What is the effect of Gaussian filtering applied before curve fitting?

Ionut-Liviu Moanta[1]
Responsible Professor: Tom Viering[1]
Supervisors: Cheng Yan[1]**, Taylan Turan**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 26, 2025

Name of the student: Ionut-Liviu Moanta
Final project course: CSE3000 Research Project
Thesis committee: Tom Viering, Taylan Turan, Cheng Yan, Arie van Deursen

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

Learning curves are graphical representations of the relationship between dataset size and error rate in machine learning. Curve fitting is the process of estimating a learning curve using a mathematical formula. This paper analyzes two ways of performing curve fitting: interpolation and extrapolation. The accuracy of the curve-fitting procedure might be negatively influenced by the irregular shape of the curve and the presence of noise. Our study investigates the effects of the Gaussian filter on curve fitting and the potential to improve its performance. This is done by analyzing multiple values of the Gaussian filter's standard deviation parameter(Sigma) and also a wide variety of learning curves(both smooth and noisy ones). The main finding of this research states that the Gaussian filter can generate significant improvements in the extrapolation process, especially when it is applied to noisy curves. On the other hand, for the interpolation procedure, its impact is reduced, even negligible for smooth curves. An important takeaway from this paper is that selecting the most suitable pre-processing method for the type of curve analyzed might generate valuable findings in the field of learning curves used in machine learning.

1 Introduction

Learning curves are study plots that analyze and interpret the correlation between the training dataset size and the corresponding error rate in machine learning. By understanding this relationship, learning curves provide useful insights in achieving the general purpose of answering the question of how much data we need to acquire certain performance levels[13] since collecting data takes time and money [9]. In particular, these curves are highly useful in determining the efficiency of a machine learning model in shifting from the training to the validation and testing processes by helping the analysts understand whether the model is overfitting(it performs well on the watched data but poorly on the inconspicuous data)[16].

Curve fitting is an essential method utilized for interpreting learning curves, where a mathematical formula is used to estimate the relationship between training dataset size and error rate. In our particular case study, we start with a set of points provided by a learner trained on a dataset, and the curve fitting is performed based on these observed points. By fitting a curve to the observed data points, researchers can model the underlying trend, making it easier to analyse and predict gaps both within and beyond those observed points. Curve fitting serves as the foundation for further tasks such as interpolation and extrapolation [5].

Two crucial aspects related to learning curves which will be analyzed during this research are interpolation and extrapolation. Interpolation is the process of estimating the model's performance based on the intermediate dataset size values falling within the range of observed points[12]. In contrast, extrapolation is the process of estimating the model's performance for larger dataset sizes. Both interpolation and extrapolation are highly useful in the decision-making process which determines the most suitable dataset sizes needed to gain a predefined performance or error rate.

However, both interpolation and extrapolation, as estimations, are not perfect and their performance is measured by calculating a mean squared error(MSE). In the case of a noisy learning curve, the fitting process, along with interpolation/extrapolation, might generate inaccurate results thus altering the whole prediction process. One proposed solution for solving this issue is applying the Gaussian filter to the raw points of the learning curve. Afterwards, we will analyze its impact by calculating the improvement provided to the mean squared errors of both the interpolation and the extrapolation.

An initial guess which determines us to choose the Gaussian filter as a main preprocessing method applied to learning curves is its ability to reduce the noise of the initial data. In our case, this might lead to smoother slopes[11] that better capture the true shape of the curve during the fitting process. The performance of this technique also depends on the Gaussian filter's hyperparameter, Sigma, which represents the standard deviation of the kernel and determines the degree of smoothing applied to the curve. Therefore, we will find the impact of applying the Gaussian filter in improving the performance of interpolation and extrapolation. In machine learning, this can be helpful to acquire the most reliable dataset sizes which provide the required performance of the models.

The main research question of this paper is "What is the impact of the Gaussian filter applied to the initial set of points before curve fitting for both interpolation and extrapolation?". In order to make the path of the research clearer, it was divided into the following 2 sub-questions:

1. **RQ1:** Under which values of Sigma does the Gaussian filter improve the MSE of the interpolation/extrapolation fitting process?
2. **RQ2:** Under which conditions (type of the learning curve - presence of noise) does the Gaussian filter improve the MSE of the interpolation/extrapolation fitting process?

The remainder of this paper is structured as follows: Section 2 presents our research using the necessary information extracted from previous scientific resources; Section 3 highlights the whole process designed by us in order to create a suitable methodology reflected in the code-based environment; Section 4 reveals the results discovered by our study related to the initial main research question and its sub-questions; Section 5 analyses the results presented in the above-mentioned section by highlighting the advantages and disadvantages of applying the Gaussian filter; Section 6 underscores this research's ethical principles of transparency and reproducibility; Section 7 presents the potential areas of improvement that we suggest for further research; Section 8 presents a summary of the entire research process, focusing on the results obtained and their potential impact.

2 Related Works

The two starting points without which this research would have been impossible are the scientific articles by Mohr et al.[14] and the one by Viering and Loog[19]. The first one introduces the Learning Curves Database(LCDB), an essential tool which provides us with a wide range of 246 datasets and 20 learners. A specific learner trained on a specific dataset generates a learning curve which will be used as an input to conduct our experiments throughout this project. Moreover, this article proves the benefits of LCDB by explaining how it contributes to examining learning curves' properties(monotonicity, convexity, and peaking) and to performing the curve-fitting process using both interpolation and extrapolation. The paper also leaves several open questions, such as whether smoothing learning curves aids in model selection, how meta-features can predict curve behaviour, and which non-parametric extrapolation techniques are most effective.

Now that we are familiar with how a learning curve can be generated and what it looks like, the second article [19] explains the significance of the learning curves and the ways of estimating them. It also presents the parametric learning curve models which are useful for the fitting process. Furthermore, Viering and Loog provide an in-depth explanation of ill-behaved and well-behaved learning curves and therefore helped us to understand that

the learning curves might take different shapes from case to case. The paper concludes by highlighting the need for consistent use of learning curves in research in order to improve algorithm evaluation and guide future studies.

Furthermore, the book by Achim Zielesny[22] played a significant role in offering us all the necessary information related to curve fitting. It starts from illustrative examples, such as that presented in Figure 1 and continues with evaluating the goodness of a fit, guessing the most suitable model function and finally with problems and pitfalls of the fitting process.

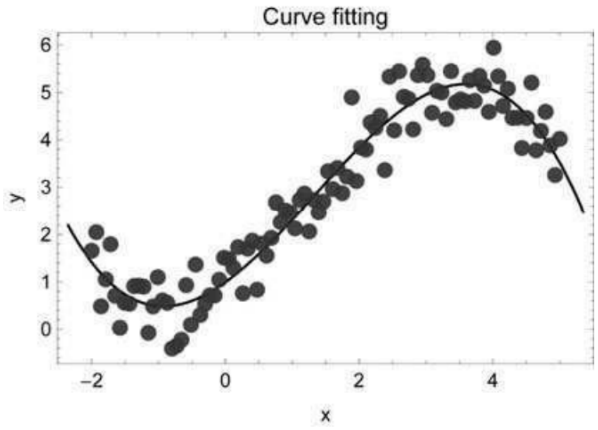

Figure 1: Concrete example of curve fitting [22]

The concepts of interpolation and extrapolation are described with formal definitions, examples and formulas by Muhammad Abdul Wahab[21]. A reliable plot presenting both interpolation and extrapolation extracted from his paper can be seen in Figure 2. In addition, Muhammad Abdul Wahab presents multiple ways of performing interpolation, supported by formulas and related Matlab functions, and their practical usages in Computer Graphics and Image Processing.

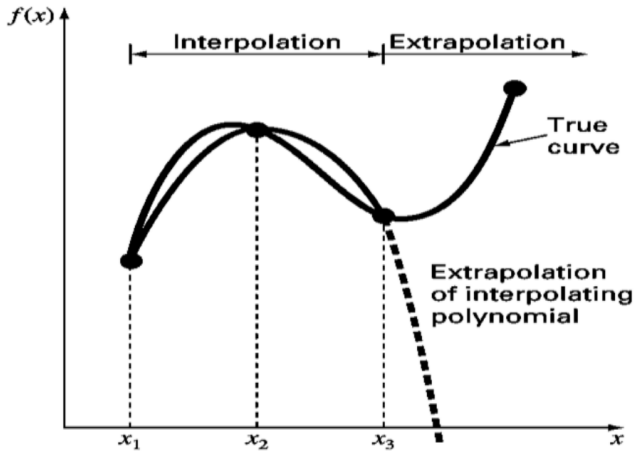

Figure 2: Concrete example of interpolation and extrapolation [21]

In addition, after dealing with the concepts of interpolation and extrapolation as ways of performing curve fitting, Mohr et al.[14] introduce the metric of Mean Squared Error(MSE) as a way of evaluating the performance of all these processes.

Last but not least, the whole research idea of this project evolves around the application of the Gaussian filter (with a well-determined standard deviation hyperparameter Sigma, as explained in Section 4) in order to obtain some benefits in the curve fitting process and thus improve its performance. Starnes and Webster[18] present in their article the enhanced efficiency generated by applying Gaussian smoothing to Stochastic Gradients. The benefits of the Gaussian filter highlighted in the above-mentioned paper include attenuating the fluctuations, boosting robustness to noise and improving generalization. As a result, their research represented a strong motivation to choose the Gaussian filter as the main proposed improvement strategy in our case and analyze its impact on all types of learning curves(both smooth and noisy ones).

3 Methology and Experimental Setup

The research plan was divided into six key steps to systematically address the initial question proposed by the academic staff which supervises the progress:

Software tools: The starting point of the whole methodology process was determined by creating a software environment which best addresses our desires - as a programming language we chose Python due to its ability to help us model and solve scientific problems[20] by a wide range of libraries which support the easy implementation of useful methods. As a way of developing and running the code, the best solution was the Jupyter Notebook, which ensures the reproducibility of our scientific results[1] and helps us keep code and research output together. In terms of libraries and imports used, the most significant ones are numpy for numerical computations and matplotlib.pyplot for creating visualizations together with the imports necessary for performing curve fitting(scipy.optimize.curve_fit), applying the Gaussian filter (scipy.ndimage.gaussian_filter) or calculating the Mean Squared Error(sklearn.metrics.mean_squared_error).

Framework: The second step consisted of developing a framework which enables us to import, interpret and analyze the learning curves. In the initial version, it only permitted to visualize a raw learning curve generated by selecting a dataset index and a learner index from the Learning Curve Database(LCDB) mentioned in Section 2.

Gaussian filter: After finding a learning curve that needs further investigation, the following step involves applying the Gaussian filter to the raw data of this curve. Consequently, the initial version of the framework accommodated the addition of the Gaussian filter functionality. In order to better analyze the impact of the smoothing, we decided to apply the filter using 4 values of its standard deviation hyperparameter Sigma(1.0, 2.5, 5.0, 10.0). A concrete example which presents the progress until this point can be seen in Figure 3. As a short explanation, the initial shape of the learning curve can be noticed in purple and its change after applying the Gaussian with Sigma 1.0, 2.5, 5.0 and 10.0 in blue, orange, green and red respectively. We remark that a larger sigma smooths more drastically but may alter the initial shape of the curve, while a smaller sigma aims to retain it but may leave some of the noise unfiltered.

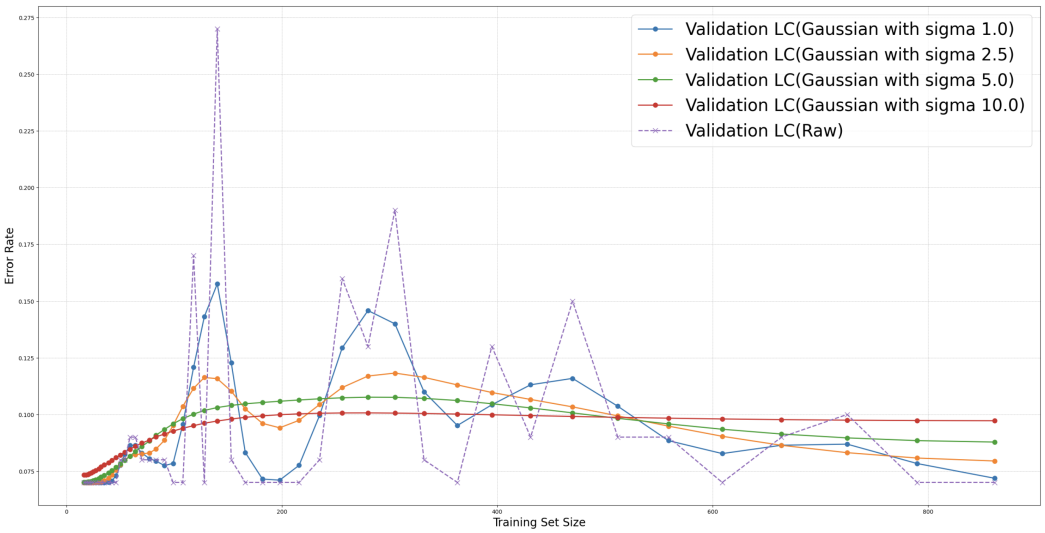

Figure 3: Concrete example of a learning curve and its transformation after applying the Gaussian filter with different values of Sigma

Curve fitting: After we identified the set of points representing the learning curve or the filtered version of them (after applying the Gaussian), the next step consists of performing the curve fitting by selecting a percentage of points which are passed to the selected model in order to generate the best possible fitted curve which suits the respective model's formula. Within the scope of this research, this fitting percentage was selected to be 80 per cent, since it is a simple and stable technique used for curve fitting[3]. The fitting model was chosen to be pow3. This model is considered to offer a good fit in many situations[19] and has the formula $pow3(x) = a - b \cdot x^{-c}$, where a, b and c are parameters selected by the curve fitting algorithm.

Interpolation and Extrapolation evaluation metric: Furthermore, we have two strategies already mentioned: interpolation and extrapolation. The next aspect that needs clarification for the methodology stage is how exactly we evaluate an interpolation/extrapolation of the curve. This is done using the Mean Squared Error(MSE) using the points seen / not seen points during the fitting for interpolation and extrapolation respectively [14]. This method is considered a fair measure of a fitting's performance[7] and has two main advantages in our case study: it provides a penalty for large errors and is useful for quantitative data analysis.

Mann-Whitney U test: To answer the research question, we will propose reliable hypotheses and test their assumptions by conducting experiments with the datasets and learners provided. The significance of the Gaussian filter applied to the learning curves will be measured by collecting two sets of mean squared errors(one generated from the fitting using the original curves and the second one generated from the fitting using the filtered curves). Afterwards, we will use a Mann-Whitney U test [4] which checks if the two samples come from the same distribution. It also provides a significance value which guides us in determining the impact of the Gaussian filter, as mentioned in Section 4. As a methodology

discussion, it is worth mentioning that the Mann-Whitney U test was preferred to the T-Test [10], because the second one requires that the input data follows the normal distribution, which was not suitable for our mean squared errors samples.

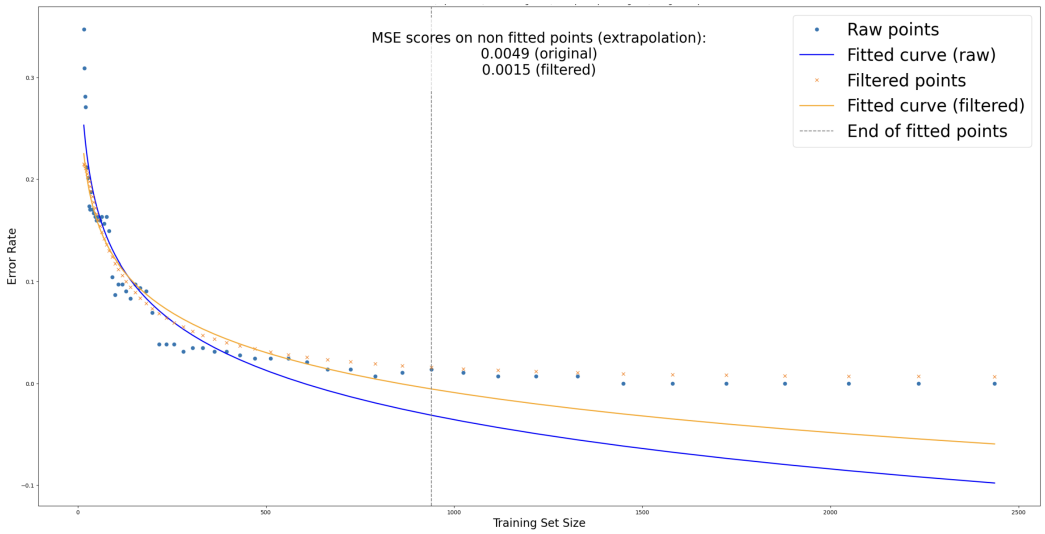

Figure 4: Extrapolation of a raw LC VS Extrapolation of a fitted LC

The plot which summarizes this section can be seen in Figure 4. As a short explanation, it analyzes how the Gaussian filter with Sigma = 10 impacts the fitting process. In blue one can see the original points of the learning curve and what a learning curve fitted to them looks like. In orange, we plot the filtered points of the curve and also the fitted curve obtained from them. Then, the evaluation process was conducted as mentioned above and we calculated the MSE using the original points, resulting in a value of 0.0049 and also using the filtered points leading to an MSE value of 0.0015. The next section, which highlights all the code-based experiments, aims to answer the question of whether the improvement (in terms of lowering the MSE) produced by applying the Gaussian smoothing is significant.

4 Experiments and Results

In this section, we are going to reveal the experiments conducted and the hypotheses initiated and further validated/invalidated by the research process. The section will be split based on the two main sub-questions mentioned in Section 1, as follows:

4.1 Under which values of Sigma does the Gaussian filter improve the MSE of the interpolation/extrapolation fitting process?

For the first sub-question, we designed two experiments to find out the best possible values for the standard deviation parameter of the Gaussian filter when applied for interpolation and extrapolation processes respectively. The general strategy used for both experiments was to perform a search using powers of 2 from the set S = {0.125, 0.25, 0.5, 1.0, 2.0, 4.0, 8.0, 16.0, 32.0, 64.0} as potential values for Sigma. In this way, we cover a wide range of

numbers by performing a limited amount of calculations, obtaining a logarithmic execution time for our algorithm. To continue, we calculate the average MSE of the fitting process across all curves corresponding to each combination of LCDB learner and Sigma value in the set S. For each learner, we identify the Sigma with the lowest average MSE. The Sigma that performs best for the most learners will be declared the winner.

Learner	**Best Sigma for Interpolation**	**Best Sigma for Extrapolation**
SVC_linear	2.0	16.0
SVC_poly	2.0	16.0
SVC_rbf	2.0	16.0
SVC_sigmoid	2.0	16.0
DecisionTree	4.0	16.0
ExtraTree	0.5	16.0
LogisticRegression	4.0	16.0
PassiveAggressive	1.0	16.0
Perceptron	2.0	16.0
RidgeClassifier	1.0	16.0
SGDClassifier	2.0	16.0
MLPClassifier	2.0	16.0
LDA	2.0	16.0
BernoulliNB	1.0	16.0
GaussianNB	2.0	16.0
KNN	4.0	16.0
NearestCentroid	4.0	16.0
ens.ExtraTrees	2.0	16.0
ens.RandomForest	0.125	16.0
ens.GradientBoosting	4.0	16.0
DummyClassifier	1.0	16.0

Table 1: Comparison of learners based on the best Sigma Values obtained for the interpolation and extrapolation processes.

Best Sigma for Interpolation: From Table 1, it becomes clear that the best Sigma Value for the interpolation has the value of 2, as for 10 out of 21 learners it provides the lowest average Mean Squared Error.

Best Sigma for Extrapolation: From Table 1, it becomes clear that the best Sigma Value for the extrapolation has the value of 16, as for all 21 out of 21 learners it provides the lowest average Mean Squared Error.

Before moving on to the next subsection, we would like to point out that these Sigma values will be used for all the following experiments during this project. So, for the interpolation process, we will always use Sigma = 2 and for extrapolation Sigma = 16.

4.2 Under which conditions (type of the learning curve - presence of noise) does the Gaussian filter improve the MSE of the interpolation/extrapolation fitting process?

This subsection will highlight four experiments presenting the impact of the Gaussian filter on the curve-fitting process(both interpolation and extrapolation) for the two corresponding

types of learning curves(noisy and smooth ones). Before presenting the concrete experiments and their results, an important question that arises is: "How can we distinguish a noisy learning curve from a smooth one?". In order to solve this problem, we developed an algorithm which measures the noisiness of the curves by computing the Median Absolute Deviation (MAD) [8] of all the slopes which compose the actual curve. As a way of interpreting the results of this algorithm, we would like to mention that a high MAD indicates significant variability in the slopes, which is characteristic of a noisy curve with irregular changes. On the other hand, a low MAD reflects consistency in the slopes, indicating a smoother curve.

Using the MAD values of all the learning curves, we performed a sorting algorithm in order to determine the noisiest and the smoothest learning curves. We selected the first 60 and the last 60 learning curves from this sorted array as the inputs for the upcoming experiments as the 60 noisiest and the 60 smoothest learning curves from LCDB.

For each experiment, we will present two ways of interpreting the data. The first one consists of plotting a histogram that visually showcases the differences between the Mean Squared Errors generated by the original curves in blue versus the Mean Squared Errors generated by the filtered curves in red. The bar charts are designed to plot the correlation between the MSES values from a range and the frequency of them in both the original and the filtered sets of errors. The second one is defined by the results generated by a Mann-Whitney U test[15]. It compares the two groups of mean squared errors(original vs filtered) under the null hypothesis that they follow the same distribution. The alternative hypothesis states that the two groups differ in distribution (e.g., medians are different). The way of validating one of the two hypotheses is by comparing the p-value generated by the test with the threshold of 0.05. If it is greater than the threshold, we validate the null hypothesis. Otherwise, we validate the alternative one. For each of the four experiments presented below, we will analyze initially the corresponding histogram and then will validate the visual commitments by conducting a Mann-Whitney U test.

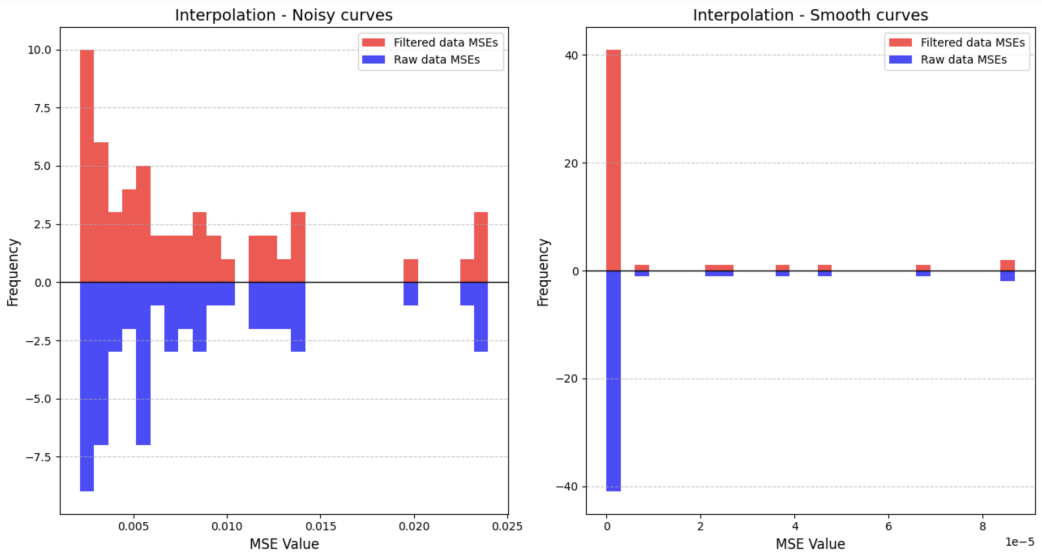

Figure 5: The impact of the Gaussian filter on the interpolation process

The impact of the Gaussian filter on the interpolation process: The histogram presented in the left part of Figure 5 indicates a high probability that the Gaussian filter

does not generate a significant performance improvement in the case of interpolation using noisy curves, since for a lot of intervals the frequency of original and filtered Mean Squared Errors is equal and for the other ones it mostly alternates with one interval dominated by Blue(original MSEs), the next one by red(filtered MSES) and so on. This assumption is further approved after performing a Mann-Whitney U test which provides us a p-value of 0.90. Therefore, the impact of the Gaussian filter in this case is considered insignificant due to the above-mentioned aspects. The bar chart offered by the right side of Figure 5 strongly suggests that the Gaussian filter has basically no impact on the Mean Squared Errors of the interpolation process applied on smooth curves since for all ranges of Mean Squared Errors the frequency of original MSES is equal to the frequency of "filtered" MSES. The Mann-Whitney U test applied in this case generates the same conclusion with a p-value of 0.97. This extremely high p-value (1.0 is the maximum possible value) strongly supports the conclusion that the Gaussian filter has no noticeable impact on the interpolation process using smooth curves.

The impact of the Gaussian filter on the extrapolation process: The histogram displayed by the left part of Figure 6 hints at a high chance of obtaining a significant improvement generated by the Gaussian filter when dealing with the extrapolation procedure applied on noisy curves. This statement comes from the "red" bars higher than the blue ones at the beginning of the chart indicating that there are low values of "filtered" MSES with a high frequency. Moreover, since in the last part of the plot we can see a strong dominance of the blue bars, it highlights the high values of the "original" MSEs relative to the "filtered" ones. The Mann-Whitney U test generates a p-value of 0.02 in this case which is lower than the threshold of 0.05 validating the alternative hypothesis. Therefore, in this case, we achieve a significant improvement produced by applying the Gaussian filter. The bar chart shown in the right side of Figure 6 denotes a substantial probability of obtaining an inconsequential impact provided by applying the Gaussian filter in the scope of improving the performance of the extrapolation process using smooth curves. This hypothesis is verified by the Mann-Whitney U test calculating a p-value of 0.63, greater than 0.05. Consequently, we can validate the null hypothesis and state that the Gaussian filter strategy cannot significantly increase the performance in this particular case.

5 Discussion

Figure 7 briefly summarizes this thesis' results by highlighting the meaningful reduction in terms of mean MSES generated by applying Gaussian smoothing in the case of extrapolation using the noisiest curves. As an analysis, this phenomenon occurs as the noisy curves in their original version are prone to generate high errors for fitting to unseen data. This happens because the fitted curve follows a mathematical formula which provides it with a smooth trajectory while our initial curve might continue to possess a lot of noisy data points. Moreover, the Mean Squared Error, as a way of evaluating curve fitting, also favours the high errors to become even higher, as it squares the differences before summing them.

When we are talking about interpolation using noisy curves, the MSES values still remain high, but they are lower than in the previous case, as we have a mathematical model used for fitting to seen data points and then evaluating the fit using the same points. This showcases lower MSES in this case, which also reduces the impact of the filtering as its main strength of eliminating high errors is diluted.

Switching to smooth curves, it is obvious that the Gaussian filter's main improvement in noise reduction is insignificant. This is because our research analyzes the smoothest

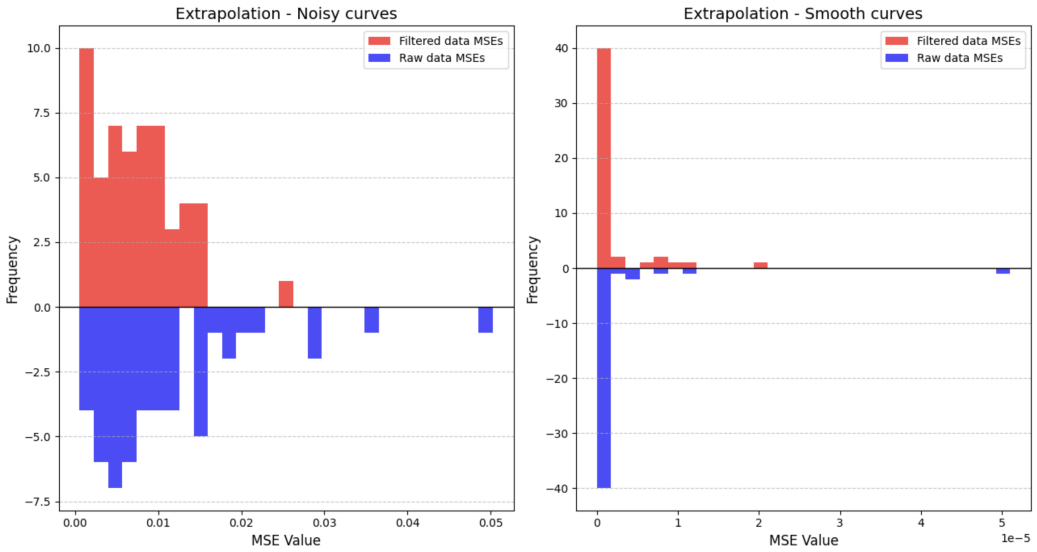

Figure 6: The impact of the Gaussian filter on the extrapolation process

learning curves from LCDB. Therefore, the "original" and the "filtered" curves look very similar, then their corresponding fitted curves will also be similar, and finally, the computed errors will have close values. As a result, the impact of the Gaussian filtering on smooth curves' fitting is negligible, because their shape neutralizes the benefits of the smoothing technique.

One of the main goals of this project was to analyze the impact of applying pre-processing methods to learning curves. As a particular technique which was deeply studied we chose the Gaussian filter. A comprehensive analysis includes discussing both its advantages and disadvantages, as follows:

Advantages: The already-known benefits of the Gaussian filter in the Machine Learning field include attenuating the fluctuations, boosting robustness to noise and improving generalization. In order to compare these known results to the ones obtained during this research, we can easily see that the first two improvements are also maintained, as Figure 3 also highlights the significant noise reduction in the initial curve and the removal of the fluctuations. In addition, the correlation between curves' noise reduction generated by the Gaussian filter and the improved performance in terms of Mean Squared Error was fully described at the start of this section. Moreover, the third advantage of improving generalization was also considered during this Research and it was addressed by showing the increased performance for the extrapolation curve fitting, which is indeed a generalization task performed in the machine learning field(T. Viering, personal communication, January 20, 2025).

Weaknesses: A pitfall of this method might be the undesired possibility of over-smoothing the curve. This can happen when the Gaussian applied has a Sigma value(standard deviation) too high. As a result, the obtained curve might completely lose its initial shape and thus lead to unreliable values of the Mean Squared Error. This might result in some confusion during the experiments, as the performance of the fitting algorithm might seem to be improved, but this is probably caused by a favourable situation. We added Figure 8 in order to highlight the comparison between the curve filtered with a Sigma of 1.0 in blue and the

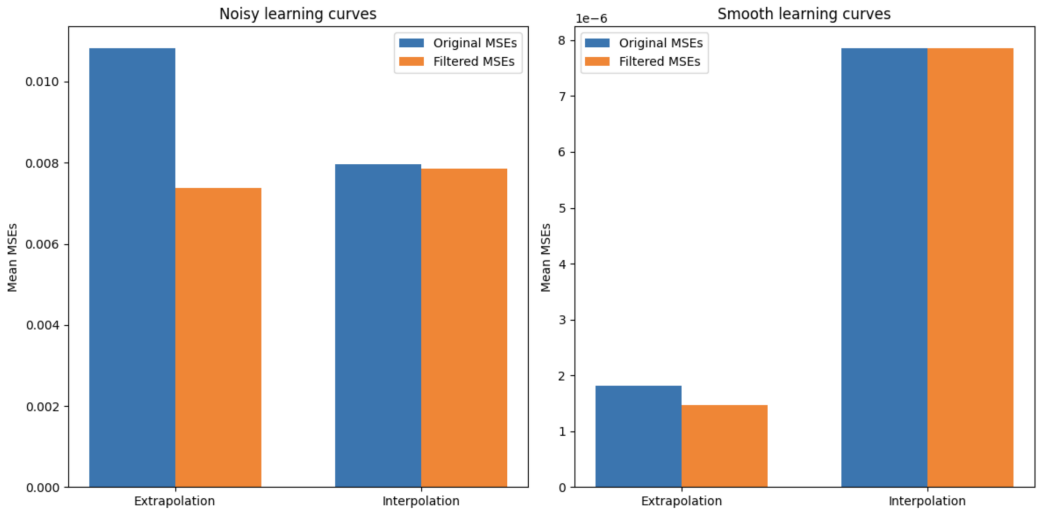

Figure 7: The comparison of "original" and "filtered" MSES

same curve filtered with a Sigma of 20.0 in orange. The second one loses the characteristics of the initial curve and its results during the fitting process might be inaccurate. Thus, this aspect deserves increased attention from those who wish to continue this research.

It is worth mentioning that the conclusions presented in this paper are not universally valid and they depend on the concrete learning curve(s) which need(s) to be studied. As an example, for a collection of learning curves, the impact of applying the Gaussian filter might be significant while for another collection the impact can be insignificant or even negative(reducing the performance). This happens because the learning curves possess really different shapes one from another, so the influence of the Gaussian depends on the type of learning curve and also on the technique chosen for the fitting(interpolation or extrapolation).

6 Responsible Research

This research project was designed to follow all the applicable ethical considerations in the field of Computer Science, ensuring transparency and reproducibility. This assumption is supported by the following statements:

Section 3 of this paper which describes all the required steps to understand and rebuild the code behind the actual research, including the Gaussian filter Sigma values utilized, the model and the percentage of fitting points used for the curve fitting and the metric used for evaluating the fitting.

Section 4 of this paper presents the relevant results of the code-based experiments which might be extremely useful for the readers interested in recreating or even continuing the actual journey of the project. Therefore, the results obtained by the reader can be easily compared with the above-mentioned ones in order to ensure reproducibility, which according to Gundersen[6] does not only require being able to replicate the desired experiments but also collecting the same results.

The Gitlab repository contains the full version of the code and all the experimental

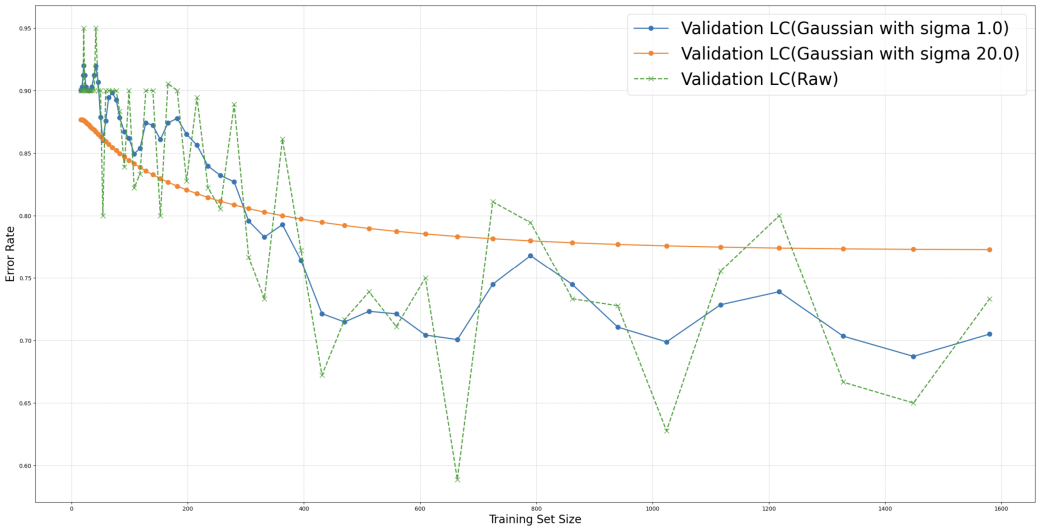

Figure 8: Concrete example of the over-smoothing issue

data(datasets and learners) utilized during this research together with a ReadMe file which provides all the necessary instructions and required dependencies for running the Jupyter-Notebook file.

Since all the necessary data was collected from a public source(LCDB), our study also guarantees privacy, as we do not access any type of personal or sensible data.

Last but not least, this research also preserves the principles of the TU Delft Vision on Integrity 2018-2024[1] and the Netherlands Code of Conduct for Research Integrity 2018[2], since all scientific resources that were utilized are properly referenced and the academic integrity of the research process was also legitimately defended.

7 Future work

This section will provide some useful recommendations for further research, as follows:

The first open issue which may consist of entirely new research is analyzing the impact of other filters(such as the Savitzky-Golay filter[17] or the Low Pass Filter[2]) on the performance of both interpolation and extrapolation processes. These filters are specifically interesting to analyze since the first one is created to smooth the curve while preserving its peaks and edges and the second one is to reduce fluctuations by removing high-frequency noise. In this way, one can obtain a better understanding of how pre-processing methods through kernel filtering can enhance learning curve fitting. Therefore, one can obtain a comprehensive analysis that states in which situations a specific filter provides the best possible results.

The second extension of the actual research might be considering different fitting percentages(0.8 is the one currently used), different model functions utilized for the fitting

[1] `https://www.tudelft.nl/over-tu-delft/strategie/integriteitsbeleid/tu-delft-vision-on-integrity-2018-2024`

[2] `https://www.nwo.nl/en/netherlands-code-conduct-research-integrity`

process(pow3 model is used for the purpose of the current project) and also different ways of calculating the Mean Squared Error of the fitting process(for example extrapolating to the last anchor[14]). To effectively enhance the generalization of this research, we also recommend validating the hypotheses of this paper by including a wider variety of learning curves coming from different fields or studies. This paper only focuses on the learning curves provided by LCDB.

The third way of completing this work is by paying increased attention to the values of Sigma used for the curve fitting. We only used Sigma = 2 for all the interpolation fitting and Sigma = 16 for all the extrapolation fitting, but from Table 1 we can derive that in some isolated situations, the best possible results would have been obtained by using Sigma = 0.125 or Sigma = 1. This might occur in the case of a smooth curve that already produces a low MSE value during the fitting process. Under these circumstances, even a Gaussian with Sigma = 2 could negatively impact the fitting's performance. Therefore, we believe that some further research may be directed at the technique used for selecting the best possible standard deviation values.

8 Conclusion

This thesis presents the impact of the Gaussian filter applied before curve fitting as a preprocessing method. We analyzed its impact on both interpolation and extrapolation techniques in order to determine if its main attribute of noise reduction can potentially improve the fitting's performance. This performance is measured using a mean squared error formula for an individual learning curve. To generalize our findings by determining the significance of applying Gaussian filtering, we performed multiple Mann-Whitney U tests for both interpolation and extrapolation using the 60 smoothest and the 60 noisiest curves from the Learning Curves Database(LCDB). The first result is determined by the values of Gaussian's hyperparameter Sigma which gives the best results for interpolation (Sigma = 2.0) and for extrapolation (Sigma = 16.0). In addition, using the above-mentioned values of Sigma, we measured the significance of applying the Gaussian filter in four main cases and obtained the following outcomes: for interpolation, when using smooth curves, the impact of the Gaussian is insignificant and, by switching to noisy curves, it increases a bit, but only minor progress can be seen; for extrapolation, when analyzing noisy curves, a significant improvement was obtained while for smooth curves we achieved a smaller impact, which can not be considered significant according to Mann Whitney U test performed.

Last but not least, this research process was designed to follow the principles of transparency and reproducibility and leaves room for improvement in the areas of validating and generalizing its findings by applying other kernel filtering methods, ways of evaluating the results, Sigma values or creating a new set of learning curves as the input data.

.

References

[1] Marijan Beg, Juliette Belin, Thomas Kluyver, Alexander Konovalov, Min Ragan-Kelley, Nicolas Thiery, and H. Fangohr. Using jupyter for reproducible scientific workflows. *Computing in Science Engineering*, 23(2):36–46, 2021.

[2] Jonathan M. Blackledget. Chapter 4 - the fourier transform. In Jonathan M. Blackledget, editor, *Digital Signal Processing (Second Edition)*, pages 75–113. Woodhead Publishing, 2006.

[3] Bostjan Brumen, Ales Cernezel, and Leon Bosnjak. Overview of machine learning process modelling. *Entropy*, 23(9):1123, 2021.

[4] George Divine, H. Norton, Anna Baron, and Elizabeth Juarez. The wilcoxon-mann-whitney procedure fails as a test of medians. *The American Statistician*, 72(3):278–286, 2017.

[5] Jeroen M. Goedhart, Thomas Klausch, and Mark A. van de Wiel. Estimation of predictive performance in high-dimensional data settings using learning curves. *Computational Statistics Data Analysis*, 180:107622, 2022.

[6] Odd Erik Gundersen. The fundamental principles of reproducibility. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379, 2021.

[7] Timothy Hodson, Thomas Over, and Sydney Foks. Mean squared error, deconstructed. *Journal of Advances in Modeling Earth Systems*, 13, 2021.

[8] Hassan H. Khalil, Rahmita O. K. Rahmat, and Waleed A. Mahmoud. Chapter 15: Estimation of noise in gray-scale and colored images using median absolute deviation (mad). In *2008 3rd International Conference on Geometric Modeling and Imaging*, pages 92–97, 2008.

[9] Donghwi Kim and Tom Viering. Different approaches to fitting and extrapolating the learning curve. In *Proceedings of the BNAIC/BeneLearn 2022*, 2022.

[10] Kim Tae Kyun. T test as a parametric statistic. *Korean J Anesthesiol*, 68(6):540–546, 2015.

[11] Mingyu Liu, Benny C.F. Cheung, Xiaobing Feng, Lai Ho, and Shu Yang. Gaussian process machine learning-based surface extrapolation method for improvement of the edge effect in surface filtering. *Measurement*, 137:214–224, 2019.

[12] Michael McCartney, Matthias Haeringer, and Wolfgang Polifke. Comparison of machine learning algorithms in the interpolation and extrapolation of flame describing functions. *Journal of Engineering for Gas Turbines and Power*, 142(6):061009, 2020.

[13] Felix Mohr and Jan N. van Rijn. Fast and informative model selection using learning curve cross-validation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9669–9680, 2023.

[14] Felix Mohr, Tom J. Viering, Marco Loog, and Jan N. van Rijn. LCDB 1.0: An Extensive Learning Curves Database for Classification Tasks. In Massih-Reza Amini, Stephane Canu, Asja Fischer, Tias Guns, Petra Kralj Novak, and Grigorios Tsoumakas, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 3–19, Cham, 2023. Springer Nature Switzerland.

[15] Nadim Nachar. The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology*, 4(1):13–20, 2008.

[16] Swathi Pothuganti. Review on over-fitting and under-fitting problems in machine learning and solutions. *International Journal of Advanced Research in Electrical Electronics and Instrumentation Engineering*, 7(9):3692–3695, 2018.

[17] Abraham Savitzky and M. J. E. Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.*, 36(8):1627–1639, 1964.

[18] Andrew Starnes and Clayton Webster. Improved performance of stochastic gradients with gaussian smoothing, 2024.

[19] Tom Julian Viering and Marco Loog. The shape of learning curves: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7799–7819, 2023.

[20] Pauli Virtanen et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272, 2020.

[21] Muhammad Abdul Wahab. Interpolation and extrapolation. *Proceedings of Topics in Systems Engineering, Winter Term*, 17:1–6, 2017.

[22] Andreas Zielesny. From curve fitting to machine learning - an illustrative guide to scientific data analysis and computational intelligence. In *Studies in Computational Intelligence*, volume 18. Springer, Berlin, Heidelberg, 2011.