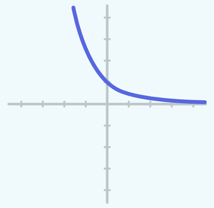
"How Much Data is Enough?" Learning Curves for Machine Learning

Investigating alternatives to the Levenberg-Marquardt algorithm for learning curve extrapolation.

By Lucian Negru (L.Negru@student.tudelft.nl)

Supervisors Dr. Jesse Krijthe (J.H.Krijthe@tudelft.nl)
Dr. Tom Viering (t.j.viering@tudelft.nl)

1. Learning Curves and the LCDB



Learning curves display the **error rate** of an ML model with respect to the **size of the training set**. The Learning Curve Database (**LCDB**) [1] contains a large number of these Learning Curves. It showcases how these can have various shapes.

2. The Research Goal

The research goal, *investigating alternatives to the Levenberg-Marquardt algorithm for learning curve extrapolation*, considers 2 alternative algorithms to **Levenberg-Marquardt** (LM) — **Gradient Descent** and **BFGS**.

- Is Gradient Descent a suitable alternative to LM?
- Is BFGS a suitable alternative to LM?

3. The Algorithms

- Levenberg-Marquardt: a method which implements functionality of Gradient Descent and Newton's method [2].
- Gradient Descent: a method which initially gets close to the optimum fit very quickly. Widely used in machine learning [3].
- BFGS: a method which improves its gradient calculations without the need of matrix transformations, decreasing its complexity [4].

4. Methodology

Sampling 10,000 learning curves with the maximum number of training anchors, we:

Implement Gradient Descent and BFGS, then run them on the sample, comparing the results to LM.

→ Optimise Gradient Descent by testing different step sizes and changing the tolerance of termination, then run it on the samples.

→ Optimise BFGS by changing the calculation of the gradient from a numerical one to an analytical one, then run it on the samples.

5. Results

The following results show the performance of the 2 alternative algorithms compared to the original LM implementation. The graphs illustrate performance before and after optimisation.

Performance is defined by the mean-squared error (MSE) on the testing anchors, average MSE of each parametric model, and computation time.

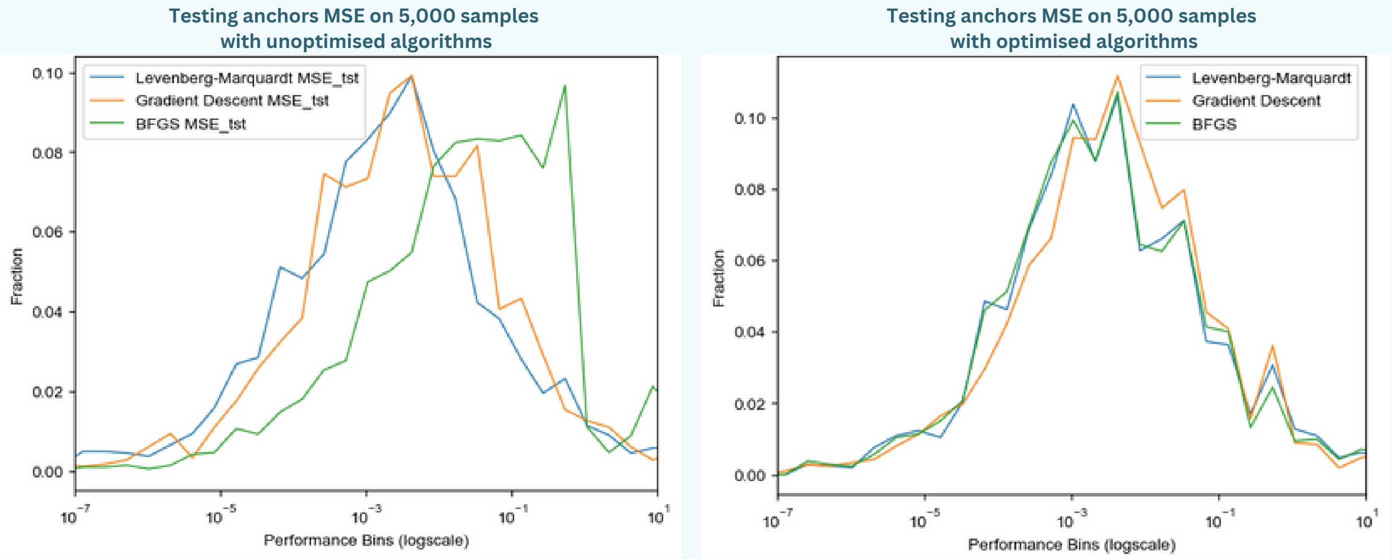


Figure 1: The testing anchors MSE of the three algorithms prior to (left) and after (right) optimisation. Displays the distribution (Y-axis) of the errors along different values of MSE (X-axis). BFGS can be seen improving after optimisation, but Gradient Descent showed. trade-off between accuracy and computation time. (Right-sided distribution is worse).

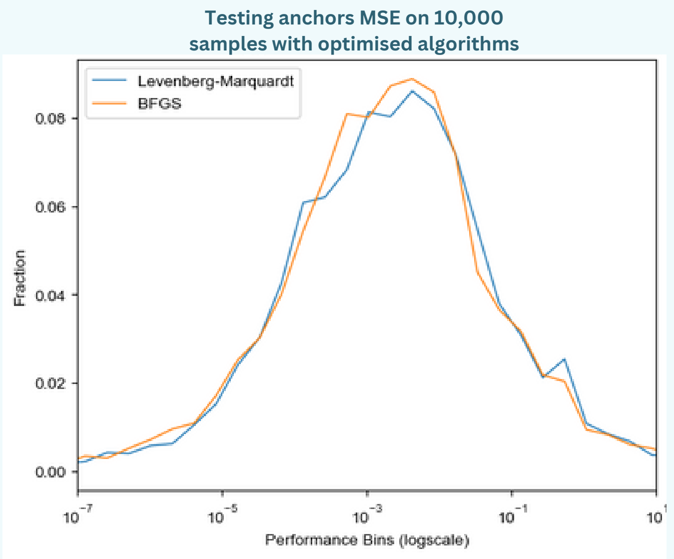


Figure 2: The Testing anchor MSE of LM and BFGS optimised. **Note**: Gradient Descent performance on 10,000 samples is not available as the Jupyter Notebook crashes from the number of iterations.

The performance of the individual parametric models follows closely between LM and BFGS (Table 1). Gradient Descent had a trade-off between MSE accuracy and computation time (Table 2).

Parametric Model Performance

Levenberg-Marquardt		Gradient Descent		BFGS	
Curve Model	MSE	Curve Model	MSE	Curve Model	MSE
last1	0.005832	last1	0.005832	last1	0.005832
wbl4	0.092318	wbl4	0.586516	wbl4	0.120457
exp4	0.094922	exp4	0.693752	exp4	0.094035
mmf4	0.113522	mmf4	0.846390	mmf4	0.135952
pow4	0.130181	pow4	1.136965	logpower3	0.207976

Table 1: Average testing anchor MSE of the 5 top-performing parametric model of every algorithm on 5,000 samples. Results include optimisation.

Computation Time

Algorithm	Time on sample size		
	128 size	5,000 size	10,000 size
Levenberg-Marquardt	4.73 s	88.86 s	477.74 s
Gradient Descent	64.00 s	593.09 s	N/A
BFGS	3.88 s	66.98 s	406.69 s

Table 2: Computation time of the algorithms on various sample sizes.
Note: Gradient Descent time on 10,000 samples is not available as the Jupyter Notebook crashes from the number of iterations.

7. Conclusions and Further Research

- Gradient Descent is not a suitable alternative; performance is similar but the trade-off with computation time makes it impractical.
- BFGS is a suitable alternative, having nearly identical MSE accuracy while being faster.
- BFGS shows narrower distribution of MSE than LM.
- Parametric models may be more important than the fitting algorithms.

Further research:

- Investigate the significance of the narrower distribution of BFGS.
- More parametric models can be developed.

6. Limitations

1. Parameters are initialised randomly.
2. Not the whole LCDB was tested.
3. Samples of learning curves with fewer training anchors were not considered.

[1] Felix Mohr, Tom J Viering, Marco Loog, and Jan N van Rijn. Lcdb 1.0: An extensive learning curves database for classification tasks. In *Machine Learning and Knowledge Discovery in Databases. Research Track - European Conference, ECML PKDD 2022, Grenoble, France, September 19-24, 2022*, 2022.
[2] Manolis IA Lourakis et al. A brief description of the levenberg-marquardt algorithm implemented by levmar. *Foundation of Research and Technology*, 4(1):1–6, 2005.
[3] Haskell B Curry. The method of steepest descent for non-linear minimization problems. *Quarterly of Applied Mathematics*, 2(3):258–261, 1944.
[4] E Dennis Jr and Robert B Schnabel. Secant methods for unconstrained minimization. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Englewood Cliffs, NJ: Prentice-Hall*, pages 194–215, 1983.