Enhancing Multi-Objective Machine Learning with an Optimized Lexicographic Approach

Determining the Tolerance Threshold

Guilherme G. D. Fernandes[1] **, Talles H. Medeiros**[2]

[1]Department of Physics, Instituto Superior Técnico, Lisbon, Portugal

[2]Department of Computing and Systems, Universidade Federal de Ouro Preto, Brazil

guilhermegrancho@tecnico.ulisboa.pt, talles@ufop.edu.br

Abstract. *Determining optimal tolerance in lexicographic multi-objective optimization is crucial for robust results. This paper introduces a machine learning algorithm that automatically sets the appropriate tolerance, optimizing the lexicographic strategy and enhancing the analysis of tolerance impact on outcomes. Applied to various datasets, our method consistently provides insights into the relationship between tolerance and model performance. Results show that automatic tolerance optimization improves computational efficiency and accuracy. These findings highlight the importance of addressing tolerance in multi-objective optimization.*

1. Introduction

Multi-objective optimization problems, prevalent in fields such as engineering, finance, and healthcare, involve the simultaneous optimization of multiple conflicting objectives. This inherent complexity poses significant challenges to traditional optimization methods, necessitating the development of advanced methodologies.

Recently, the integration of machine learning with optimization algorithms has demonstrated considerable potential in addressing complex multi-objective problems. The adaptability of machine learning offers novel ways to enhance the efficiency and effectiveness of these optimization processes. This research focuses on improving the lexicographic method, a renowned multi-objective optimization technique, through the application of machine learning [Ahlert and Kliemt 2001].

The lexicographic method prioritizes objectives based on their importance and solves the problem iteratively. A critical aspect of this approach is the selection of the optimal tolerance parameter, which balances solution quality with computational efficiency [Andrade 2020]. Traditionally, tolerance values have been chosen in an ad-hoc manner, significantly limiting the potential of the lexicographic approach [Basgalupp et al. 2014]. This arbitrary selection introduces uncertainties, making it difficult to achieve consistent and reliable results. Despite the critical importance of tolerance values, there is a notable lack of literature addressing systematic methods to explore or resolve this issue. The absence of comprehensive studies on optimizing tolerance parameters undermines the robustness and efficiency of the lexicographic method, often leading to sub-optimal solutions and increased computational costs.

Nevertheless, the lexicographic method stands out in multi-objective optimization for several compelling reasons. Firstly, it provides a clear hierarchical prioritization of objectives, ensuring that the most critical objectives are not sacrificed for lesser ones. For example, in medical diagnosis systems, ensuring high accuracy might be more important than minimizing computational time. This hierarchical prioritization is essential in applications where certain objectives must take precedence over others, such as safety-critical systems or financial decision making [Basgalupp et al. 2009].

Secondly, the lexicographic method is relatively straightforward to understand and implement compared to more complex techniques like evolutionary algorithms [Biswas et al. 2021]. Its simplicity enhances the interpretability of results, which is crucial in fields such as healthcare and finance, where stakeholders must comprehend the trade-offs made. The method's transparency helps in communicating the decision-making process to non-expert users, facilitating trust and adoption.

Thirdly, by focusing on one objective at a time in a pre-determined order, the lexicographic method ensures that each solution is feasible for the most critical objectives before considering others. This approach is particularly beneficial in resource allocation problems where certain constraints must be met before others can be optimized [Boriwan et al. 2020]. For example, in supply chain management, ensuring timely delivery might be prioritized over cost reduction [Boriwan et al. 2021].

The inclusion of a tolerance parameter adds flexibility in solution comparison, allowing for more practical solutions in scenarios where strict adherence to one objective is less important than achieving a balance between several objectives. This flexibility can lead to more robust solutions that perform well across various conditions, enhancing the method's applicability in real-world scenarios [Ahlert and Kliemt 2001].

The lexicographic method also compares favorably to other multi-objective techniques. The weighted sum approach, which combines all objectives into a single composite objective using predetermined weights, is simple but selecting appropriate weights is challenging and may not capture the nuances of trade-offs between objectives as effectively as the lexicographic method [Andrade 2020]. This approach often leads to solutions that might not be truly optimal in a multi-objective sense, as it reduces the problem to a single-objective one.

Pareto optimization seeks to find a set of Pareto-optimal solutions where no other solution is better in all objectives. While comprehensive, it often requires more computational resources and can produce a large set of solutions, making decision-making more complex. Users are left with the task of selecting from potentially many non-dominated solutions, which can be overwhelming and impractical in time-sensitive situations [Basgalupp et al. 2014].

Evolutionary algorithms are powerful for finding diverse solutions but can be computationally intensive and harder to interpret [Basgalupp et al. 2009]. They often require fine-tuning of parameters and extensive computational resources, which may not be feasible in all applications. Furthermore, the stochastic nature of these algorithms can lead to variability in the solutions, making it harder to guarantee consistency [Biswas et al. 2021].

This study proposes a novel approach to improve the lexicographic method by empowering a machine learning algorithm to automatically determine the optimal tolerance parameter tailored to the specific characteristics of each problem. By automating this process, we aim to enhance the robustness and effectiveness of the lexicographic method. Our extensive experiments with benchmark datasets illustrate the efficacy of our approach in identifying Pareto-optimal solutions even with simple neural network structures.

2. Theoretical Framework

The concept of tolerance plays a crucial role in the multi-objective algorithm [Andrade 2020]. It is important to distinguish between absolute tolerance and relative tolerance. Tolerance refers to the absolute value of the tolerance applied during the optimization process. The absolute tolerance is obtained by multiplying the relative tolerance by the value of the objective of interest, which in this case it will be the categorical cross-entropy. Relative tolerance is a relative number, expressed with decimal places.

Using relative tolerance instead of absolute tolerance ensures that the different numerical scales of the objectives, such as categorical cross-entropy and norm L2, do not pose a problem. This is particularly important because objectives can have vastly different numerical meanings and magnitudes. By applying relative tolerance, we normalize the impact of tolerance across objectives, facilitating a more balanced and meaningful optimization process.

This approach addresses a common challenge in multi-objective optimization methods, such as the weighted sum method, where the different scales of the objectives can lead to biased or sub-optimal solutions [Boriwan et al. 2020]. By leveraging relative tolerance, our algorithm effectively navigates these numerical disparities, enabling a more robust and equitable optimization process [Boriwan et al. 2021].

In summary, the lexicographic method offers a balanced approach to multi-objective optimization by providing hierarchical prioritization, simplicity, feasibility assurance for critical objectives, and flexibility through tolerance parameters [Ahlert and Kliemt 2001]. These advantages make it a suitable and effective method for various real-world applications, where interpretability, resource constraints, and clear prioritization are paramount [Basgalupp et al. 2014].

3. Materials and Methods

3.1. Datasets

To explore the robustness of the multi-objective algorithm, we applied it to various data analysis challenges. To this end, we employ four distinct datasets, each representing a unique challenge: MNIST; FashionMNIST; PneumoniaMNIST; and BreastMNIST. By confronting the algorithm with these different datasets, we aim to consolidate its effectiveness and generalization in the face of diverse recognition and classification tasks. This approach allows us to examine the algorithm's capability to handle a variety of problems and validate its utility in various machine learning contexts.

3.2. Artificial Neural Network Structure

The same neural network architecture was used for all datasets to ensure consistency and focus on the multi-objective algorithm. The choice of a simple architecture was deliberate to minimize the impact of the neural network structure on the results, thereby concentrating on the evaluation and optimization of the multi-objective algorithm itself.

The neural network used consists of an input layer of 784 nodes, corresponding to a flat 28x28 pixel MNIST image, followed by two dense layers with 50 and 10 neurons, respectively. The *rectified linear unit* activation function was used for the first dense layer, while the *softmax* activation function was employed in the output layer for multi-class classification. The network weights were randomly initialized using a uniform distribution between -0.1 and 0.1, with biases initialized to zeros. This approach ensured that the neural network structure did not introduce bias into the results and allowed for a more accurate assessment of the multi-objective algorithm performance across different datasets.

This comprehensive description of the neural network architecture used is essential for understanding the context and details of the implemented model for solving multi-objective problems.

3.3. Training the Artificial Neural Network

3.3.1. Objectives

In this study, we adopt a multi-objective approach to train the artificial neural network, considering two key parameters: categorical cross-entropy (CE) and L2 complexity (L2). The choice of these objectives is motivated by their complementary roles in model optimization and their importance in achieving a balance between accuracy, generalization, and complexity of the model [Andrade 2020]. This multi-objective approach allows us to explore the trade-offs between these two objectives and identify solutions that exhibit optimal performance across multiple evaluation criteria.

3.3.2. The Multi-Objective Algorithm

We propose an algorithm that operates through an iterative process that dynamically adjusts the relative tolerance threshold while training a neural network model to determine the optimal relative tolerance to be used in the lexicographic method for multi-objective problems. To achieve this, we investigate the influence that different relative tolerance values have on crucial key metrics, in hopes of finding the most suitable relative tolerance value to be employed. Here's how the algorithm works:

1. Establish a lexicographic order to manage all objectives of the multi-objective problem. In this study, categorical cross-entropy is prioritized as the primary objective, followed by norm L2.
2. Define a sequence of loss functions corresponding to the lexicographic order of objectives. Specifically, we create loss functions for categorical cross-entropy and norm L2.
3. Initialize all necessary variables, including the neural network structure and the relative tolerance value range to be investigated, which varies based on the objectives and datasets.
4. Conduct a complete training of the neural network for each relative tolerance value of interest.
5. For the training process of the neural network, executed for each relative tolerance value:

   (a) Begin training with the loss function dedicated to the primary lexicographic objective (categorical cross-entropy).

   (b) Perform training epochs until the optimization of the lexicographic objective on the validation dataset is achieved.

   (c) Proceed to the next lexicographic objective (norm L2) and repeat the training process.

   (d) From the second lexicographic objective onwards, consider the relative tolerance value to manage the compromise between optimizing the current objective and previously optimized objectives.

(e) Change the loss function when further optimization of the lexicographic objective value is not possible.

(f) Iterate through the every lexicographic objective.

6. Extract important data from each trained neural network, including accuracy, categorical cross-entropy, norm L2, training epochs, and training time.
7. Plot the extracted data as a function of the relative tolerance value to analyze the Pareto curve.
8. Analyze the data to identify the best solution, corresponding to the optimal relative tolerance for the problem.

By iteratively adjusting the relative tolerance threshold and prioritizing objectives using lexicographic ordering, the algorithm effectively navigates the optimization landscape to generate competitive solutions for multi-objective tasks. This systematic and adaptive approach offers a promising framework for optimizing neural network models across various domains and datasets.

3.3.3. Best Solution Analysis

Following the training of the artificial neural network using the multi-objective algorithm with varying relative tolerances, a comprehensive analysis was conducted to determine the best-performing solutions. This analysis aimed to identify configurations that achieved superior performance across multiple evaluation metrics.

To initiate the analysis, we transformed the evaluation metrics, including categorical cross-entropy and L2 complexity, into tensors for efficient manipulation. Subsequently, the data were normalized to ensure consistency and comparability between metrics, allowing a fair assessment of model performance.

The mean of the normalized values was calculated to obtain an aggregate measure of performance. Subsequently, the three configurations with the smallest mean normalized values were identified, representing the top-performing solutions across both metrics.

A combined graph was plotted to illustrate the normalized values of categorical cross-entropy, L2 complexity, and mean performance across different relative tolerances (center of Fig. 1). This visualization provided a comprehensive overview of how each configuration performed relative to others, emphasizing the top-performing solutions. To enhance clarity, distinct markers were utilized to identify the points corresponding to the three configurations with the smallest mean normalized values, ensuring their visibility and significance in the analysis.

Furthermore, the models were sorted according to the ideal tolerance value discovered, the categorical cross-entropy, the L2 complexity, and the training time, providing a detailed analysis of the best-performing configurations. This step provided valuable information on the optimal model parameters and the relative tolerance values associated with the most effective solutions.

This algorithm for analyzing the best solution is significantly different from a typical analysis of a Pareto curve or solving a multi-objective problem by summing values into a single function. The crucial distinction is that the solutions being analyzed were generated by an optimized lexicographic algorithm. As will be shown in the practical results, this approach greatly impacts the characteristics of the solutions. Essentially, what is being analyzed is the impact of the tolerance value in the optimized lexicographic method previously utilized. This provides a deeper understanding of the robustness and efficiency of the multi-objective algorithm in optimizing the performance of the artificial neural network across various evaluation metrics.

4. Results and Discussion

4.1. MNIST Dataset

For the MNIST dataset, 33 solutions were generated within the relative tolerance range of 20.25 to 24. This interval was meticulously selected after numerous iterations of the algorithm, which allowed the identification of a region where meaningful variations in optimization metrics occurred. Below the tolerance threshold of 20.25, the algorithm seemed to converge to a stable regime, where the accuracy plateaued at a commendable 0.96 (left of Fig. 1), indicating robust performance in digit recognition tasks. Similarly, categorical cross-entropy, a measure of the model's prediction accuracy, remained consistently low at around 0.1 (center of Fig. 1), signifying the reliability of the model's predictions. The L2 norm, representing the complexity of the model, exhibited a relatively stable behavior, hovering around 3400 (center of Fig. 1). These observations suggest that within this range, the model's performance remained largely unaffected by variations in the tolerance value.

In contrast, beyond the upper relative tolerance limit of 24, significant deviations in the optimization metrics were observed. The accuracy sharply declined to a mere 0.09 (left of Fig. 1), indicating a substantial drop in the predictive power of the model. Concurrently, the categorical cross-entropy increased to 2.30 (center of Fig. 1), reflecting a deterioration in the model's ability to make accurate predictions. Interestingly, the L2 norm decreased drastically to 4.0×10^{-4} (center of Fig. 1), suggesting a reduction in model complexity. These findings underscore the critical role of tolerance value in influencing the optimization process, as deviations from the optimal range can lead to considerable degradation in model performance.

Exploring the interval of interest, 20.25 to 24, revealed intriguing insights into the behavior of optimization metrics. Variations in the tolerance value within this range caused irregular fluctuations in the parameters: categorical cross-entropy; L2 norm; and accuracy. Although these fluctuations may initially appear disruptive, they present an opportunity to finely balance the trade-off between categorical cross-entropy and L2 complexity. It is within these irregular disturbances that the optimal tolerance value emerges, enabling the algorithm to prioritize the optimization of categorical cross-entropy while maintaining an acceptable norm L2.

Taking into account these observations, the algorithm has shown exceptional results, pinpointing 23.5313 as the optimal relative tolerance value for this lexicographic approach. Consequently, the top 3 solutions identified (center of Fig. 1) can be identified in Table 1.

Rank	Relative Tolerance	Accuracy	CE	L2	Training Time (s)
1º	23.5313	0.9732	0.0949	33694.9570	94.8731
2º	21.8906	0.9664	0.1145	28961.2168	46.9999
3º	20.8359	0.9668	0.1118	29041.4043	41.4505

Table 1. Best Models for MNIST.

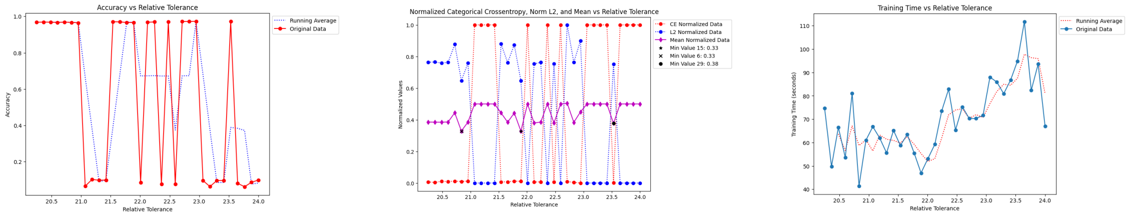

Figure 1. Algorithmic Analysis of MNIST.

4.2. FashionMNIST Dataset

For the FashionMNIST dataset, a distinct set of 30 solutions was generated within a specific relative tolerance range of 4.75 to 5.20. As observed in the MNIST dataset, below this tolerance threshold, the algorithm stabilized, yielding an admirable accuracy rate of 0.87 (left of Fig. 2). Consequently, categorical cross-entropy consistently hovered around 0.35 (center of Fig. 2), underscoring the reliability of the model's predictions. Furthermore, the L2 norm exhibited relative stability, maintaining a value near 4000 (center of Fig. 2).

Similarly, beyond the upper tolerance limit of 5.20 for the FashionMNIST dataset, notable deviations were observed in the optimization metrics. The accuracy was sharply reduced to 0.1 (left of Fig. 2), signaling a significant reduction in the predictive capacity of the model. Simultaneously, the categorical cross-entropy spiked to 2.30 (center of Fig. 2), indicating a deterioration in the model's predictive accuracy. Interestingly, the L2 norm decreased substantially to 5.0×10^{-4} (center of Fig. 2), implying a reduction in the complexity of the model. These trends are consistent with those observed in the MNIST dataset, emphasizing the sensitivity of the model's performance to tolerance values.

The analysis of the FashionMNIST dataset, within the 4.75 to 5.20 interval, revealed a pattern of irregular fluctuations in optimization metrics similar to those observed in the MNIST dataset. These variations in tolerance value also caused fluctuations in categorical cross-entropy, L2 norm, and accuracy. This similarity suggests that, as with the MNIST dataset, these irregular fluctuations help identify the optimal tolerance value, balancing the trade-off between categorical cross-entropy and L2 complexity.

With these observations in mind, the algorithm has achieved highly commendable outcomes, identifying 4.9983 as the ideal relative tolerance value to employ in this lexicographic method. As a result, the top 3 solutions (center of Fig. 2) can be seen in Table 2.

Rank	Relative Tolerance	Accuracy	CE	L2	Training Time (s)
1º	4.9983	0.8772	0.3474	35053.9336	96.1841
2º	5.2000	0.8766	0.3539	35210.9961	79.7574
3º	4.8586	0.8746	0.3561	35234.4648	78.7583

Table 2. Best Models for FashionMNIST.

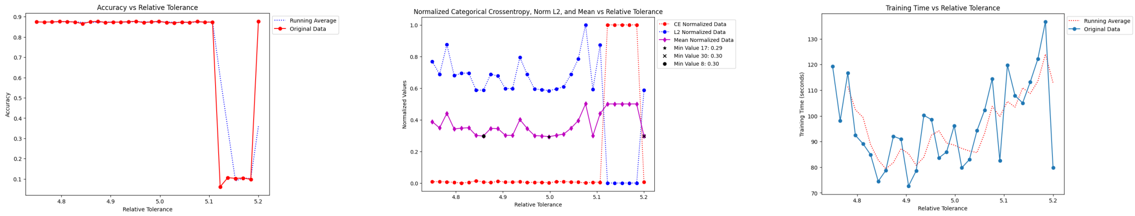

Figure 2. Algorithmic Analysis of FashionMNIST.

4.3. PneumoniaMNIST Dataset

For the PneumoniaMNIST dataset, a set of 33 solutions was generated within a specific relative tolerance range spanning from 14 to 18. Similarly to the MNIST and FashionMNIST datasets, below this tolerance threshold, the algorithm exhibited stability, achieving a commendable accuracy rate of 0.87 (left of Fig. 3). Similarly, categorical cross-entropy consistently maintained a value around 0.4 (center of Fig. 3), indicating reliable model predictions. Furthermore, the L2 norm remained relatively stable, hovering around 1400 (center of Fig. 3).

For the PneumoniaMNIST dataset, beyond the upper tolerance limit of 18, noticeable deviations were also observed in the optimization metrics. The accuracy sharply declined to 0.01 (left of Fig. 3), indicating a significant deterioration in the predictive capacity of the model. Simultaneously, the categorical cross-entropy increased to 2.30 (center of Fig. 3), reflecting a degradation in the predictive accuracy of the model. Interestingly, the L2 norm decreased substantially to 2.0×10^{-8} (center of Fig. 3), implying a reduction in the complexity of the model. These observations are in line with the findings from the MNIST and FashionMNIST datasets, highlighting the critical impact of tolerance value deviations on model performance.

In the PneumoniaMNIST dataset, exploring the interval from 14 to 18 resulted in findings consistent with the MNIST and FashionMNIST dataset. Here, variations in tolerance value induced irregular fluctuations in categorical cross-entropy, L2 norm, and accuracy. This consistency across datasets reinforces the notion that these irregular fluctuations are key to identifying the optimal tolerance value, allowing the algorithm to effectively prioritize categorical cross-entropy optimization while maintaining acceptable model complexity.

Given these observations, the algorithm has shown remarkable effectiveness, determining that 14.1250 is the optimal relative tolerance value for using this lexicographic strategy. The algorithm's top 3 solutions (center of Fig. 3), therefore, can be observed in Table 3.

Rank	Relative Tolerance	Accuracy	CE	L2	Training Time (s)
1º	14.1250	0.8622	0.4602	1421.7051	11.2353
2º	14.7500	0.8599	0.4679	1420.6168	9.9428
3º	14.6250	0.8478	0.4779	1411.7041	7.8627

Table 3. Best Models for PneumoniaMNIST.

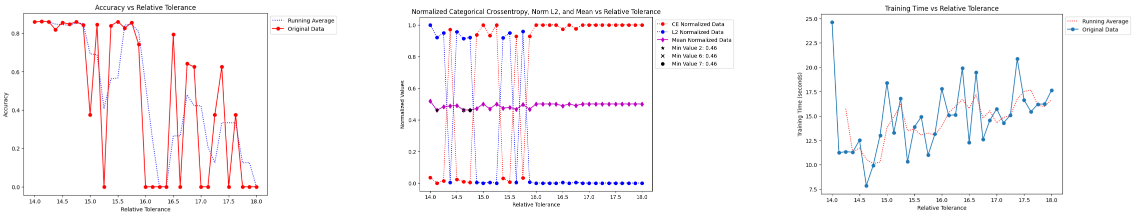

Figure 3. Algorithmic Analysis of PneumoniaMNIST.

4.4. BreastMNIST Dataset

For the BreastMNIST dataset, a total of 33 solutions were generated within a specific relative tolerance range spanning from 0 to 5. This range was carefully chosen after multiple iterations of the algorithm, identifying it as the region of interest where meaningful variations in optimization metrics occurred. Below the tolerance threshold of 0, no results were obtained as tolerance values cannot be negative. In contrast, beyond the upper tolerance limit of 5, significant deviations in the optimization metrics were observed. The accuracy sharply declined to 0.006 (left of Fig. 4), indicating a considerable degradation in the predictive capacity of the model. Simultaneously, the categorical cross-entropy increased to 2.30 (center of Fig. 4), reflecting a deterioration in the model's predictive accuracy. Interestingly, the L2 norm decreased substantially to 0.005 (center of Fig. 4), implying a reduction in model complexity.

Exploring the interval of interest, from 0 to 5, revealed intriguing insights into the behavior of optimization metrics. Unlike other datasets, the variations in parameters such as accuracy, categorical cross-entropy, and the L2 norm within this tolerance range were significantly more regular. Specifically, categorical cross-entropy exhibited an exponential relationship with relative tolerance, while the L2 norm showed an inverse proportionality to relative tolerance. This consistent behavior suggests a unique optimization landscape for BreastMNIST, where the interplay between tolerance values and optimization metrics follows distinct patterns. Leveraging these observed patterns within the tolerance range presents an opportunity to uncover an optimized solution for the multi-objective problem, where the fine balance between categorical cross-entropy and L2 complexity can be achieved more systematically.

In light of these observations, the algorithm has delivered outstanding results, establishing 0.7813 as the best relative tolerance value to apply this lexicographic approach. Consequently, the top 3 solutions identified by the algorithm (center of Fig. 4) can be seen in Table 4.

Rank	Relative Tolerance	Accuracy	CE	L2	Training Time (s)
1º	0.7813	0.7436	0.6568	101.3206	11.6942
2º	1.0938	0.7372	0.6730	99.1314	11.9251
3º	0.9375	0.7564	0.6794	99.8176	12.7922

Table 4. Best Models for BreastMNIST.

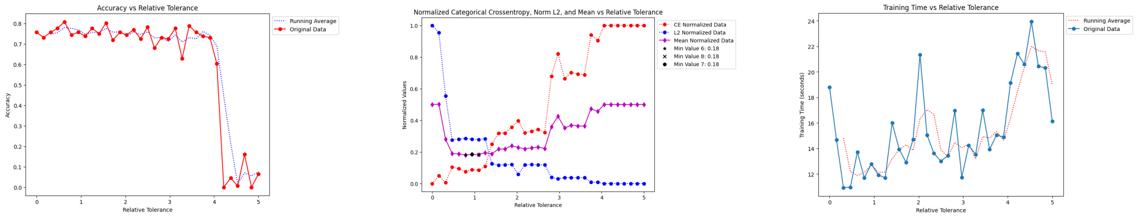

Figure 4. Algorithmic Analysis of BreastMNIST.

4.5. Results Overview

As demonstrated, intriguing phenomena consistently emerged in all data sets, providing valuable insights into the intricacies of the tolerance value. Firstly, a notable pattern was observed in the behavior of critical metrics, including accuracy, norm L2, and categorical cross-entropy, outside a specific range of tolerance values. It became apparent that beyond this range, these metrics exhibited stability, suggesting a transition towards a single-objective optimization scenario. This behavior stems from extreme tolerance values, where the optimization focus shifts exclusively towards either categorical cross-entropy or norm L2, disrupting the delicate balance between them. Consequently, beyond the identified range of interest, accuracy and categorical cross-entropy are optimized while norm L2 is unduly penalized for low tolerance values. In contrast, for high tolerance values, an optimized L2 value is achieved at the expense of accuracy and cross-entropy, demonstrating the inverse relationship. Thus, our analytical focus centers on the delineated tolerance value ranges, ensuring a holistic examination of the multi-objective optimization landscape.

Secondly, a noticeable trend was observed in relation to training time in all datasets (right of Fig. 1; Fig. 2; Fig. 3 and Fig. 4). On average, an irregular escalation in training time was observed with higher tolerance values. This phenomenon can be attributed to the increase in the number of training epochs required to optimize the norm L2 as tolerance values increase. With increased tolerance, the neural network training algorithm gains flexibility to explore additional iterations optimizing norm L2, thereby elongating the training duration.

These consistent phenomena underscore the intricate interplay between tolerance values and algorithmic performance, highlighting the importance of selecting an appropriate tolerance range for effective optimization. While computational resources constraints posed challenges to the depth of analysis, the elucidated observations offer profound insights into the importance of the tolerance value across diverse datasets.

4.5.1. Exploring Limitations

Despite the constraints imposed by the limited RAM memory capacity available, our investigation made significant strides. Although exploration of the Pareto curve was curtailed, we managed to generate up to 33 solutions, showcasing the remarkable potential of our approach. By iteratively refining the tolerance value range, we maximized the efficiency of our analysis. However, the constrained number of solutions did influence the depth of our investigation, but also underscores the effectiveness of our methodology in optimizing resource utilization.

It is crucial to note that we employed a simple neural network structure to maintain focus on the multi-objective algorithm, but the approach can be applied to any machine learning model. Achieving satisfactory results with such simplicity is noteworthy, as it underscores the algorithm's capability to deliver exceptional outcomes. This reinforces the robustness and effectiveness of our approach.

Additionally, our optimization function targeted two objectives: categorical cross-entropy and the L2 norm. In situations where more objectives of different natures need to be optimized simultaneously, the effectiveness of the proposed lexicographic algorithm requires validation. Although, it was designed to perform with any optimization function and set of objectives. While initial findings are promising, additional research is required to fully assess its potential under such conditions. Future studies could explore the robustness of our method in more diverse contexts, such as optimising fairness, interpretability, and energy efficiency metrics in neural networks.

Addressing these aspects could not only mitigate current limitations but also enhance the applicability and effectiveness of our developed multi-objective algorithm. This would further solidify its contribution to advancing multi-objective optimization in machine learning.

5. Conclusion

A meticulous analysis of the results of the implementation and evaluation of our proposed algorithm reveals extremely promising and insightful results. Our exploration into determining the optimal tolerance threshold within the lexicographic framework has unveiled innovative pathways for addressing multi-objective optimization in machine learning.

A key revelation from this study is the efficacy of our approach in identifying the correct tolerance level, which has traditionally been determined in an ad-hoc manner. With no academic literature relevant to the subject, we are exploring uncharted ground. By systematically varying the tolerance values and rigorously analyzing their impact on key performance metrics, such as categorical cross-entropy, norm L2, and training time, we were able to pinpoint the optimal tolerance levels with precision, for each specific group of objectives and dataset. This methodological approach not only enhances the robustness of the multi-objective algorithm but also ensures a more reliable and consistent determination of tolerance levels, leading to improved model performance and better generalization. Such a systematic approach eliminates the guesswork and subjectivity often associated with ad-hoc methods, providing a clear and reproducible pathway to fine-tuning neural network training processes.

The findings of this study reveal profound insights into the behavior of our algorithm in different datasets. Firstly, despite maintaining the same neural network structure and lexicographic objectives, the optimal tolerance threshold varied significantly with each dataset. This underscores the sensitivity of the optimization process to dataset-specific nuances, highlighting the need for a customized approach in determining the tolerance threshold. We show that the previous ad-hoc approach can significantly and detrimentally impact the task at hand.

Furthermore, our findings highlight the existence of a specific tolerance threshold range for each dataset, beyond which crucial parameters such as accuracy, categorical cross-entropy, and L2 norm reach a plateau. This phenomenon arises from extreme tolerance values, transforming the multi-objective problem into a single-objective one. By utilizing the tolerance value determined by our algorithm for each scenario, we demonstrated excellent results, showcasing the potential of our approach in effectively addressing multi-objective challenges in machine learning.

Within the identified tolerance threshold range, we discovered two distinct behavioral patterns. The first pattern, observed in the MNIST, FashionMNIST, and PneumoniaMNIST datasets, exhibited irregular and abrupt fluctuations in key parameters, which presented challenges in the analysis and selection of the optimal tolerance value. Despite this complexity, our proposed algorithm proved effective in navigating and resolving these challenges. In contrast, the second pattern, observed in BreastMNIST, showed a more regular behavior, with parameters such as accuracy, categorical cross-entropy, and the L2 norm showing consistent trends. This facilitated the identification of the optimal tolerance value, as evidenced by our findings.

Moreover, our study elucidates that, on average, the training time of the generated neural networks increases with the tolerance value. This increase can be attributed to a greater number of training epochs required to optimize the L2 norm. With higher tolerance values allowing for a more significant trade-off between categorical cross-entropy and the L2 norm, the algorithm explores a broader optimization space, resulting in longer training times.

In conclusion, our research highlights the significant impact of our algorithm in addressing multi-objective challenges within machine learning. By providing fresh insights and innovative methodologies, this study establishes a solid foundation for future explorations in the field.

Ultimately, our work has resulted in the creation of a potent tool for handling multi-objective optimization. As proven, this algorithm has the remarkable ability to empower even the simplest neural network structures, significantly amplifying their effectiveness and enabling them to achieve commendable results in complex tasks. Further testing against an expanded set of objectives will ensure its broad applicability and deployment in real-world settings, maximizing the potential of machine learning to confront intricate challenges and foster a new era of innovation and discovery.

References

[Ahlert and Kliemt 2001] Ahlert, M. and Kliemt, H. (2001). A lexicographic decision rule with tolerances. *Analyse Kritik*, 23.

[Andrade 2020] Andrade, F. (2020). Metodologia multicritério de apoio à decisão - a gestão da informação no processo decisório.

[Basgalupp et al. 2014] Basgalupp, M., Barros, R., de Carvalho, A., and Freitas, A. (2014). Evolving decision trees with beam search-based initialization and lexicographic multi-objective evaluation. *Information Sciences*, 258.

[Basgalupp et al. 2009] Basgalupp, M., Barros, R., de Carvalho, A., Freitas, A., and Ruiz, D. (2009). Legal-tree: A lexicographic multi-objective genetic algorithm for decision tree induction. pages 1085–1090.

[Biswas et al. 2021] Biswas, A., Fuentes, C., and Hoyle, C. (2021). A mo-bayesian optimization approach using the weighted tchebycheff method. *Journal of Mechanical Design*, 144.

[Boriwan et al. 2020] Boriwan, P., Ehrgott, M., Kuroiwa, D., and Petrot, N. (2020). The lexicographic tolerable robustness concept for uncertain multi-objective optimization problems: A study on water resources management. *Sustainability*, 12.

[Boriwan et al. 2021] Boriwan, P., Kuroiwa, D., and Petrot, N. (2021). On the properties of lexicographic tolerable robust solution sets for uncertain multi-objective optimization problems. *Carpathian Journal of Mathematics*, 37:25–34.