Multi-Objective Optimization in Machine Learning
Seminar

Enhancing COSMOS with Augmented Chebyshev Scalarization: Incorporating SoftMax for Advanced Pareto Front Discovery in Multi-Objective Deep Learning

Department of Statistics
Ludwig-Maximilians-Universität München

Davit Martirosyan

Munich, April, 2024

Contents

1 Abstract 1

2 Introduction 2

3 Related Work 3

4 Proposed Method 5

4.1 Augmented Chebyshev Scalarization 5

4.2 Augmented Chebyshev Scalarization with Softmax 6

4.3 Methodological Enhancements and Expected Outcomes 6

5 Data & Experimental Setup 6

5.1 Datasets 7

5.2 Reproducibility 7

5.3 Comparative Framework 8

5.4 Implementation Details 8

6 Results & Analysis 9

6.1 Fairness on tabular data and Multi-MNIST 9

6.2 Pareto fronts comparison on tabular data and Multi-MNIST 11

6.3 Convergence of all COSMOS variants 12

7 Discussion & Conclusions 14

A Appendix 16

A.1 Multi-MNIST+Fashion 16

A.1.1 COSMOS ACS SoftMax Pareto front throughout different epochs . . . 16

A.1.2 COSMOS ACS Pareto front throughout different epochs 17

A.1.3 Original COSMOS Pareto front throughout different epochs 18

1 Abstract

In the dynamic and complex field of deep learning (DL), the effective management of Multi-Objective Optimization (MOO) is crucial for the development of sophisticated models that are capable of addressing a diverse array of often conflicting performance metrics. Such challenges are prevalent across various domains, including autonomous driving, where the trade-off between decision accuracy and reaction time is paramount; medical diagnostics, where the balance between precision and recall can determine life-altering outcomes; and in natural language processing, where the efficiency and accuracy of models must be judiciously balanced against the constraints of computational resources. Traditional approaches to MOO, notably linear scalarization (LS), have laid the groundwork for this endeavor but exhibit pronounced limitations in their capacity to accurately map the complex and multifaceted landscapes of Pareto fronts—particularly, their inefficacy in representing concave regions effectively. This inadequacy presents a significant impediment to the deployment of DL models in a variety of real-world applications, where achieving a harmonious balance among multiple objectives is pivotal. Addressing this critical gap, this paper introduces the augmented Chebyshev scalarization (ACS) method within the COSMOS framework, further enriched by a distinct and novel variant that incorporates a SoftMax function to smooth the scalarization process. This dual-faceted advancement not only showcases a more nuanced and robust approximation of Pareto fronts, a concept central to MOO that delineates the frontier of optimal trade-offs, beyond which no improvements can be made on one objective without conceding performance on another (Deb et al., 2000); (Ehrgott, 2005), but also underscores the potential of these enhanced scalarization methods to significantly elevate the application and effectiveness of MOO in DL in case of difficult optimization landscapes. This study demonstrates the capability of the ACS alternative of COSMOS—both with and without the SoftMax modification—in navigating the complexities of MOO landscapes.

2 Introduction

Within the domain of machine learning (ML), MOO emerges as a foundational pillar, addressing the multifarious and critical challenges inherent in various applications, from multi-task learning to the nuances of fairness considerations. These domains necessitate the delicate balancing of multiple, often conflicting objectives, such as the minimization of classification errors while concurrently ensuring adherence to fairness criteria. Historically approached through evolutionary algorithms, the field of ML has witnessed a progressive shift towards the adoption of gradient-based MOO algorithms. These methods, exemplified by the work of Fliege and Svaiter (2000) and Desideri (2012), have heralded improvements in convergence rates, albeit unveiling complexities in the comprehensive delineation of Pareto fronts that comprise non-dominating solutions—a process that enables practitioners to select optimally balanced solutions post hoc based on the trade-offs they embody.

The intersection of DL with MOO has predominantly been navigated through the lens of multi-task learning, with initial methodologies focusing on identifying singular Pareto optimal solutions via gradient descent. Evolutionary strategies, in their quest to populate the Pareto front with a multiplicity of solutions, have adapted neural networks to be contingent upon preference vectors, delineating the relative importance of each objective (Lin et al., 2019); (Mahapatra and Rajan, 2020). Such methodologies, despite their innovation, grapple with scalability challenges, particularly as the depth and complexity of neural networks expand, leading to an exponential increase in the number of trainable parameters.

In an endeavor to transcend these limitations, this study introduces an enhancement to the novel COSMOS methodology, encapsulating two significant advancements: the integration of ACS and a separate, innovative implementation that incorporates a SoftMax function into this scalarization framework. The ACS stands as a robust mechanism, enabling a more accurate and comprehensive exploration of the Pareto front in case of complex optimization landscapes and difficult tasks, thereby addressing the limitations inherent in previous scalarization approaches. In a parallel and distinct advancement, the inclusion of a SoftMax function seeks to refine this exploration further. By smoothing the scalarization process, the SoftMax variant introduces a more granular and continuous mechanism for navigating the intricacies of MOO landscapes.

This bifurcated approach not only facilitates a deeper and more nuanced understanding of the competing objectives inherent in MOO but also signifies a substantial advancement in the capability of DL models to judiciously balance these objectives. This comprehensive experimental framework illustrates the efficacy of both methodologies, showcasing not only their capacity to rival and surpass existing state-of-the-art MOO techniques in terms of quality but also their remarkable enhancement of computational efficiency and scalability. This study, thus, stands not merely as a rectification of the shortcomings presented by LS but as a leap forward in propelling the efficiency and applicability of MOO within the DL landscape, paving the way for more sophisticated and effective optimization strategies in tackling complex, MOO datasets.

3 Related Work

The domain of DL presents a fertile ground for MOO, driven by an array of applications that necessitate navigating the trade-offs between conflicting objectives. From enhancing the performance of multi-task learning frameworks to embedding fairness considerations into algorithmic decisions, MOO stands as a cornerstone in the advancement of ML methodologies that are both robust and equitable (Zhang and Yang, 2018); (Sener and Koltun, 2018).

The complexity of real-world problems often demands solutions that cater to multiple criteria simultaneously, marking a departure from traditional single-objective optimization approaches. This shift underscores the need for methods capable of identifying a set of Pareto optimal solutions—each representing a unique compromise among competing objectives. Such solutions are encapsulated within the Pareto front.

Despite its critical role, accurately approximating the Pareto front remains a formidable challenge, particularly in the context of DL. Conventional strategies like LS, while computationally straightforward, often yield Pareto fronts that are inadequately diverse or fail to capture the full spectrum of optimal trade-offs. These limitations are partly attributed to the simplistic assumption of linear relationships among objectives, which does not hold in many complex scenarios (Ehrgott, 2005).

Recognizing these challenges, recent advancements have introduced novel approaches aimed at more effectively approximating Pareto fronts in DL settings. Deist et al. (2021) propose a

method rooted in the concept of hypervolume maximization, leveraging this metric to dynamically adjust the weighting of objectives during the training of neural networks. The hypervolume, defined as the space encompassed by a set of solutions within the objective space, serves as a comprehensive measure that accounts for both the closeness of solutions to the Pareto front and their distribution. By optimizing for hypervolume maximization, this approach not only seeks Pareto optimal solutions but also ensures their diversity, addressing critical shortcomings in previous MOO methods (Deist et al., 2021).

In parallel, Ruchte and Grabocka (2021) explore a scalable solution to MOO that circumvents the need for training multiple networks or the excessive parameterization introduced by hyper-networks. Their methodology conditions the neural network directly on preference vectors, effectively tailoring the network's predictions to the desired trade-offs. This preference-conditioned approach allows for the approximation of the entire Pareto front in a single training iteration, significantly reducing computational overhead. Furthermore, by incorporating a novel penalty term that minimizes the angle between solutions and their corresponding preference vectors, the method ensures a well-spread Pareto front, mitigating one of the key drawbacks of LS (Ruchte and Grabocka, 2021).

The contributions of these studies to the field of MOO in DL are manifold. They not only offer more efficient means of approximating Pareto fronts but also extend the applicability of MOO to a broader range of problems. By addressing the limitations of traditional MOO strategies, such as the lack of solution diversity and the computational inefficiency of generating comprehensive Pareto fronts, these methodologies pave the way for advancements in multi-task learning, fairness in ML, and beyond.

4 Proposed Method

Extending the COSMOS framework, this methodology integrates ACS and ACS with SoftMax to address the limitations of LS, particularly for exploring complex Pareto fronts such as concave or disjoint ones. This extension is underpinned by the ACS principle, enhanced with a cosine similarity measure to ensure the diversity and distribution of solutions, drawing theoretical support from Emmerich and Deutz (2018).

4.1 Augmented Chebyshev Scalarization

The ACS aims to minimize the maximum weighted deviation from an ideal point, augmented by a summation of all objectives, thereby promoting a more equitable exploration of the Pareto front. Incorporating cosine similarity into this framework enhances solution diversity by aligning solutions more closely with their preference vectors. The combined optimization objective is formulated as follows:

$$\min_{\mathbf{x}} \left\{ \max_{i=1}^{m} [r_i \cdot (f_i(\mathbf{x}) - z_i^*)] + \rho \sum_{i=1}^{m} (f_i(x) - z_i^*) \right\} - \lambda \cdot \frac{\mathbf{r} \cdot \mathbf{f}(\mathbf{x})}{||\mathbf{r}||||\mathbf{f}(\mathbf{x})||} \tag{1}$$

where:

- $\mathbf{x}$ represents the decision variables.
- $\mathbf{r} \sim \text{Dir}(\boldsymbol{\alpha})$, with $\mathbf{r}$ being the preference vector sampled from a Dirichlet distribution characterized by the parameter vector $\boldsymbol{\alpha}$, guiding the search towards diverse regions of the Pareto front.
- $\mathbf{f}(\mathbf{x})$ denotes the vector of objective function values at $\mathbf{x}$.
- $\mathbf{z}^*$ is the ideal point, typically representing the best achievable values for each objective.
- ρ is a small positive scalar, the augmentation coefficient.
- λ is a regularization parameter for the cosine similarity measure, enhancing the diversity among solutions.

4.2 Augmented Chebyshev Scalarization with Softmax

To further refine the exploration of the Pareto front, an additional variant of ACS is proposed that incorporates a SoftMax function. This variant aims to provide a smoother approximation of the max operator by utilizing the SoftMax function, thereby potentially alleviating issues related to the abruptness of maximum selection in the original formulation. The adapted objective function with SoftMax is described as follows:

$$\min_{\mathbf{x}} \left\{ \text{SoftMax}\left([r_i \cdot (f_i(\mathbf{x}) - z_i^*)]_{i=1}^m \right) + \rho \sum_{i=1}^{m} (f_i(\mathbf{x}) - z_i^*) \right\} - \lambda \cdot \frac{\mathbf{r} \cdot \mathbf{f}(\mathbf{x})}{||\mathbf{r}||||\mathbf{f}(\mathbf{x})||} \tag{2}$$

4.3 Methodological Enhancements and Expected Outcomes

By integrating both the traditional ACS and its novel variant with SoftMax, the enhanced COSMOS framework aims to provide a robust method for identifying diverse and well-distributed sets of Pareto optimal solutions across varied optimization landscapes. These enhancements are particularly suited to DL applications in MOO, where the ability to uncover and accurately approximate the Pareto front can significantly impact the effectiveness and applicability of the solutions derived.

Expected outcomes include improved convergence properties, reduced sensitivity to outlier objectives, and a more comprehensive mapping of the Pareto front. These advancements underscore the potential of the enhanced COSMOS framework to more effectively manage the complex interplay between multiple objectives, enhancing its utility in a broad range of optimization tasks.

5 Data & Experimental Setup

Following the methodology established in the "Scalable Pareto Front Approximation for Deep Multi-Objective Learning" study (Ruchte and Grabocka, 2021), our extended COSMOS method with Chebyshev scalarization and its variant with SoftMax employs the same datasets and experimental settings to maintain consistency and comparability of results.

5.1 Datasets

Three primary datasets are utilized, each selected to challenge the model's ability to balance competing objectives effectively:

- **Multi-MNIST**: Comprising images with overlaid digits, this dataset challenges the model to simultaneously classify multiple digits, emphasizing the core principles of MOO in visual recognition tasks. For a detailed description of the Multi-MNIST dataset and its variants see Sabour et al. (2017) and Lin et al. (2019). LeNet (LeCun et al., 1999) is used with task-specific heads similar to prior work (Sener and Koltun, 2018) and the learning rate is decayed at epochs 20, 40, 80, 90 by 0.1. As losses cross-entropy has been defined for the BR and TL tasks.

- **Fairness Datasets (Adult, COMPASS)**: The data of the Adult and (Dua and Graff, 2017), Compass (Angwin et al., 2016), datasets are preprocessed. These are employed to evaluate the model's capability to minimize classification errors while ensuring fairness, these datasets highlight the societal implications and practical challenges of applying MOO in sensitive areas. As a differentiable fairness objective the hyperbolic tangent relaxation of Difference of Equality of Opportunity (DEO) is used, (Padh et al., 2020) defined as:

$$\text{DEO} = \frac{1}{N}\sum_{a=0,y=1} t(f\cdot);c) - \frac{1}{N}\sum_{a=1,y=1} t(f\cdot);c)$$

where $t(x;c) = \tanh(c \cdot \max(0,x))$ and set $c = 1$ for all experiments.

5.2 Reproducibility

For reproducibility, the hyperparameters of the baseline methods have been adjusted to align with the values provided in their original publications and official sources [1]. To enhance the competitiveness of the baselines, early stopping has been implemented based on the hypervolume metric (Zitzler et al., 2007), calculated using the validation set. Unless specified differently, the batch size has been set at 256 and the Adam optimizer (Kingma and Ba, 2015) has been employed with a learning rate of 0.001. Presented are average scores and Pareto fronts

[1] Details in references

from five separate trials for each method and dataset. For calculating hyper-volume, (2, 2) has been used as the reference point and 25 evenly spaced test rays.

5.3 Comparative Framework

To validate the effectiveness of Chebyshev scalarization within the COSMOS framework, a rigorous comparative analysis is conducted against the original COSMOS model and established baselines. This comparative approach is detailed, focusing on critical metrics such as Pareto front quality (spread and diversity) and computational efficiency (training time overhead compared to single-objective optimization).

The performance metrics from the extended COSMOS model will be directly compared to those of the original model and relevant baselines. This analysis aims to underscore the benefits of Chebyshev scalarization, and its SoftMax variant, especially in generating a well-distributed Pareto front and enhancing scalability with minimal computational overhead.

5.4 Implementation Details

Consistent with the original COSMOS study, this research has adhered to identical preprocessing, model architectures, and training protocols for each dataset. This approach ensures that observed performance differences are attributable solely to these methodological enhancements, facilitating the replication of these experiments and validation of the findings.

By maintaining alignment with the established COSMOS methodology and incorporating a robust comparative analysis, this study offers a comprehensive evaluation of the proposed enhancements.

6 Results & Analysis

6.1 Fairness on tabular data and Multi-MNIST

Table 1: Results compared to the state-of-the-art methods (HV: hyper-volume)

Method	**Adult**		**Compass**	
	HV	**Time (Sec)**	**HV**	**Time (Sec)**
Single Task	-	139	-	114
PHN-EPO	± 3.34	125	± 3.71	50
PHN-LS	± 3.34	83	± 3.70	44
ParetoMTL	± 2.93	951	± 2.03	467
COSMOS	± 3.34	60	± 3.73	80
COSMOS ACS	± 3.33	62	± 3.72	82
COSMOS ACS SoftMax	± 3.34	60	± 3.73	81
Method	**Multi-MNIST**		**Multi-Fashion**	
Single Task	± 2.88	645	± 2.21	641
PHN-EPO	± 2.88	2,370	± 2.19	1,964
PHN-LS	± 2.87	1,002	± 2.20	981
ParetoMTL	± 2.91	5,781	± 2.26	5,775
COSMOS	± 2.95	549	± 2.31	422
COSMOS ACS	± 2.94	675	± 2.30	664
COSMOS ACS SoftMax	± 2.94	682	± 2.29	679
Method	**Multi-Fashion+MNIST**			
Single Task	± 2.78	695		
PHN-EPO	± 2.81	2,074		
PHN-LS	± 2.76	1,033		
ParetoMTL	± 2.75	5,980		
COSMOS	± 2.83	752		
COSMOS ACS	± 2.80	712		
COSMOS ACS SoftMax	± 2.84	702		

The experimental results, as summarized in Table 1, reveal intriguing insights into the comparative performance of the baseline COSMOS method and its extended variants, COSMOS ACS and COSMOS ACS SoftMax, across the "Adult" and "Compass" datasets.

Interestingly, the results indicate minimal deviation in performance between the original COSMOS method and its extended counterparts for these specific datasets. Such uniformity in

performance suggests that the LS utilized in the original COSMOS method may be sufficiently adept at handling the optimization tasks presented by the "Adult" and "Compass" datasets. In terms of runtime, it is noteworthy that the original COSMOS demonstrates superior performance in the context of the "Multi-MNIST" and "Multi-Fashion datasets". However, its performance lags behind when applied to the "Multi-Fashion+MNIST" dataset. This observation hints at a potential trend: for more challenging tasks, the proposed variations of COSMOS might exhibit quicker convergence rates, as well as higher HV values (COSMOS ACS SoftMax, has a slighter higher HV value compared to original COSMOS for the "Multi-Fasion+MNIST" dataset). This observation prompts an interesting conjecture: while LS methods may encounter challenges when faced with complex Pareto front shapes, such as concave fronts, the relatively simple structures of the optimization landscapes in these datasets may not necessitate the adoption of such sophisticated scalarization techniques, such as augmented Chebyshev with and without SoftMax. Therefore, the comparable performance between the original COSMOS method and its extended variants underscores the importance of contextualizing algorithmic choices with respect to the specific characteristics of the optimization problem at hand. In this context, the suitability of LS for the "Adult" and "Compass" datasets raises key questions regarding the trade-offs between algorithmic complexity and applicability in MOO.

In essence, while the extended variants of COSMOS offer promising avenues for enhancing optimization performance, the near-parity observed in this study emphasizes the need for nuanced algorithm selection strategies tailored to the intricacies of individual optimization tasks.

6.2 Pareto fronts comparison on tabular data and Multi-MNIST

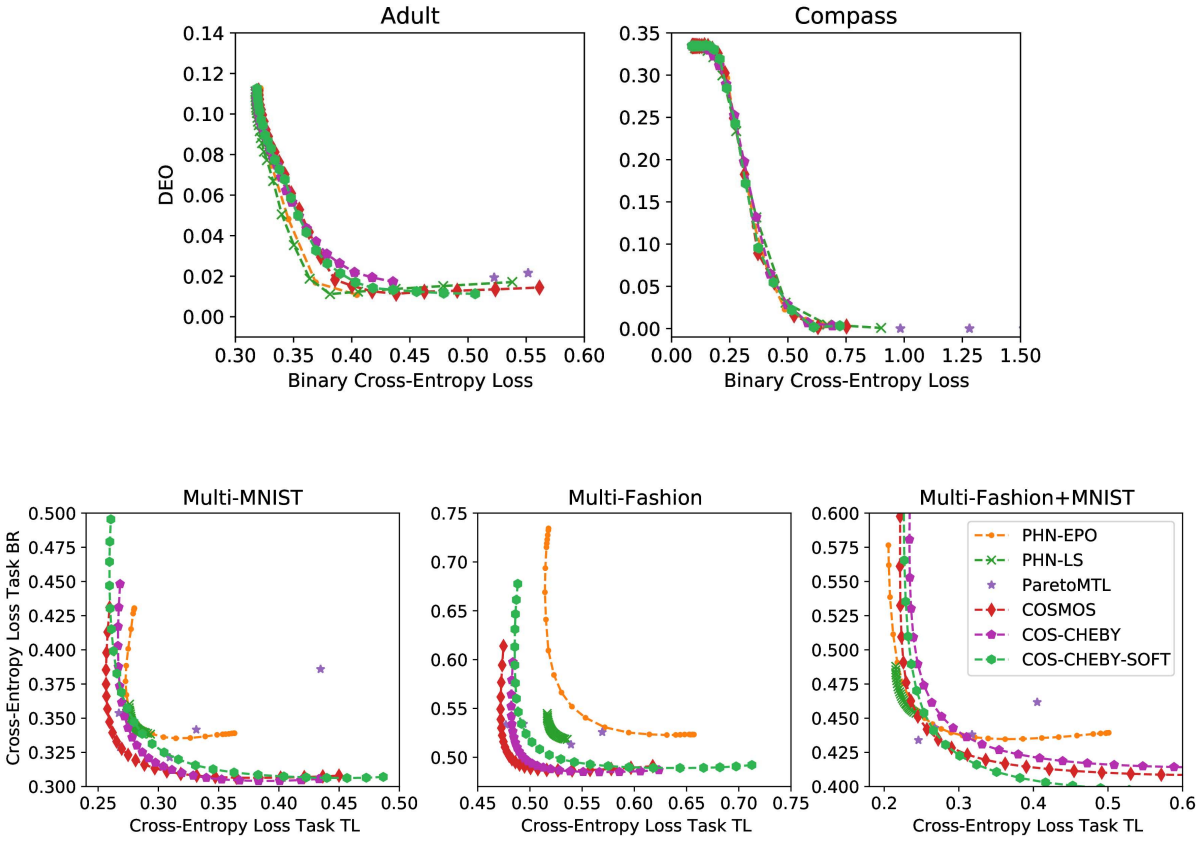

Figure 1: Comparison of the Pareto fronts on fairness and image classification datasets

In Figure 1 it is notable that for both the "Adult" and "Compass" datasets, the Pareto fronts exhibit similar characteristics across all methods. In the case of the "Multi-MNIST" dataset, the COSMOS family of methods demonstrates superior performance, with the original COSMOS method showing the most favorable loss. While the loss for the COSMOS ACS SoftMax method may slightly lag behind, its Pareto front exhibits a more favorable spread. Conversely, the COSMOS ACS method aligns closely with the original COSMOS method in terms of Pareto front, albeit with a higher loss. Similar trends are observed in the "Multi-Fashion" dataset. When analyzing the "Multi-Fashion+MNIST" dataset, it becomes apparent that the COSMOS ACS SoftMax method surpasses the original COSMOS method in certain scenarios. This suggests that for more challenging tasks, the COSMOS ACS SoftMax method might offer preferable outcomes.

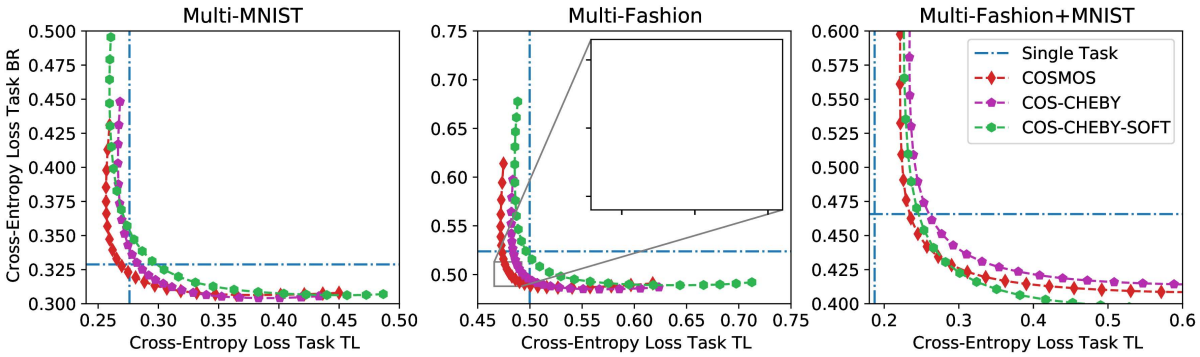

Figure 2: Comparison of the Pareto fronts between COSMOS and Single Task

In Figure 2, a comparison between the Pareto fronts of all COSMOS variants and the Single Task method can be seen. For the "Multi-Fashion" and "Multi-Fashion+MNIST" datasets, all COSMOS variants outperform the Single Task method. However, for the "Multi-MNIST" dataset, it is evident that the original COSMOS method clearly outperforms the Single Task, with the other variants of COSMOS also outperforming the latter for most solutions. However, visually, they exhibit a higher loss compared to the original COSMOS method.

6.3 Convergence of all COSMOS variants

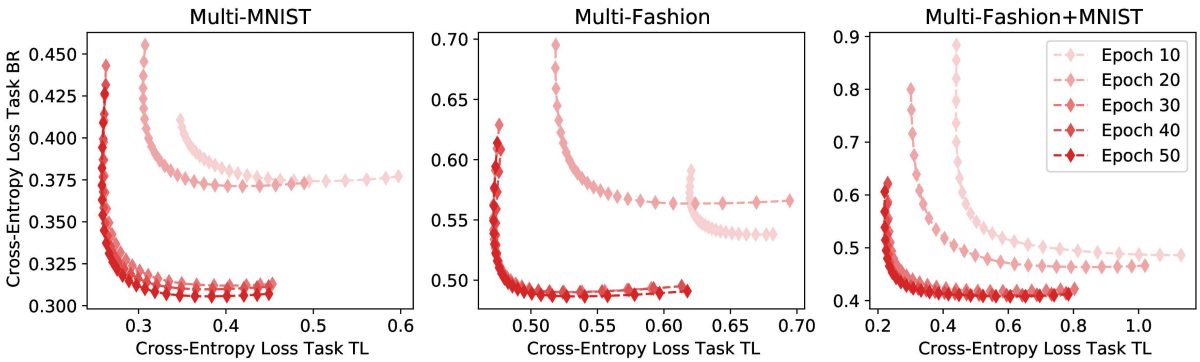

Figure 3: Convergence of the Pareto fronts generated by the original COSMOS method on the image classification datasets

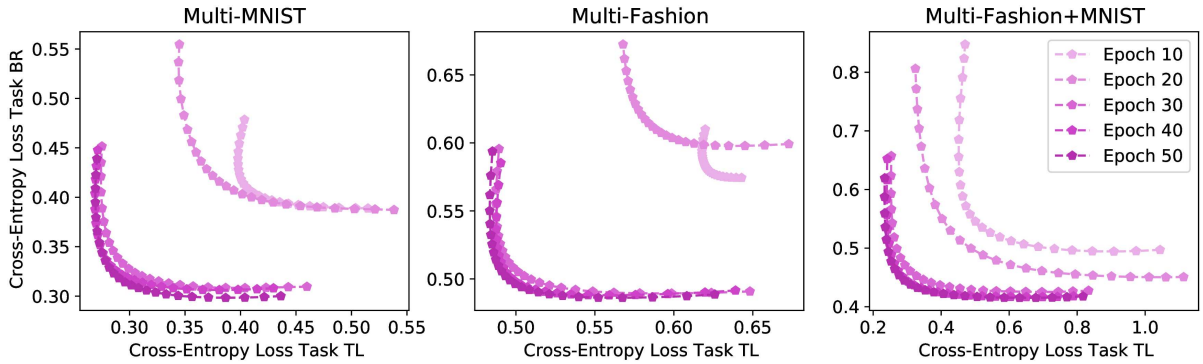

Figure 4: Convergence of the Pareto fronts generated by the COSMOS ACS method on the image classification datasets

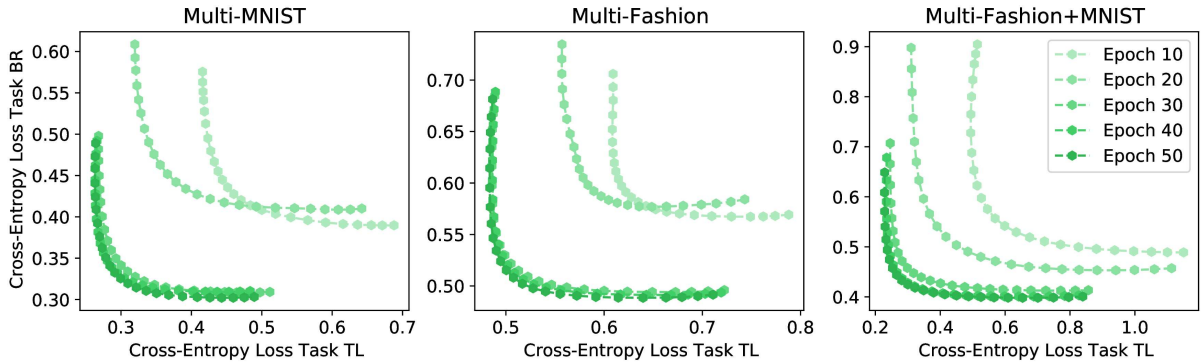

Figure 5: Convergence of the Pareto fronts generated by the COSMOS ACS SoftMax method on the image classification datasets

Figures 3, 4, and 5 demonstrate that all COSMOS methods converge fast and are able to generate well-spread Pareto fronts after 10 epochs.

7 Discussion & Conclusions

This study extends the foundational COSMOS method by introducing two novel scalarization variants: COSMOS ACS and COSMOS ACS with SoftMax, aiming to explore the efficacy of different scalarization techniques in handling MOO tasks across various datasets. Our investigation focuses on whether these extended scalarization methods can outperform the original COSMOS's LS in capturing complex Pareto front shapes, such as those which are concave.

The results presented in this paper highlight several key findings. First, the performance of COSMOS ACS and COSMOS ACS SoftMax is largely comparable to that of the original COSMOS across the "Adult" and "Compass" datasets. This similarity suggests that the simpler, LS approach of the original COSMOS is adequate for the types of optimization landscapes present in these datasets. For slightly more complex datasets like "Multi-Fashion+MNIST," where challenges typically arise in approximating the Pareto front, COSMOS ACS SoftMax shows a marginal improvement in the Hypervolume (HV) indicator. This observation implies that while the proposed alternative scalarization methods may offer theoretical advantages, their practical impact can be limited by the inherent characteristics of the dataset and the optimization landscape. It is, however, noteworthy to mention that the COSMOS ACS SoftMax variant outperformed COSMOS ACS across all datasets.

Regarding runtime efficiency, the alternative methods implemented align closely with the original COSMOS method, maintaining computational feasibility for calculating cosine similarity, which is crucial for penalizing the angle between achieved solutions in objective space and the inputted preference vectors. Notably, the convergence speeds of these methods vary with task complexity; they demonstrate faster convergence in the challenging "Multi-Fashion+MNIST" dataset and slightly slower rates in the simpler "Multi-MNIST" and "Multi-Fashion" datasets. This suggests that the alternative methods might be better suited for more complex and nuanced MOO problems where the diversity of solutions is critical.

The observation that all variants of COSMOS, including the original method, demonstrate robustness and effectiveness across diverse scenarios indicates the strength of the COSMOS framework in general. However, the nuanced differences observed suggest that while extended variants like COSMOS ACS and COSMOS ACS SoftMax provide valuable tools for optimiza-

tion, their benefits are most pronounced in scenarios where the complexity of the Pareto front challenges traditional methods.

The extended methods introduced here add to the toolkit available to practitioners, offering more flexible solutions depending on the specific requirements of their MOO problems. In essence, while these variants offer promising avenues for enhancing optimization performance, the near-parity observed in this study emphasizes the need for nuanced algorithm selection strategies tailored to the intricacies of individual optimization tasks.

In conclusion, this study not only reaffirms the versatility and efficacy of the COSMOS method but also enhances our understanding of how different scalarization approaches can be strategically utilized to optimize MOO learning tasks. While the benefits of COSMOS ACS and COSMOS ACS SoftMax are dataset-dependent, they represent meaningful additions to the field of MOO, particularly for complex tasks where traditional methods may falter. Future work could further explore the integration of other scalarization techniques and extend the application of these methods to a wider range of problems to fully ascertain their potential and limitations.

A Appendix

A.1 Multi-MNIST+Fashion

A.1.1 COSMOS ACS SoftMax Pareto front throughout different epochs

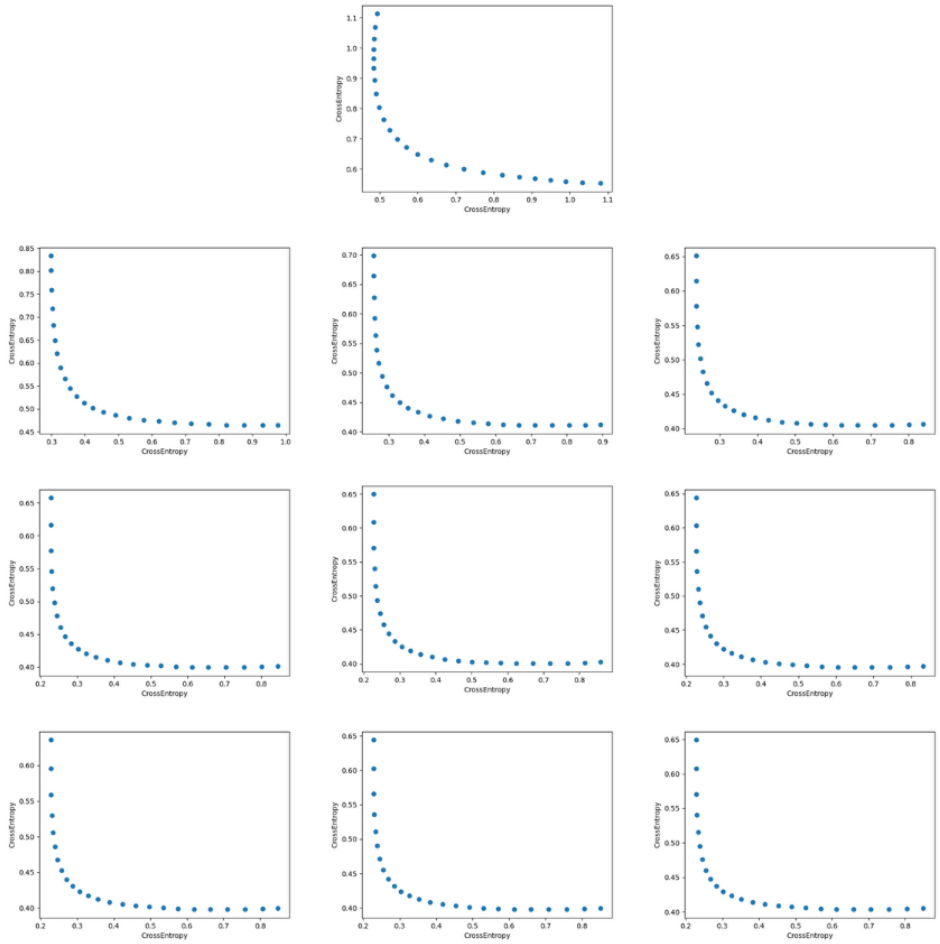

Figure 6: COSMOS ACS SoftMax Pareto fronts across epochs 1-100

A.1.2 COSMOS ACS Pareto front throughout different epochs

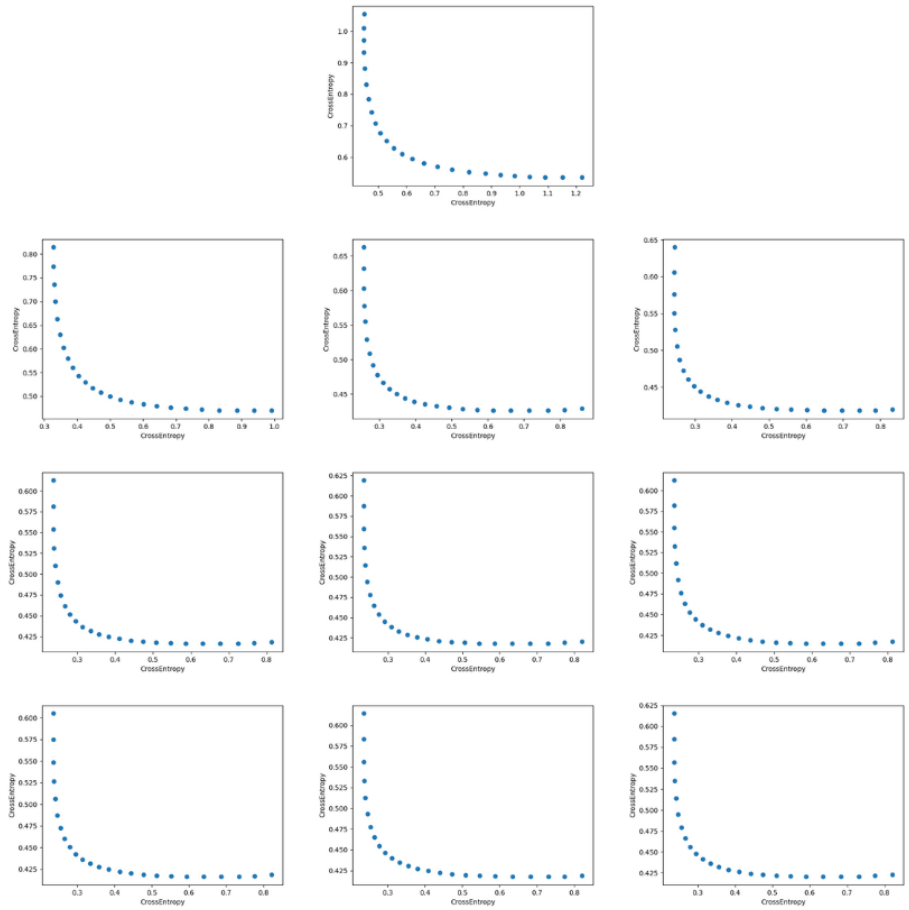

Figure 7: COSMOS ACS Pareto fronts across epochs 1-100

A.1.3 Original COSMOS Pareto front throughout different epochs

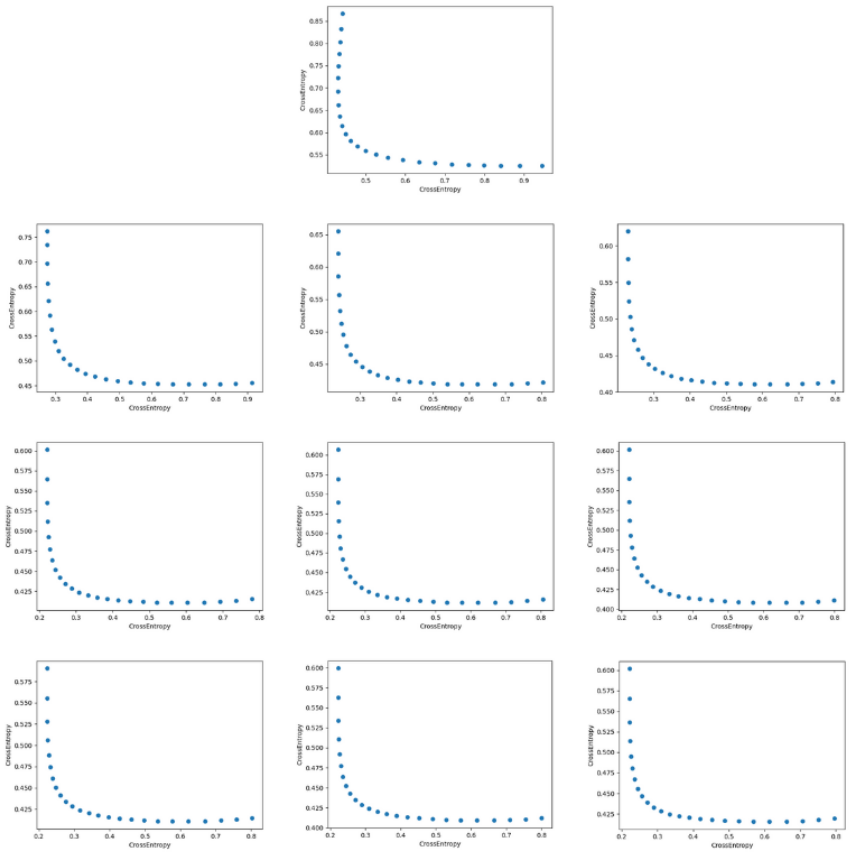

Figure 8: Original COSMOS Pareto fronts across epochs 1-100

References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2000). A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6:182–197.

Deist, T. M., Grewal, M., Dankers, F. J. W. M., Alderliesten, T., and Bosman, P. A. N. (2021). Multi-objective learning to predict pareto fronts using hypervolume maximization. *arXiv preprint arXiv:2102.04523*.

Desideri, J. (2012). Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6):313–318.

Dua, D. and Graff, C. (2017). UCI machine learning repository. `http://archive.ics.uci.edu/ml`.

Ehrgott, M. (2005). *Multicriteria Optimization*. Springer.

Emmerich, M. T. and Deutz, A. H. (2018). A tutorial on multiobjective optimization: Fundamentals and evolutionary methods. *Natural Computing*, 17(3):585–609.

Fliege, J. and Svaiter, B. F. (2000). Steepest descent methods for multicriteria optimization. *Mathematical Methods of Operations Research*, 51(3):479–494.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

LeCun, Y., Haffner, P., Bottou, L., and Bengio, Y. (1999). Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*, pages 319–345. Springer.

Lin, X. et al. (2019). Pareto multi-task learning. In *Advances in Neural Information Processing Systems*.

Mahapatra, D. and Rajan, V. (2020). Multi-objective deep learning for simultaneous optimization of classification accuracy and fairness criteria. *Journal of Machine Learning Research*, 21:1–36.

Padh, K., Antognini, D., Glaude, E. L., Faltings, B., and Musat, C. (2020). Addressing fairness in classification with a model-agnostic multi-objective algorithm. *arXiv preprint arXiv:2009.04441*.

Ruchte, M. and Grabocka, J. (2021). Scalable pareto front approximation for deep multi-objective learning. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 1306–1311, Auckland, New Zealand. IEEE.

Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic routing between capsules. *arXiv preprint arXiv:1710.09829*.

Sener, O. and Koltun, V. (2018). Multi-task learning as multi-objective optimization. In *Neural Information Processing Systems*.

Zhang, Y. and Yang, Q. (2018). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 30:1819–1837.

Zitzler, E., Brockhoff, D., and Thiele, L. (2007). The hypervolume indicator revisited: On the design of pareto-compliant indicators via weighted integration. In *International Conference on Evolutionary Multi-Criterion Optimization*, pages 862–876. Springer.