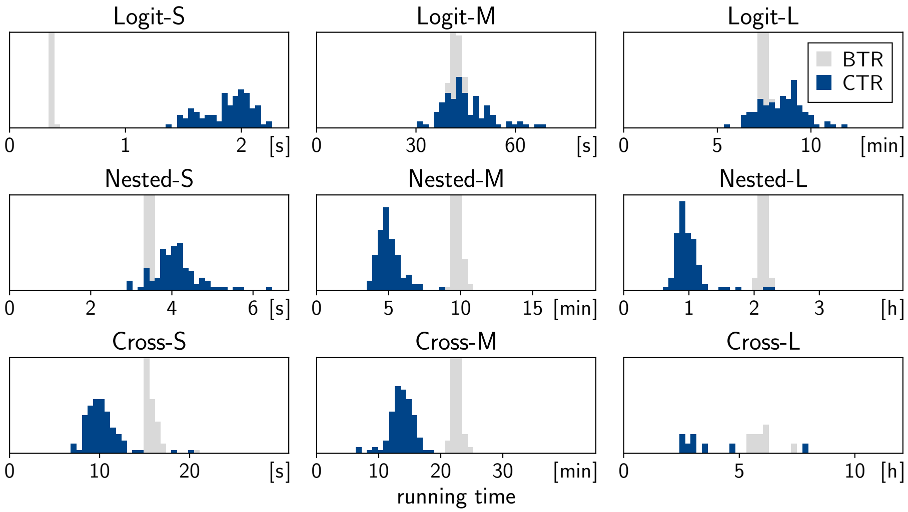
Nested-L
- 31 expl. variables.
- **101 parameters**.

Cross-S
- 10 expl. variables.
- **15 parameters**.

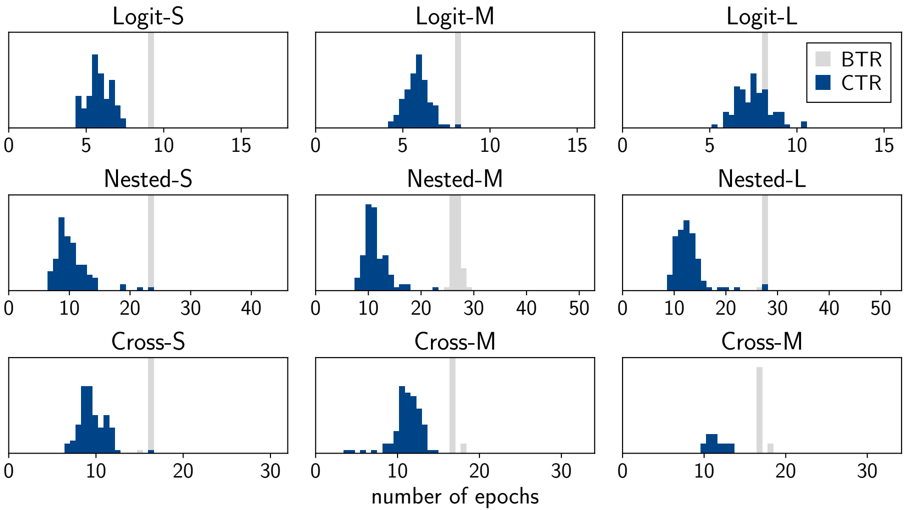
Cross-M
- 26 expl. variables.
- **55 parameters**.

Cross-L
- 31 expl. variables.
- **102 parameters**.

Preliminary results



Preliminary results



Conclusion

Summary

- Stochastic trust-region method for DCMs.
- Substantial savings for relatively complex model formulations.
- Importance of starting small.

Conclusion

Summary

- Stochastic trust-region method for DCMs.
- Substantial savings for relatively complex model formulations.
- Importance of starting small.

Future work

- Test more complex model formulations.
- Approximate the Hessian using BFGS.
- Extend to Monte Carlo simulation.

IIDE
Institut
interdisciplinaire
du Développement
de l'entreprise

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

16th Workshop on Discrete Choice Models
6–8 June 2024 | EPFL, Switzerland

Nicola Ortelli,[1,2] Matthieu de Lapparent,[1] Michel Bierlaire[2]

[1] IIDE, HEIG-VD, Switzerland
[2] TRANSP-OR, EPFL, Switzerland

References

LSH-DR

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

Direct precedents

- Bastin, F., Cirillo, C. and Toint, P. L. (2006). Application of an adaptive monte carlo algorithm to mixed logit estimation, Transportation Research Part B: Methodological 40(7): 577–593.
- Lederrey, G., Lurkin, V., Hillel, T. and Bierlaire, M. (2021). Estimation of discrete choice models with hybrid stochastic adaptive batch size algorithms, Journal of choice modelling 38.

Dataset

- Hillel, T., Elshafie, M. Z. and Jin, Y. (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).

Bucket width update