Stochastic optimization with adaptive batch size

Discrete choice models as a case study

Gael Lederrey, Virginie Lurkin,
Tim Hillel, and Michel Bierlaire

STRC | 19th Swiss Transport Research Conference
Monte Verità / Ascona, May 15-17, 2019

Outline

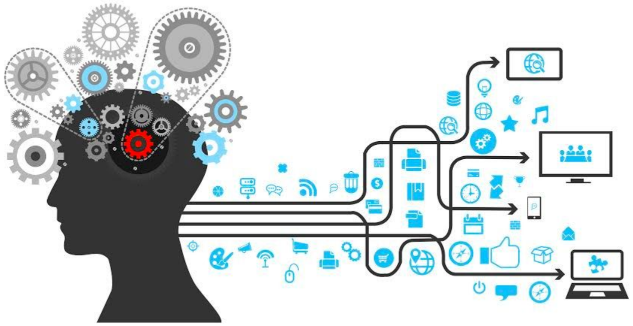
1. Motivation
2. In the literature
3. WMA-ABS
4. Results
5. Conclusion

Motivation

DCMs in a nutshell

- Many success stories until now...

Solution: Machine Learning



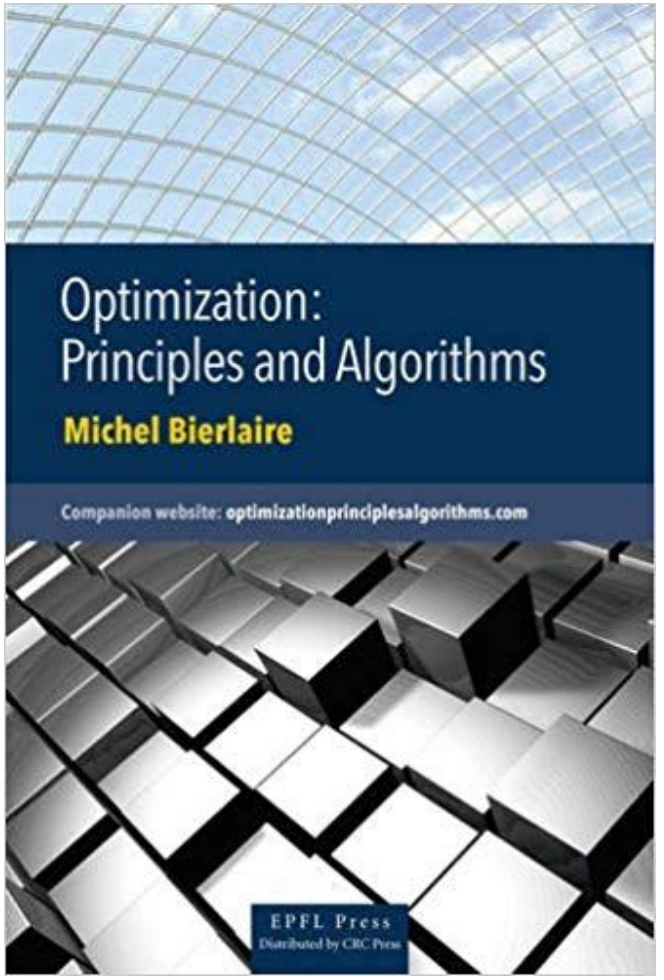
- ML is **THE** field dealing with a lot of data!

Third direction

- Mix of Choice Modeling and Machine Learning:

In the literature

Optimization of DCM and ML



Stochastic Iterative Optimization Algorithms

```
optimize():
    while stop_crit == False:
        Select a batch of data
        Compute p_k
        Compute alpha
        x_k+1 = x_k + alpha*p_k
        Check stop_crit
    return x_k+1
```

1.5th order:

BFGS

2nd Order:

$$\nabla^2 f(x_k) \cdot p_k = -\nabla f(x_k)$$

Some state-of-the-art algorithms

- 1st order:
  Adagrad (**Duchi et al., 2011**), AMSGrad (**Reddi et al., 2019**)

- 1.5th order:
  adaQN (**Keskar & Berahas, 2016**), SGD-QN (**Bordes et al., 2009**)

- 2nd order:
  Stochastic Hessian (**Byrd et al., 2011**)

WMA-ABS

Window Moving Average - Adaptive Batch Size

WMA-ABS: the algorithm

```
optimize_IOA():
    while stop_crit == False:
        Select a batch of data
        Compute p_k and alpha
        x_k+1 = x_k + alpha*p_k
        Check stop_crit
        USE WMA-ABS
    return x_k+1
```

Algorithm 2 Window Moving Average - Adaptive Batch Size (WMA-ABS)

1. Compute the WMA using previous function values → Window (10)
2. Compute the improvement using previous averages
3. Check if under threshold → Threshold (1%)
4. If yes, update counter. Otherwise, reset counter.
5. Check if counter finished → Count (1 or 2)
6. If yes, update the batch size → Factor (2)

Results

Case Study

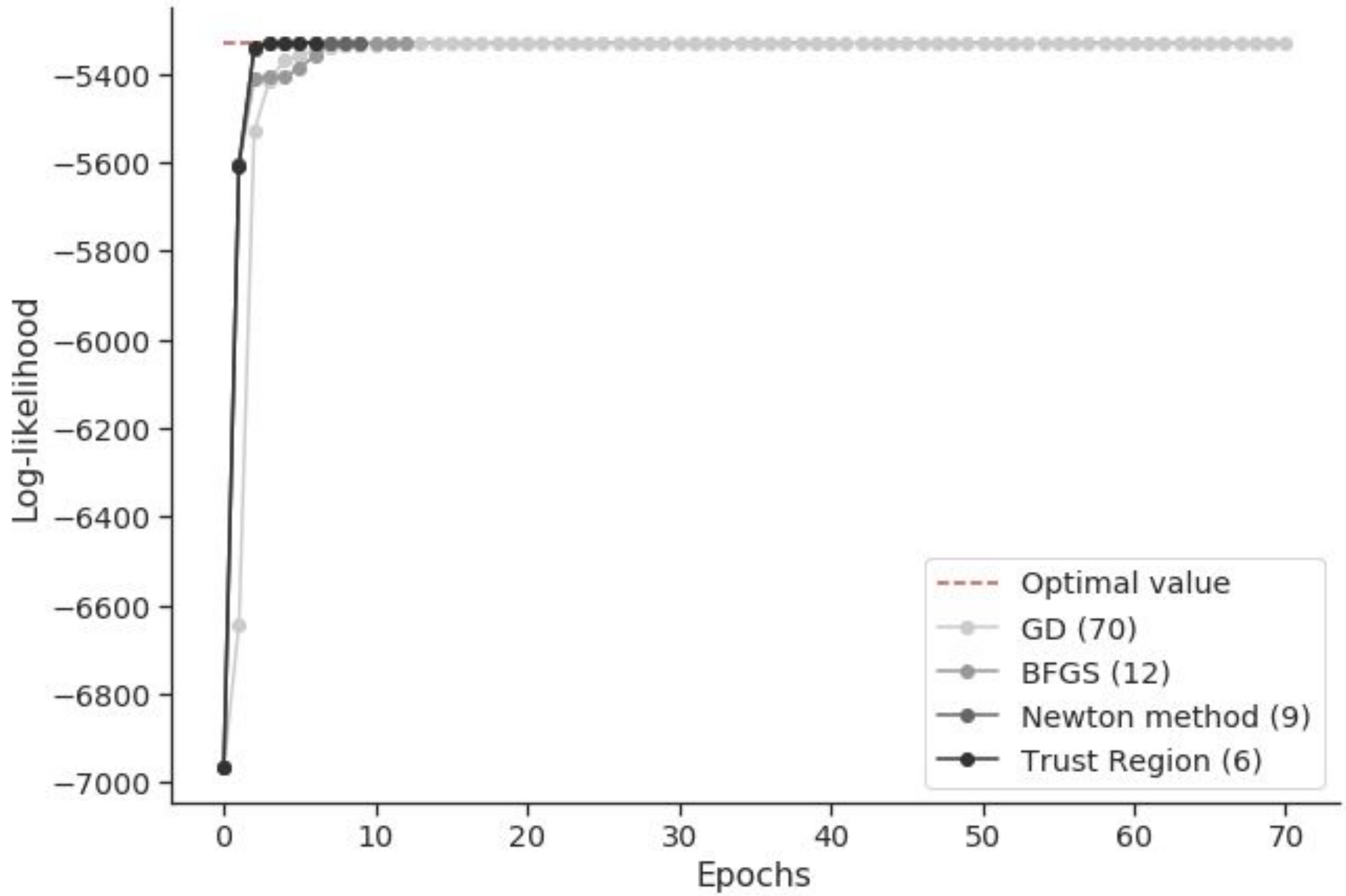
- 5 different models

Name	Type	#Parameters	Data set	#Observations
`MNL-SM`	MNL	4	Swissmetro	6’768
`Nested-SM`	Nested	5	Swissmetro	6’768
`LogReg-BS`	Logistic Regression	12	Bike Sharing	16’637
`small MNL-CLT`	MNL	100	London Passenger Mode Choice	27’478
`MNL-CLT`	MNL	100	London Passenger Mode Choice	54’766

- 4 different IOAs:

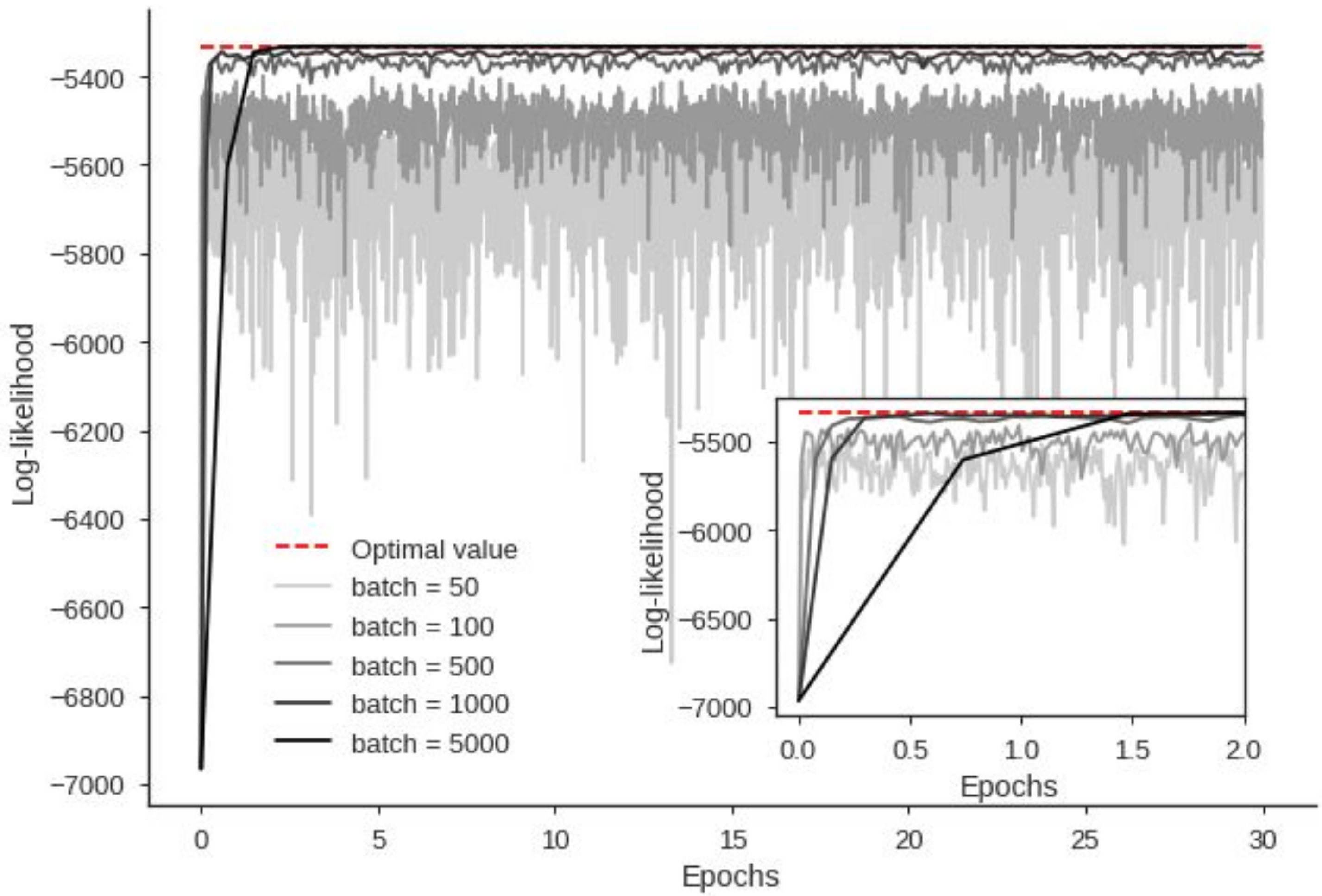
GD, BFGS, Newton Method, Trust Region

Iterative Optimization Algorithms



Optimization of model `MNL-SM`. (# of epochs until convergence)

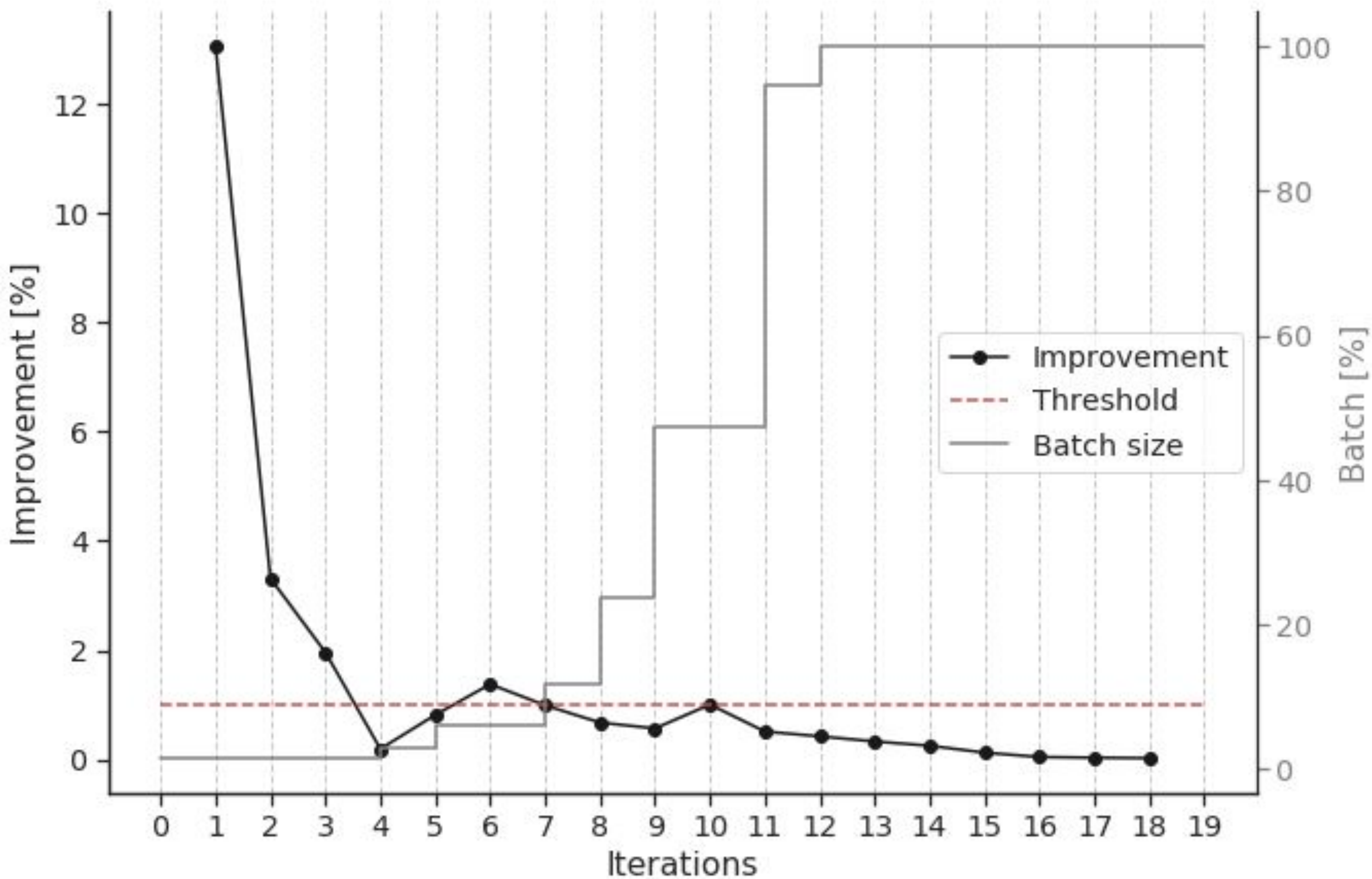
Stochastic Newton Method



Optimization of model `MNL−SM`. Lines = avg value over 10 runs.

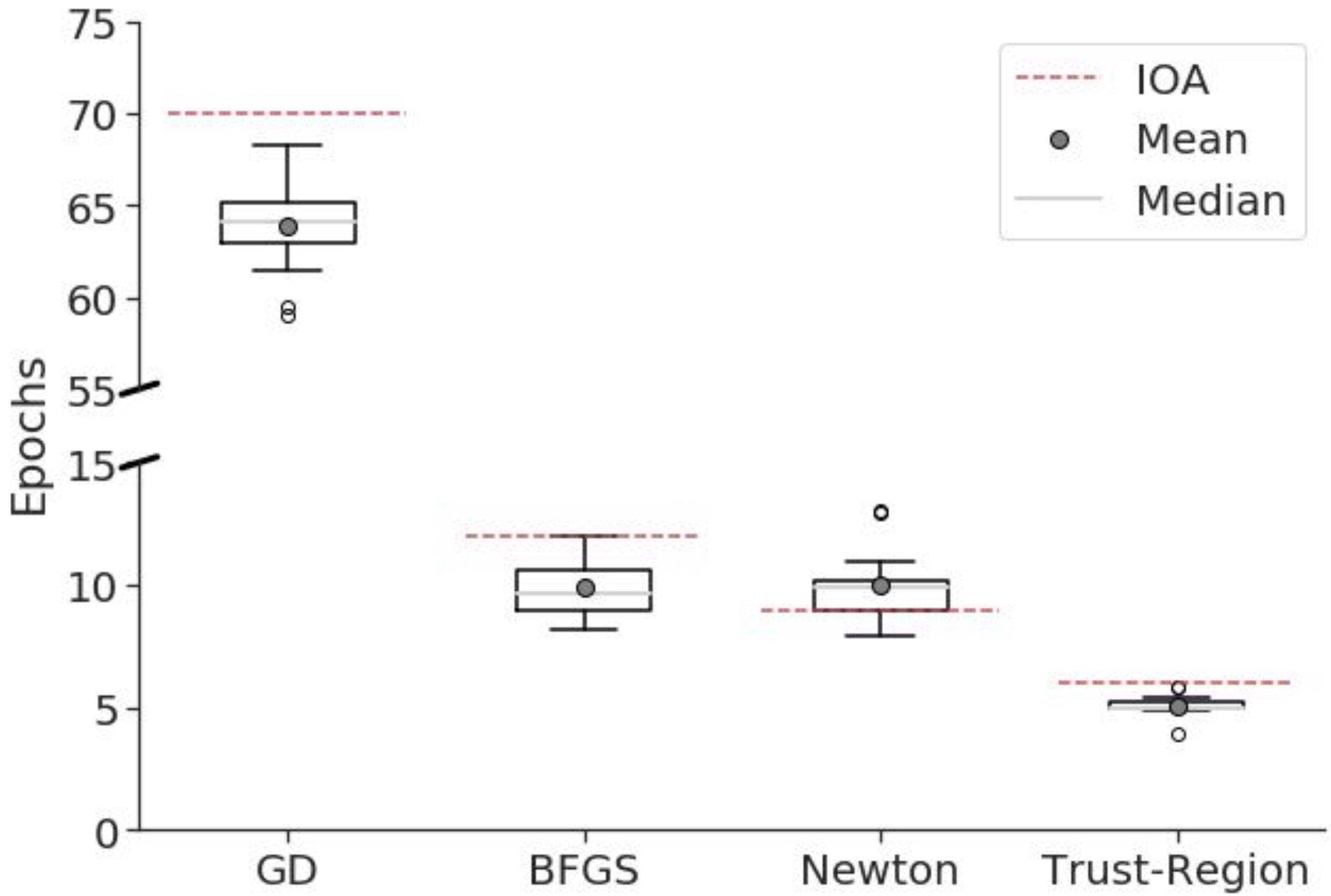
Lederrey *et al.*, 2018 - IEEE ITSC 2018

Improvements using WMA-ABS



Optimization of model `MNL-SM` with SNM and WMA-ABS.

Results of WMA-ABS (The more interesting ones!)



Optimization of model MNL−SM with WMA-ABS. (20 runs)

Results of WMA-ABS - Summary

- Stochastic IOA + WMA-ABS **<** standard IOA: **15/16**

- Improvements:

  GD: 2-10%

  BFGS: 10-20%

  NM: 0-25%

  TR: 20-54%

- bigger/complex the model => better improvements!

Parameter study

- Optimized parameters different from suggested parameters (expect for Count) …
- Often suggesting higher threshold and higher factor.
- Suggested parameters seem good though!

Conclusion

Conclusion

- WMA-ABS works with many stochastic IOAs
- Bigger improvements on complex models (**up to 55%**)
- No need to optimize the hyper parameters

Future work

- Write a paper (in progress...)
- Works on a heuristic to choose between stochastic and standard IOAs
- Implement the WMA-ABS + IOAs in  -Biogeme!

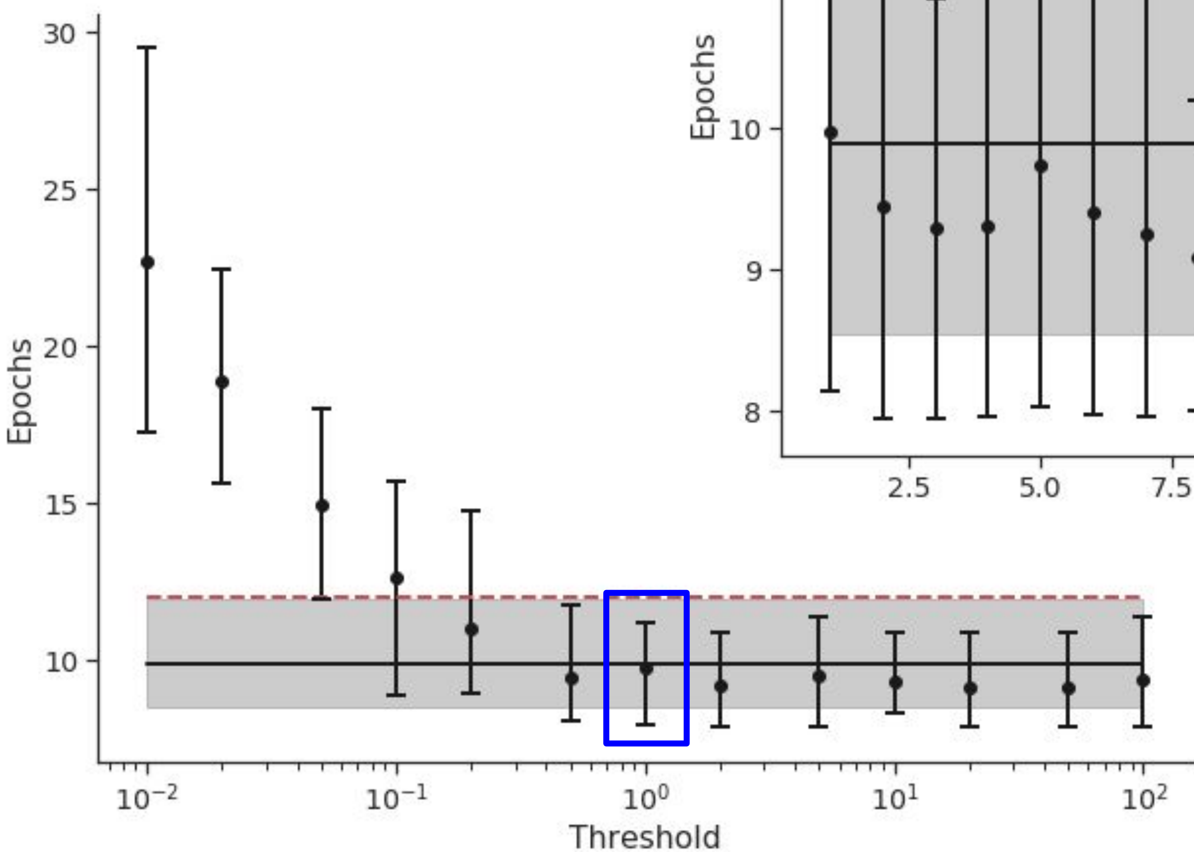

https://knowyourmeme.com/memes/improvise-adapt-overcome
Make a step. Adapt. Optimize!

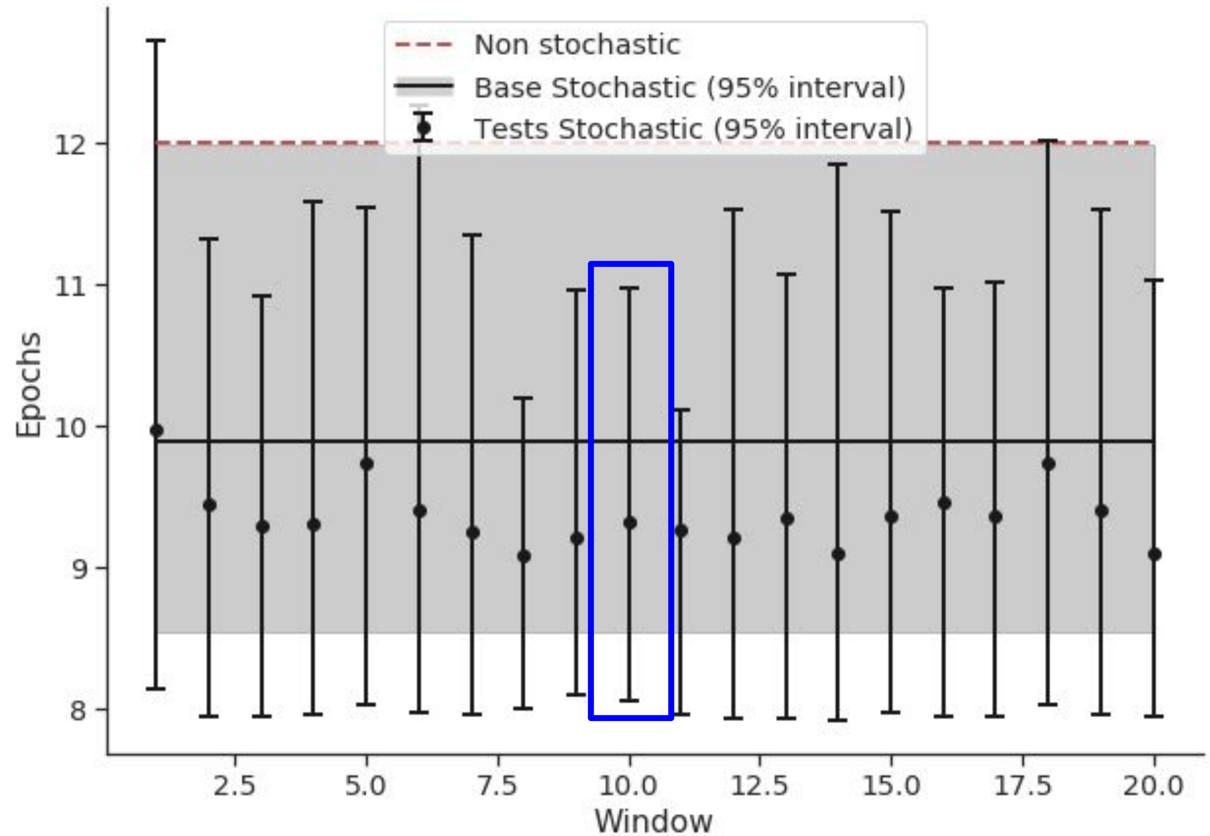
References

- **Ortelli, Rodrigues, Pereira & Bierlaire**, *Automatic Utility Specific in Discrete Choice Models*, 2019 (Master thesis; Not published yet)
- **Duchi, Hazan & Singer**, *Adaptive Subgradient Methods for Online Learning and Stochastic Optimization*, 2011
- **Reddi, Kale & Kumar**, *On the Convergence of Adam and Beyond*, 2019
- **Keskar & Berahas**, *adaQN: An Adaptive Quasi-Newton Algorithm for Training RNNs*, 2016
- **Bordes, Bottou, Gallinari, Chang & Smith**, *SGD-QN: Careful Quasi-Newton Stochastic Gradient Descent*, 2009
- **Byrd, Chin, Neveitt & Nocedal**, *On the Use of Stochastic Hessian Information in Optimization Methods for Machine Learning*, 2011
- **Balles, Romero & Hennig**, *Coupling Adaptive Batch Sizes with Learning Rates*, 2016
- **Devarakonda, Naumov & Garland**, *AdaBatch: Adaptive Batch Sizes for Training Deep Neural Networks*, 2017

Backup slides

Effects of the parameters (1/2)





Effects of the parameters (2/2)