The Journal of Open Source Software

LCPP: Learning Curve Plus Plus

Ozgur Taylan Turan [1¶] **and David M. J. Tax**[1]

1 Delft University of Technology, The Netherlands ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 30 September 2025
Published: unpublished

Summary

A learning algorithm is said to learn if its performance on a given task improves with experience (Mitchell, 2013). This fundamental definition links the size of training data to the generalization performance of the model. In supervised learning, a **learning curve** depicts how generalization performance evolves as a function of the training set size. Collections of such data, known as **learning curve databases**, track the performance of diverse machine learning algorithms (learners) across multiple tasks as they observe increasing amounts of training data.

Learning curve databases are valuable for **model selection** and for **estimating the amount of data needed** to achieve a target performance. These applications typically assume learning curves are monotonic and convex. However, findings of (Yan et al., 2025), (Mohr et al., 2023) and (Viering & Loog, 2022) suggest that learning curves often exhibit more complex and irregular behavior. Sparse sampling of training sizes limits the ability to fully characterize these behaviors, highlighting the need for **high-fidelity learning curves** to investigate them.

LCPP (Learning Curve Plus Plus) is a C++ library that allows for learning curve creation of machine learning models. LCPP enables its users to obtain learning curves for variety of learners on any supervised learning dataset with/out hyper-parameter tuning, enabling model selection and training data requirement determination.

Statement of need

Generally, creating learning curves is computationally expensive because it requires repeatedly training algorithms on many subsets of varying training sizes. Consequently, learning curves are often computed for a limited number of training set sizes. For example, while creating learning curve databases (Mohr et al., 2023) and (Yan et al., 2025) limited number of training set-sizes are investigated, moreover, these generations are done only for fixed learners without hyper-parameter tuning.

To empower the machine learning community to generate richer, more detailed learning curves, we propose LCPP, a C++ library for scalable learning curve generation. LCPP offers several features; first several approaches for splitting a given dataset into training and test sets of varying sizes (where training sets can be drawn randomly or incrementally, where test sets can be fixed or vary in size). Next, unlike most existing tools that fix hyper-parameters during learning curve creation, LCPP integrates hyperparameter optimization routines from mlpack, enabling more realistic and optimized learner evaluations.

LCPP also includes a simple dataset container for access to OpenML datasets (Vanschoren et al., 2013), with built-in support for complete dataset transformations and train/test splits, allowing users to directly measure the generalization performance of models available in mlpack and some other learning algorithms included in itself, such as kernel ridge regression, discriminant classifiers, multi-class classification extensions of binary classifiers.

Designed for easy deployment on high-performance computing (HPC) environments, LCPP

can efficiently run large-scale experiments in parallel, ensuring reproducibility and scalability. This combination of features has already been demonstrated in (Turan, Tax, et al., 2025) and (Turan, Loog, et al., 2025), where large-scale learning curve databases are generated.

By capturing generalization performance across many learners and tasks, LCPP facilitates systematic benchmarking, fair algorithm comparisons, and meta-analysis for understanding broader patterns in learning behaviors of machine learning models.

References

Mitchell, T. M. (2013). *Machine learning* (Nachdr.). McGraw-Hill. ISBN: 978-0-07-115467-3

Mohr, F., Viering, T. J., Loog, M., & van Rijn, J. N. (2023). LCDB 1.0: An extensive learning curves database for classification tasks. In M.-R. Amini, S. Canu, A. Fischer, T. Guns, P. Kralj Novak, & G. Tsoumakas (Eds.), *Machine learning and knowledge discovery in databases* (pp. 3–19). Springer Nature Switzerland. ISBN: 978-3-031-26419-1

Turan, O. T., Loog, M., & Tax, D. M. J. (2025). *Generalization performance distributions along learning curves*.

Turan, O. T., Tax, D. M. J., Viering, T. J., & Loog, M. (2025). Learning Learning Curves. *Pattern Analysis and Applications*, *28*(1), 15. https://doi.org/10.1007/s10044-024-01394-6

Vanschoren, J., Rijn, J. N. van, Bischl, B., & Torgo, L. (2013). OpenML: Networked science in machine learning. *SIGKDD Explorations*, *15*(2), 49–60. https://doi.org/10.1145/2641190.2641198

Viering, T., & Loog, M. (2022). *The shape of learning curves: A review*. https://arxiv.org/abs/2103.10948

Yan, C., Mohr, F., & Viering, T. (2025). *LCDB 1.1: A database illustrating learning curves are more ill-behaved than previously thought*. https://arxiv.org/abs/2505.15657