

FIRAS LAAKOM

# Feature Diversity in Neural Networks

Theory and Algorithms



FIRAS LAAKOM

# Feature Diversity in Neural Networks

## Theory and Algorithms

ACADEMIC DISSERTATION

To be presented, with the permission of  
the Faculty of Information Technology and Communication Sciences  
of Tampere University,  
for public remotely,  
on 18 March 2024, at 12 o'clock.

## ACADEMIC DISSERTATION

Tampere University, Faculty of Information Technology and Communication Sciences  
Finland

*Responsible  
supervisor  
and Custos*

Professor  
Moncef Gabbouj  
Tampere University  
Finland

*Supervisors*

Assistant Professor  
Jenni Raitoharju  
University of Jyväskylä  
Finland

Professor  
Alexandros Iosifidis  
Aarhus University  
Denmark

*Pre-examiners*

Associate Professor  
Sotirios Chatzis  
Cyprus University of Technology  
Cyprus

Assistant Professor  
Arno Solin  
Aalto University  
Finland

*Opponent*

Associate Professor  
Saikat Chatterjee  
KTH – Royal Institute of Technology  
Sweden

The originality of this thesis has been checked using the Turnitin Originality Check service.

Copyright ©2024 author

Cover design: Roihu Inc.

ISBN 978-952-03-3354-6 (print)

ISBN 978-952-03-3355-3 (pdf)

ISSN 2489-9860 (print)

ISSN 2490-0028 (pdf)

<http://urn.fi/URN:ISBN:978-952-03-3355-3>



Carbon dioxide emissions from printing Tampere University dissertations have been compensated.

PunaMusta Oy – Yliopistopaino  
Joensuu 2024

# PREFACE

The research work in this dissertation was conducted within the Signal Analysis and Machine Intelligence (SAMI) research group at Tampere University, Finland. Funding for this research was provided by the National Science Foundation (NSF), Center for Visual and Decision Informatics (CVDI), and Business Finland. The financial support of INTEL and XIAOMI is sincerely appreciated.

I want to express my gratitude to my supervisors, Prof. Moncef Gabbouj, Prof. Jenni Raitoharju, and Prof. Alexandros Iosifidis for their invaluable guidance throughout my doctoral studies. I am genuinely thankful for their sustained support during my academic journey. Their mentorship and steadfast belief in my capabilities have left an indelible mark, and I will remain eternally grateful for their unwavering support.

I extend my appreciation to Prof. Sotirios Chatzis from Cyprus University of Technology, Cyprus, and Prof. Arno Solin from Aalto University, Finland, for their detailed review of this dissertation, as well as for offering constructive feedback and valuable comments. I am also grateful to Prof. Saikat Chatterjee from KTH Royal Institute of Technology, Sweden, for agreeing to serve as the Opponent during the thesis defense.

I would also like to thank Prof. Yuheng Bu from the University of Florida, USA, for his support and warm welcome during my research visit at the University of Florida. I am also thankful to my other collaborators, Dr. Nikolaos Passalis and Dr. Jarno Nikkanen, who have provided great opportunities, resources, and suggestions to our joint works.

I would like to show my deepest appreciation to various members of our research group: Anton Muravev, Ali Senhaji, Aysen Degerli Ahishali, Bilge Can Pullinen, Cem Entok, Dr. Dat Thanh Tran, Emmi Antikainen, Dr. Farhad Pakdaman, Dr. Fahad Sohrab, Ilke Adalioglu, Junaid Malik, Kateryna Chumachenko, Lei Xu, Mehmet Yamaç, Mete Ahishali, Dr. Mohammad Fathi

Al-Sa'd, Muhammad Uzair Zahid, Mert Duman, Oumaïma Hamila, Özer Devecioglu, Sanaz Nami, and Tuomas Jalonen, with whom I have had a great working atmosphere, long coffee breaks, and discussions.

Finally, I would like to express my deepest gratitude to my family for their continued encouragement and support. I dedicate this thesis to them.

Tampere, February 2024

Firas Laakom

# ABSTRACT

The main strength of neural networks lies in their ability to generalize to unseen data. ‘*Why* and *when* do they generalize well?’ are two extremely important questions for a full understanding of this phenomenon and for developing better and more robust models. Several studies have explored these questions from different perspectives and proposed multiple measures/bounds that correlate well with generalization. The dissertation proposes a new perspective by focusing on the ‘feature diversity’ within the hidden layers. From this standpoint, neural networks are seen as a two-stage process, with the first stage being feature (representation) learning through the intermediate layers, followed by the final prediction layer. Empirically, it has been observed that learning a rich and diverse set of features is critical for achieving top performance. Yet, no theoretical justification exists. In this dissertation, we tackle this problem by theoretically analyzing the effect of the features’ diversity on the generalization performance. Specifically, we derive several Rademacher-based rigorous bounds for neural networks in different contexts and we demonstrate that, indeed, having more diverse features correlates well with better generalization performance. Moreover, inspired by these theoretical findings, we propose a new set of data-dependent diversity-inducing regularizers and we present an extensive empirical study confirming that the proposed regularizers enhance the performance of several state-of-the-art neural network models in multiple tasks. Beyond standard neural networks, we also explore different diversity-promoting strategies in different contexts, e.g., Energy-Based Models, autoencoders, and bag-of-features pooling layers and we show that learning diverse features helps consistently.





# CONTENTS

1	Introduction . . . . .	1
1.1	Motivation and Objectives . . . . .	1
1.2	Publications and Author’s Contributions . . . . .	6
1.3	Dissertation Outline . . . . .	9
2	Research Background . . . . .	11
2.1	Convolutional Neural Networks . . . . .	11
2.2	Bag-of-Features Pooling . . . . .	11
2.3	Autoencoders . . . . .	12
2.4	Energy-based Models . . . . .	13
2.5	Diversity in Machine Learning . . . . .	14
3	Contributions . . . . .	17
3.1	Feature Diversity in Neural Networks: Theory . . . . .	17
3.1.1	Problem Formulation . . . . .	17
3.1.2	Learning Distinct Features Helps, Provably . . . . .	20
3.1.3	Discussion . . . . .	22
3.2	Feature Diversity in Neural Networks: Algorithms . . . . .	23
3.2.1	WLD-Reg: A Data-Dependent Within-Layer Diversity Regularizer . . . . .	23
3.2.1.1	Methodology . . . . .	24
3.2.1.2	Empirical Results . . . . .	26
3.2.1.3	Discussion . . . . .	32
3.2.2	Diversity in BoF Pooling . . . . .	33

3.2.2.1	Methodology . . . . .	33
3.2.2.2	Empirical Results . . . . .	34
3.2.2.3	Discussion . . . . .	35
3.2.3	Feature Diversity in Autoencoders . . . . .	36
3.2.3.1	Methodology . . . . .	36
3.2.3.2	Empirical Results . . . . .	38
3.2.3.3	Discussion . . . . .	42
3.3	Feature Diversity in Energy-Based Models . . . . .	43
3.3.1	Feature Diversity in Energy-Based Models: Theory . . .	44
3.3.1.1	Problem Formulation . . . . .	44
3.3.1.2	Feature Diversity Improves the Generalization of EBMs . . . . .	45
3.3.1.3	Discussion . . . . .	49
3.3.2	Feature Diversity in Energy-Based Models: Algorithms .	49
3.3.2.1	Methodology . . . . .	49
3.3.2.2	Empirical Results . . . . .	50
3.3.2.3	Discussion . . . . .	56
3.4	Feature Diversity vs Class-wise Overfitting . . . . .	57
3.4.1	Feature Diversity and Class-wise Generalization . . . .	58
3.4.2	Class-wise Generalization: an Information-Theoretic Per- spective . . . . .	59
3.4.2.1	Problem Formulation . . . . .	60
3.4.2.2	Main Results . . . . .	61
3.4.2.3	Extra Applications . . . . .	65
3.4.3	Discussion . . . . .	68
4	Conclusions . . . . .	71
	References . . . . .	75
	Publication I . . . . .	93

Publication II . . . . .	113
Publication III . . . . .	125
Publication IV . . . . .	133
Publication V . . . . .	143
Publication VI . . . . .	157

## List of Figures

2.1	An illustration of an EBM used to solve (a) a regression (b) a classification (c) an implicit regression [Publication V] . . . . .	13
3.1	Visual illustration of the hypothesis class . . . . .	19
3.2	Generalization gap, i.e., train error - test error, and the theoretical bound, i.e., $(C_5^2 - d_{min}^2)/\sqrt{N}$ , as a function of the number of training samples on MNIST dataset for neural networks with intermediate layer sizes from left to right: 128 (correlation=0.9948), 256 (correlation=0.9939), and 512 (correlation=0.9953). The theoretical term has been scaled in the same range as the generalization gap. All results are averaged over 5 random seeds. [Publication I] . . . . .	22
3.3	Illustration of ‘ <b>within-layer</b> ’ feedback and the ‘ <b>between-layer</b> ’ feedback . . . . .	24
3.4	An illustration on how the BoF-based CNN model loss is computed using our approach. The standard loss can be least squares or cross entropy and the similarity loss corresponds to equation 3.10 [Publication III]. . . . .	34
3.5	An illustration of how the autoencoder loss is computed using our approach. . . . .	37
3.6	Average ( $K = 3$ )-NN accuracy as a function of the dimension of the bottleneck size $d$ on Isolet dataset. Results are averaged over 10 random seeds [Publication IV]. . . . .	40
3.7	Visualization of the training data for the 1-D regression tasks: The dataset [37] on the left and the dataset [129] on the right [Publication V]. . . . .	51

3.8	Qualitative results of our approach ( $\beta = 1e^{-13}$ ) : Few intermediate samples of the MCMC sampling (Langevin Dynamics). . .	54
3.9	Test classification accuracy vs number of observed tasks on CIFAR10 using the boundary-aware (left) and boundary-agnostic (right) setting. The results are averaged over ten random seeds.	57
3.10	Class-generalization errors of the different classes in ImageNet of different approaches: ResNet50 with no diversity regularizer (top left), ResNet50 with WLD-Reg (Direct) (top right), ResNet50 with WLD-Reg (Det) (bottom left), and ResNet50 with WLD-Reg (Logdet) (bottom right). . . . .	58
3.11	Experimental results of class-generalization error and our bounds in Theorems 6 and 7 for the class of “trucks” (left) and “cats” (right) in CIFAR10 (top) and noisy CIFAR10 (bottom), as we increase the number of training samples [Publication VI]. . . .	64
3.12	The scatter plots between the bound in Theorem 7 and the class-generalization error of the different classes for CIFAR10 (left) and noisy CIFAR10 (right) [Publication VI]. . . . .	65
3.13	Illustration of the Subtask problem. The source domain is composed of 5 different classes and the target domain is composed of data from two particular classes encountered during training in the source domain. . . . .	66

## List of Tables

3.1	Classification errors of the different methods on CIFAR10 and CIFAR100. Results are averaged over three random seeds [Publication II]. . . . .	28
3.2	Classification errors of different models with different diversity strategies on ImageNet dataset [Publication II]. . . . .	28
3.3	The generalization gap, i.e., training error - test error, of different approaches on ImageNet dataset. * denotes WLD-Reg variants [Publication II]. . . . .	29
3.4	Classification errors of ResNet50 using different diversity strategies on CIFAR10 and CIFAR100 datasets with different label noise ratios. Results are averaged over three random seeds [Publication II]. . . . .	30
3.5	Transfer learning performance on CIFAR10 and CIFAR100 of ResNet50 models pre-trained on ImageNet with the different diversity approaches. . . . .	31
3.6	Average error rates and standard deviation of different approaches for different number of filters in the last convolutional layer on the MNIST dataset. Results are averaged over 5 random seeds. The top results for each approach are in bold and the best global result is underlined [Publication III]. . . . .	35
3.7	Average error rates and standard deviation of different approaches for different numbers of filters in the last convolutional layer on the fashionMNIST dataset. Results are averaged over 5 random seeds. The top results for each approach are in bold and the best global result is underlined [Publication III]. . . . .	35

3.8	Average error rates and standard deviation of different approaches for different number of filters in the last convolutional layer on the CIFAR10 dataset. Results are averaged over three random seeds. Top results for each approach are in bold and best global result is underlined [Publication III]. . . . .	36
3.9	Statistics of the three datasets used in the dimensionality reduction experiments. # Dim: dimensionality of the data. # Train: number of training samples. # Test: number of test samples. d: projection dimension [Publication IV]. . . . .	38
3.10	Classification accuracy of Nearest Neighbor classifier applied on the bottleneck representations (average and standard deviation over 10 repetitions) [Publication IV]. . . . .	39
3.11	Autoencoder topology used for CIFAR10. * denotes the bottleneck representation [Publication IV]. . . . .	40
3.12	RMSE, PSNR, and SSIM on the image compression task with CIFAR10 dataset (average and standard deviation over 5 repetitions) [Publication IV]. . . . .	42
3.13	RMSE, PSNR, and SSIM on the image denoising task with CIFAR10 dataset (average and standard deviation over 5 repetitions)	43
3.14	Results of the EBM trained with NCE (EBM) and the EBM trained with NCE augmented with our regularizer (ours) for the 1-D regression tasks. We report the approximate KL divergence for the first dataset [37], and the approximate NLL for the second dataset [129]. For each dataset, we report the results for three different values of $\sigma_1$ . . . . .	52
3.15	Results in terms of approximate NLL for the EBM age estimation experiments. The results are reported as the mean/SEM over these runs. . . . .	53

3.16	Simple CNN model used in the example. * refers to the features' layer. . . . .	54
3.17	Table of FID scores and NLL loss of different approaches for generations of MNIST images. Each experiment was performed three times with different random seeds, the results are reported as the mean/SEM over these runs. . . . .	55
3.18	Evaluation of class-incremental learning on both the boundary-aware and boundary-agnostic setting on CIFAR10 and CIFAR100 datasets. Each experiment was performed ten times with different random seeds, the results are reported as the mean/SEM over these runs. . . . .	56
3.19	Different quantitative measures of the effect of different approaches on the class-generalization error of the model. . . . .	59



# ABBREVIATIONS

ANN	Artificial Neural Network
CL	Continual Learning
CMI	conditional mutual information
CNN	Convolutional Neural Network
e-CMI	evaluated conditional mutual information
EBM	Energy-Based Model
f-CMI	functional conditional mutual information
FID	Fréchet Inception Distance
i.i.d.	independent and identically distributed
KL	Kullback-Leibler
MI	Mutual Information
MLP	Multilayer Perceptron
MSE	Mean Squared Error
NCE	noise contrastive estimation
PDF	probability density function
PSNR	Peak Signal-to-Noise Ratio
ReLU	Rectified Linear Unit
RMSE	Root Mean Squared Error
SEM	Standard deviation of the mean
SGD	Stochastic Gradient Descent
SSIM	Structural Similarity Index

VC      Vapnik-Chervonenkis

# ORIGINAL PUBLICATIONS

- Publication I      F. Laakom, J. Raitoharju, A. Iosifidis and M. Gabbouj. Learning distinct features helps, provably. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 2023, 206–222. DOI: 10.1007/978-3-031-43415-0\_13.
- Publication II     F. Laakom, J. Raitoharju, A. Iosifidis and M. Gabbouj. WLD-Reg: A Data-Dependent Within-Layer Diversity Regularizer. *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 2023, 8421–8429. DOI: 10.1609/aaai.v37i7.26015.
- Publication III    F. Laakom, J. Raitoharju, A. Iosifidis and M. Gabbouj. Efficient CNN with uncorrelated Bag of Features pooling. *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*. 2022, 1082–1087. DOI: 10.1109/SSCI51031.2022.10022157.
- Publication IV    F. Laakom, J. Raitoharju, A. Iosifidis and M. Gabbouj. Reducing redundancy in the bottleneck representation of the autoencoders. *Pattern Recognition Letters* 178 (2024), 202–208. ISSN: 0167-8655. DOI: <https://doi.org/10.1016/j.patrec.2024.01.013>.
- Publication V      F. Laakom, J. Raitoharju, A. Iosifidis and M. Gabbouj. On Feature Diversity in Energy-based models. *Energy Based Models Workshop-ICLR*. 2021.

Publication VI    F. Laakom, Y. Bu and M. Gabbouj. Class-Wise Generalization Error: An Information-Theoretic Analysis. *arXiv preprint arXiv:2401.02904*. Submitted to ICML (2024).

# 1 INTRODUCTION

## 1.1 Motivation and Objectives

Over the past decade, neural networks in general and deep learning models, in particular, have emerged as powerful tools capable of learning complex patterns and representations from data [1, 2]. One of their key strengths lies in their remarkable capacity to generalize effectively to unseen data, a characteristic that underlies their success in several applications, e.g., image/text classification [3, 4], compression [5, 6], and generative tasks [7, 8, 9].

*“Why and When do neural networks exhibit a strong generalization?”* represents a pivotal inquiry essential for a comprehensive understanding of this phenomenon and more importantly for developing more efficient neural network models that do not overfit the training data [10, 11, 12]. Several studies have approached these questions from diverse perspectives, investigating the impact of dataset characteristics [13, 14], model structure [15, 16, 17], and optimization algorithms [18, 19]. This multifaceted exploration has led to the formulation of theoretical frameworks and empirical studies that contributed to unraveling the dynamics of neural networks’ generalization [12, 20]. In particular, to answer the aforementioned questions, several studies have proposed theoretical measures and bounds, based on Rademacher complexity [21, 22, 23], VC dimension [24, 25], and margin-based metrics, which have shown strong correlations with the generalization performance of neural networks.

This dissertation proposes a new research direction, which complements prior studies, by focusing on the “feature diversity” within the hidden layers of neural networks. Conceptually, deep networks learn hierarchical representations, capturing complex features at different levels of abstraction. This hierarchical structure allows them to model intricate patterns in the data, contributing to their ability to generalize. “Feature diversity”, as defined in this

dissertation, refers to the idea that the features, i.e., activations’ output, learned by different units/ neurons within an intermediate layer in a neural network should be diverse or distinct from each other. In other words, each unit should capture unique aspects or patterns of the input data, contributing complementary information to the overall representation.

Diversity has been extensively explored in the machine learning context [26]. In the domain of deep learning generalization, previous works have predominantly concentrated on studying the effect of diversity within the set of weights on generalization both theoretically [27, 28] and practically [29, 30, 31, 32]. However, the diversity of activations, i.e., feature diversity, has received comparatively limited attention. Here, we argue that due to the presence of non-linear activations: (i) weights can not capture the intrinsic properties of the learned mapping, and (ii) diversity based on weights does not guarantee a diverse feature representation. Thus, we advocate directing attention toward the diversity at the level of feature mapping to study generalization.

In essence, the main hypothesis in this dissertation posits that

*“The success of deep learning models hinges not only on their architecture and training algorithms but also on the diversity of the features they can effectively capture.”*

In other words, feature diversity in neural networks should play a pivotal role in their ability to generalize and avoid overfitting. The principal part of the research conducted in this dissertation delves into the critical importance of feature diversity, exploring how it influences the generalization of neural networks. By unraveling the dynamics of feature diversity, the aim is to uncover novel insights that can propel advancements in the field, ultimately contributing to a deeper understanding of generalization and developing more versatile and efficient neural network models. The first step in this endeavor is to provide a theoretical substantiation of our hypothesis. Thus, the first research question that will be addressed in this dissertation regarding establishing a theoretical foundation of feature diversity is the following:

**Research Question 1:** *Can we derive rigorous generalization bounds for neural networks highlighting the role of feature diversity?*

Building upon the theoretical validation of our hypothesis, the subsequent part of this dissertation focuses on developing practical methodologies to harness the concept of feature diversity. Having diverse features can be beneficial for several reasons. Firstly, it allows the neural network to have a richer and non-redundant representation, as each unit learns to capture a unique pattern, thereby enriching the overall representation. This can be important for learning intrinsic relationships in the data and improving the network’s capacity to capture complex patterns. Secondly, having diverse features can make the model more robust to variations and noise in the input data, as the model does not rely on a single pattern to make a decision. In [33], it has been shown empirically that learning decorrelated features can reduce overfitting. As reducing the correlations between the features can be interpreted as a feature diversity-promoting approach, this showcases that indeed feature diversity is a promising research direction to boost the performance of neural networks in general and Convolutional Neural Networks (CNNs) in particular. Thus, the second research question that will be addressed in this dissertation will involve developing new strategies that leverage diversity to mitigate overfitting and enhance the generalization capabilities of state-of-the-art CNN models:

**Research Question 2:** *Can we employ feature diversity-promoting strategies to improve the performance of neural networks in different learning scenarios?*

By tackling the first two research questions, we will explore and highlight the role of feature diversity both theoretically and practically in boosting the performance of neural networks mainly in the supervised learning context. In the next step of this dissertation, the aim is to extend both our theoretical and practical findings to encompass other learning settings. To this end, we focus on the energy-based learning paradigm [34].

Energy-Based Models (EBMs) form a powerful learning framework that encapsulates various supervised [34, 35, 36, 37, 38], unsupervised [39, 40, 41], and generative [7, 8, 9, 42, 43] approaches in a unified formulation. In particular, an EBM is typically formed of inner model(s) that learn a combination of the different features to generate an energy mapping for each input configuration. The parameters of the EBM are optimized by associating

the desired configurations with small energy values and the undesired ones with higher energy values. This flexibility of modeling provides a generic framework to tackle a wide range of tasks ranging from standard regression [36, 37, 38] to image generation [9, 44, 45, 46]. The third research question in this dissertation aims to extend both our theoretical and empirical findings to the energy-based learning framework and can be formulated as follows:

**Research Question 3:** *How does feature diversity affect the generalization of EBMs?*

The first three research questions discussed so far aim to demonstrate the role of feature diversity in the generalization capabilities of neural networks and showcase how different diversity-promoting strategies can be used to mitigate overfitting. Recently, it has been empirically observed that deep learning models do not generalize equally well for the different classes and there is a noticeable disparity of overfitting among different classes [14, 47], i.e., given a task, neural networks tend to exhibit a class-bias in overfitting having high variance in generalization performance among the different classes. Moreover, [14, 47] showed that while several regularization techniques improve standard average generalization and reduce overall overfitting, these techniques inadvertently aggravate this occurrence increasing the disparity of generalization among different classes. It follows that, for a more comprehensive analysis of feature diversity’s effect, we analyze how the approaches developed in this dissertation affect class-wise generalization performance. Thus, the fourth research question can be formulated as follows:

**Research Question 4:** *How does feature-diversity promoting approaches affect class-wise generalization performance?*

The phenomenon of the disparity of overfitting among different classes poses a substantive puzzle within the domain of deep learning. Understanding the root causes of such disparity is crucial for refining and advancing deep learning theory and methodologies. However, existing generalization theories of supervised learning typically take a holistic approach and provide bounds



for the expected generalization over the whole data distribution. Thus, they can not capture the aforementioned behavior and can not provide insights into this generalization puzzle. In the last part of this dissertation, we aim to amend this gap in the literature by providing the first rigorous theoretical framework for studying and understanding this phenomenon. Thus, the last research question that will be addressed in this dissertation is the following:

**Research Question 5:** *Can we develop a theory of class-wise generalization?*

To sum up, the contributions of this dissertation are as follows:

- Theoretically, we introduce the concept of feature diversity and provide the first rigorous bounds highlighting its effect on the generalization of neural networks in different contexts.
- Methodologically, we propose a new family of regularizers that aim to encourage the ‘diversification’ of the layers’ output feature maps in neural networks.
- We extend our theoretical findings to EBMs and showcase how to practically leverage feature diversity to improve the performance of EBMs in several tasks.
- We empirically analyze how the regularizers developed in this dissertation, in particular the proposed approaches in Publication II, affect the disparity of overfitting among different classes.
- We develop theoretical tools aiming to study this generalization puzzle in general and we derive several rigorous bounds that successfully capture its behavior in deep learning models. Additionally, we show how to use the newly developed tools to provide insights into other contexts beyond this phenomenon, e.g., the subtask problem, and learning with sensitive attributes.

## 1.2 Publications and Author’s Contributions

**Publication I** This publication presents the first theoretical investigation of how feature diversity improves generalization. We study the diversity of the features learned by a two-layer neural network trained with the least squares loss. We investigate how learning non-redundant distinct features affects the performance of the network. We derive novel generalization bounds depending on feature diversity based on Rademacher complexity for such networks. Our analysis provides theoretical guarantees that more distinct features at the network’s units within the hidden layer lead to better generalization. We also show how to extend our results to deeper networks and different losses.

The candidate proposed the idea, formulated the theoretical setup, derived the analysis, and wrote the paper. The co-authors have supervised, reviewed, and edited the publication.

**Publication II** This publication presents the first methodology to promote the ‘diversification’ of the layer-wise feature map outputs in neural networks. The primary contribution is the introduction of a new family of data-dependent regularizers with the explicit purpose of encouraging feature diversity and reducing redundancy within the feature layer. Furthermore, an extensive experimental analysis has been conducted to demonstrate the efficacy of this approach. The results illustrate that implementing such a strategy significantly boosts the performance of different state-of-the-art networks across different datasets and different tasks, i.e., image classification and label noise. These findings can spark further research in diversity-based approaches to improve the performance of neural networks and mitigate overfitting.

The candidate contributed to the proposal of this work, implemented and conducted experimental analysis, and wrote

the paper. The co-authors have supervised and reviewed the publication.

**Publication III** This publication investigates feature diversity within the context of Bag of Features (BoF) pooling. Specifically, it introduces an approach that extends BoF pooling to enhance its efficiency by ensuring non-redundancy among the items in the learned dictionary. The proposed method introduces an additional loss term based on pair-wise correlations among dictionary items. This supplementary loss term, in conjunction with the standard loss, serves to explicitly regularize the model to ensure learning a more diverse and rich dictionary. Experimental results substantiate that the BoF can benefit from feature diversity regularization to enhance performance without the need for additional parameters. The candidate proposed the idea, implemented and conducted the experimental analysis, and wrote the paper. The co-authors supervised, reviewed, and edited the publication.

**Publication IV** This work investigates feature diversity within the context of Autoencoders. The main contribution is introducing a methodology for reducing redundancies at the bottleneck of an autoencoder. The proposed approach involves augmenting the training loss with an additional regularization term, specifically targeting the pair-wise covariances of units at the bottleneck, i.e., the encoder’s output. The proposed regularizer ensures learning more diverse and compact representations for input samples. Furthermore, it can be seamlessly integrated into any autoencoder-based model in a plug-and-play fashion. Through extensive empirical evaluations on multiple tasks—dimensionality reduction, compression, and denoising—we substantiate the efficacy of the approach. The candidate proposed the idea, conducted the experiments, and wrote the paper. The remaining co-authors supervised and reviewed the publication.

**Publication V** This publication explores feature diversity in the context of energy-based models. The primary contribution is introducing the concept of feature diversity in this context and using it to theoretically analyze the generalization of EBMs. Specifically, we derive different generalization bounds for various learning contexts, i.e., regression, classification, and implicit regression, with different energy functions and show that reducing the redundancy of the feature set can consistently improve the generalization ability and reduce overfitting. Furthermore, the theory developed in this work is independent of the loss function or the training strategy used to optimize the parameters of the EBM. This provides a broader theoretical guarantee that feature diversity helps. The candidate proposed the idea, conducted the theoretical analysis, and wrote the paper. The remaining co-authors supervised and reviewed the publication.

**Publication VI** This paper tackles the deep learning generalization puzzle concerning the pronounced heterogeneity of overfitting observed across different classes. Conventional generalization theories of supervised learning typically adopt a holistic approach, providing bounds for the expected generalization over the whole data distribution, implicitly assuming that the model generalizes similarly for all the classes. However, empirical observations reveal significant variations in generalization performance among different classes, which cannot be captured by the existing generalization bounds. In this work, we close this gap by introducing and exploring the concept of “class-generalization error” using information-theoretic tools. In particular, we provided the first rigorous generalization bounds for this concept. We also empirically strengthened the findings with supporting experiments validating the efficiency of the proposed bounds. Furthermore, we showcase the versatility of the theoretical tools, developed in this work, in providing tight bounds for various con-

texts, e.g., the subtask problem, generalization certificates with sensitive attributes, recall & specificity generalization. The candidate contributed to the proposal of the idea, contributed to the theoretical analysis, conducted the experiments, and wrote the paper. The co-authors have supervised and reviewed the publication.

## 1.3 Dissertation Outline

The rest of this dissertation are structured as follows:

Chapter 2 presents the research background. Chapter 3 examines the dissertation contributions and highlights their role and significance to the body of knowledge. Furthermore, this chapter serves as a valuable augmentation to the candidate's previously published works by providing supplementary insights, presenting additional results, and facilitating comparisons. Finally, Chapter 4 serves as the final segment of the dissertation, providing a comprehensive summary of the research findings and offers insights into the limitations and potential avenues for future research directions.



## 2 RESEARCH BACKGROUND

### 2.1 Convolutional Neural Networks

The field of image classification has witnessed significant advancements after the emergence of CNNs [3]. Deep CNN architectures revolutionized image classification on large datasets [4, 48, 49]. The fundamental elements of CNNs include convolutional layers, which employ learnable filters to systematically scan input data, capturing local patterns and features. Pooling layers further reduce spatial dimensions, retaining essential information while promoting translation invariance. Additionally, CNNs often incorporate fully connected layers for global abstraction and classification. The use of weight sharing in convolutional layers enhances the network’s ability to recognize spatially invariant patterns across the input space, enabling CNNs to achieve state-of-the-art performance in several tasks, e.g., image classification and object detection. Furthermore, residual learning [4, 50, 51] addressed challenges associated with training very deep networks, leading to improved performance. One of the main limitations of standard CNN models is the fact that they typically require large amounts of labeled training data to effectively learn and generalize well. To mitigate this problem, several techniques have been proposed, e.g., pretraining [52, 53, 54], data augmentation [13, 55, 56, 57]. In addition, several other hand-crafted topologies/layers have been proposed to improve the efficiency of CNNs and reduce their dependency on the availability of data, such as BoF Pooling [58].

### 2.2 Bag-of-Features Pooling

In this section, we provide a brief review of the Bag of Features (BoF) pooling mechanism [58, 59], a technique that has garnered broad applicability across

diverse domains, consistently demonstrating superior performance in various studies [59, 60, 61, 62, 63, 64]. BoF pooling is parameterized with a dictionary, and given an input, typically the output maps of the last convolutional layer in a CNN [59], it generates a histogram representation based on this dictionary. During the training phase, the optimization of dictionary items is conducted through standard back-propagation.

The BoF pooling encompasses two inner layers: a Radial Basis Function (RBF) layer, used to measure the similarity of input features to RBF centers, and an accumulation layer, used to construct a histogram of quantized feature vectors. Formally, if  $\mathbf{X}$  denotes the input image and  $T(\mathbf{X}) \in \mathbb{R}^{D \times P}$  signifies the output of the convolutional layer, the RBF layer yields a sequence of quantized representations denoted as  $\Psi = [\psi_1, \psi_2, \dots, \psi_P] \in \mathbb{R}^{K \times P}$ , where  $\psi_i$  corresponds to the representation of the  $i^{th}$  feature, i.e.,  $\psi_i = [\psi_{i,1}, \dots, \psi_{i,K}]$ .

The output of the  $i^{th}$  RBF unit is defined by the expression:

$$\psi_{n,i} = \frac{\exp\left(-\frac{\|T(\mathbf{X})_n - \mathbf{c}_i\|}{m_i}\right)}{\sum_j \exp\left(-\frac{\|T(\mathbf{X})_n - \mathbf{c}_j\|}{m_j}\right)}, \quad (2.1)$$

where  $\mathbf{c}_i$  represents the center of the  $i^{th}$  RBF neuron, and  $m_i$  is a scaling factor. The outputs of the  $P$  RBF neurons are subsequently accumulated in the ensuing layer to derive the final representation  $\Phi$  for each image:

$$\Phi = \frac{1}{P} \sum_j \psi_j. \quad (2.2)$$

In summation, BoF accepts a high-dimensional feature representation, e.g., convolution’s output map, as input and discretizes it into a fixed-size shallow histogram representation. The quantization process relies on an inner dictionary,  $\{\mathbf{c}_1, \dots, \mathbf{c}_K\}$ , which can be learned jointly with the other parameters CNN in an end-to-end manner.

## 2.3 Autoencoders

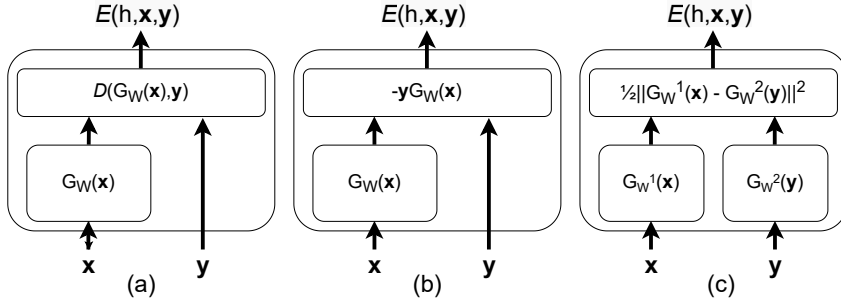
Autoencoders [5, 65] are unsupervised learning models designed for feature learning and data representation [2]. The architecture typically comprises



an encoder network responsible for compressing input data into a lower-dimensional latent space and a decoder network that reconstructs the original input from this encoded representation. The bottleneck, i.e., the output of the encoder, typically has a low dimension and is the focal part of the autoencoder. Training an autoencoder involves minimizing the reconstruction error and encouraging the model to learn a compact and informative representation of the input data. This framework has proven to be versatile, finding applications in diverse domains, such as transfer learning [66, 67, 68], dimensionality reduction [6, 69, 70], denoising [71, 72], and anomaly detection [73, 74, 75].

## 2.4 Energy-based Models

EBMs constitute a prominent class of learning models that have garnered substantial attention due to their versatile applications, such as regression [36, 37, 38], learning to rank [35], image/text generation [39, 44, 46], continual learning [76], anomaly detection [77], protein conformations [78], and reinforcement learning [79, 80].



**Figure 2.1** An illustration of an EBM used to solve (a) a regression (b) a classification (c) an implicit regression [Publication V]

Let  $E(h, \mathbf{x}, \mathbf{y})$  denote an energy-based model with an inner model  $h = G_{\mathbf{W}}(\mathbf{x})$  parameterized with  $\mathbf{W}$ . Figure 2.1 shows how different learning problems, i.e., classification, regression, and implicit regression can be solved with EBMs. Figure 2.1 shows how different learning tasks are solved with EBMs. The first input  $\mathbf{x}$  undergoes transformation through an inner model  $G_{\mathbf{W}}(\cdot)$ . Subsequently, the energy function is computed as the distance to the second input  $\mathbf{y}$  with a valid energy function. For example,  $L_1$  or  $L_2$  can be used as en-

ergy functions for regression,  $E(h, \mathbf{x}, \mathbf{y}) = -yG_{\mathbf{W}}(\mathbf{x})$  for binary classification, and  $L_2$  distance between the transformed inputs for implicit regression.

## 2.5 Diversity in Machine Learning

Diversity-based approaches have played a pivotal role across various facets of machine learning [81, 82]. In ensemble learning, the combination of models employing different learning algorithms or architectures ensures a comprehensive capture of intricate patterns within the data [32, 83]. Sampling strategies, encompassing exact sampling [84] and batch sampling [85], strategically introduce diversity in training instances, enriching the model’s understanding of the data distribution [86]. In the context of ranking, the integration of diversity is imperative to produce balanced and unbiased rankings, as exemplified by approaches such as balanced ranking [87] and methods enhancing diversity in the ranking process [81]. In the pruning context, dynamic pruning [88] and filter pruning [89, 90] leverage diverse strategies to reduce redundancy and enhance model efficiency without compromising performance [91].

Within the context of neural network regularization, several methodologies have incorporated diversity as a direct regularizer applied to the weight parameters [29, 32, 92]. This overview categorizes these approaches into two distinct groups based on the definition of diversity. The first category encompasses regularizers relying on the pairwise dissimilarity of components, asserting that the overall set of weights is diverse if every pair of weights is dissimilar. Given weight vectors  $\{\mathbf{w}_m\}_{m=1}^M$ , [32] define the regularizer as  $\sum_{mn}(1 - \theta_{mn})$ , where  $\theta_{mn}$  denotes the cosine similarity between  $\mathbf{w}_m$  and  $\mathbf{w}_n$ . [92] proposes an incoherence score expressed as

$$-\log \left( \frac{1}{M(M-1)} \sum_{mn} \beta |\theta_{mn}|^{\frac{1}{\beta}} \right), \quad (2.3)$$

incorporating a positive hyperparameter  $\beta$ . [93, 94] utilize  $\text{mean}(\theta_{mn}) - \text{var}(\theta_{mn})$  as a regularization measure for Boltzmann machines, with theoretical analyses on its impact on generalization error bounds presented in [28] and further extended to kernel space in [27]. The second group of regularizers adopts a broader perspective on diversity. For instance, in works such as [29, 95, 96],

weight regularization based on the determinant of the weights covariance is proposed, while [97, 98] explore a determinantal point process-based approach.

In contrast to the previously mentioned approaches that advocate for diversity at the level of weights, akin to our methodology, [33] introduce a novel strategy to impose dissimilarity at the feature map outputs, specifically on the activations. In pursuit of this objective, they introduced a supplementary loss function grounded in the pairwise covariance of the activation outputs. The proposed regularizer  $L_{Decov}$  is defined as the squared sum of the non-diagonal elements of the global covariance matrix  $\mathbf{C}$ :

$$L_{Decov} = \frac{1}{2}(\|\mathbf{C}\|_F^2 - \|\text{diag}(\mathbf{C})\|_2^2), \quad (2.4)$$

where  $\|\cdot\|_F$  is the Frobenius norm. The approach in [33], denoted Decov, demonstrated promising empirical performance of feature diversity. However, a theoretical substantiation was lacking. Addressing this gap, one of the main contributions of this dissertation provides theoretical underpinnings, elucidating how feature diversity can effectively reduce the estimation error bound and enhance the model’s generalization capacity. Furthermore, we introduce several approaches that leverage feature diversity to boost the performance of neural networks in multiple contexts.



# 3 CONTRIBUTIONS

This chapter gives a detailed description of the contributions of this dissertation. In Section 3.1, we present the main theoretical results of Publication I. In Section 3.2, we present the different feature diversity methodologies from Publication II, Publication III, and Publication IV, where different approaches in the context of standard neural networks, BoF-based models, and autoencoders are presented, respectively. In Section 3.3, we present the main theoretical results of Publication V extending our results to EBMs. Furthermore, in this dissertation, we also propose an approach that leverages feature diversity to boost the performance of EBMs across multiple tasks. Finally, Section 3.4 presents our study on how the proposed regularizer in Publication II affect class-wise generalization and presents the findings of Publication VI.

## 3.1 Feature Diversity in Neural Networks: Theory

The dissertation’s contribution presented in this section provides an answer to Research Question 1, which concerns providing a theoretical understanding of the effect of feature diversity on the generalization performance of neural networks. In the following, we will describe the main results which were proposed in Publication I.

### 3.1.1 Problem Formulation

Here, we aim to derive theoretical generalization bounds for neural networks depending on feature diversity. Specifically, we consider the case of a regression task with a two-layer neural network architecture comprising a hidden layer with  $M$  neurons and a one-dimensional output  $y$ .

In the regression task, we typically have access to training data  $S$  consisting

of  $N$  i.i.d. samples  $z_i = (\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y} \triangleq \mathcal{Q}$ . Given a network  $f(\cdot)$  from the hypothesis class  $\mathcal{F}$ , the main goal of generalization theory is to study the interplay between two critical metrics: the empirical loss and the anticipated risk defined respectively as follows:

$$\hat{L}(f) = \frac{1}{N} \sum_{i=1}^N l(f(\mathbf{x}_i), y_i), \quad (3.1)$$

$$L(f) = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{Q}}[l(f(\mathbf{x}), y)], \quad (3.2)$$

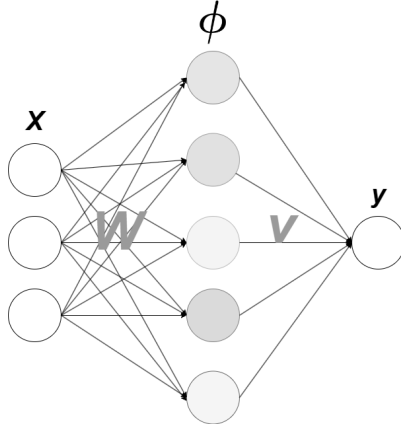
where the empirical loss  $l(\cdot, \cdot)$  measures the performance of the model on the observed dataset, while the expected risk quantifies the error on the true distribution, i.e., anticipated performance on unseen data. The objective is to discern and quantify the extent to which the model's performance on observed data is indicative of its performance on previously unseen instances. Let  $f^* = \arg \min_{f \in \mathcal{F}} L(f)$  be the expected risk minimizer and  $\hat{f} = \arg \min_{f \in \mathcal{F}} \hat{L}(f)$  be the empirical risk minimizer. We are interested in the estimation error, i.e.,  $L(f^*) - L(\hat{f})$ , defined as the gap in the loss between both minimizers [99]. This gap quantifies how well a trained model, based on a finite set of observed data, generalizes to unseen or future data [100, 101]. Several techniques have been used in the literature to study this generalization error, such as VC dimension [24] and the Rademacher complexity [102]. Our main goal in this part is to derive a bound for this quantity that highlights the role of diversity in the generalization dynamics.

In our case, the hypothesis class is formed of two-layer neural networks. So it can be expressed as follows

$$\mathcal{F} = \left\{ f \mid f(\mathbf{x}) = \sum_{m=1}^M v_m \phi_m(\mathbf{x}) = \sum_{m=1}^M v_m \rho(\mathbf{w}_m^T \mathbf{x}) \right\}, \quad (3.3)$$

where  $\rho(\cdot)$  is the activation function in the hidden layer,  $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M] \in \mathcal{R}^{D \times M}$  is the weight matrix connecting the input to the hidden layer with  $M$  units,  $\Phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \dots, \phi_M(\mathbf{x})]$  is the  $M$ -dimensional feature representation (intermediate layer output) of the input  $\mathbf{x}$ , and  $\mathbf{v} = [v_1, v_2, \dots, v_M]$  is the weight vector connecting the hidden layer to the output. In Figure 3.1, we provide a visual illustration of the hypothesis class used in

our setup.



**Figure 3.1** Visual illustration of the hypothesis class

Our hypothesis in this dissertation is that learning a rich and diverse set of features should be important to achieve good performance. Intuitively, if each unit within  $\Phi(\mathbf{x})$  captures a unique and distinct pattern of the input data, it will contribute complementary information to the overall representation yielding a more robust model with better generalization capabilities. The diversity of the features can be quantified using the lower-bound of the average pairwise  $L_2$  distance between the outputs as expressed in the following assumption:

**Assumption 1.** *Given any input  $\mathbf{x}$ , we have*

$$\frac{1}{2M(M-1)} \sum_{i \neq j}^M (\phi_i(\mathbf{x}) - \phi_j(\mathbf{x}))^2 \geq d_{min}^2. \quad (3.4)$$

The average  $L_2$  distance provides a straightforward way to quantify diversity. In case the mappings captured by two different units are redundant, then, given the same input sample, both units would return similar values. This similarity translates into low  $L_2$  distance, consequently yielding lower  $d_{min}$  and lower diversity. Conversely, when each unit learns a distinct mapping, the distances between the outputs of different units within the layer become substantial. Thus, this yields a higher lower bound  $d_{min}$  and a high global diversity. So,  $d_{min}$  can be used as a proxy for the feature diversity of the model. Through an examination of how the lower bound  $d_{min}$  influences the model’s generalization,

we can theoretically analyze the impact of diversity on the performance of neural networks. The subsequent part of this section is dedicated to deriving generalization bounds for neural networks depending on this measure, i.e.,  $d_{min}$ .

The key assumptions of the theoretical setup in Publication I can be summarized as follows:

- **Training dataset:**  $S$  consisting of  $N$  i.i.d. samples  $z_i = (\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y} \triangleq \mathcal{Q}$
- **Input/output:** The input satisfies  $\|\mathbf{x}\|_2 \leq C_1$  and the output satisfies  $|y| \leq C_2$ .
- **Hypothesis:** Two-layer neural networks:

$$\mathcal{F} \triangleq \left\{ f | f(\mathbf{x}) = \sum_{m=1}^M v_m \phi_m(\mathbf{x}) \right\}$$

- **Loss function:** Function  $\ell = \frac{1}{2} |f(\mathbf{x}) - y|^2$
- **Intermediate activation:** Positive  $L_\rho$ -Lipschitz continuous function
- **Weight norms:** First weight matrix  $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M]$  satisfies  $\|\mathbf{w}_m\|_2 \leq C_3$  and second weight vector satisfies  $\|\mathbf{v}\|_\infty \leq C_4$ .

### 3.1.2 Learning Distinct Features Helps, Provably

We derive a rigorous bound for the generalization error of a two-layer neural network. The main result is presented in Theorem 1.

#### Theorem 1: Generalization bound for regression

With probability of at least  $(1 - \delta)$ , we have

$$L(\hat{f}) - L(f^*) \leq \left( \sqrt{\mathcal{J}} + C_2 \right) \frac{A}{\sqrt{N}} + \frac{1}{2} (\sqrt{\mathcal{J}} + C_2)^2 \sqrt{\frac{2 \log(2/\delta)}{N}}, \quad (3.5)$$

where  $C_5 = L_\rho C_1 C_3 + \phi(0)$ ,  $\mathcal{J} = C_4^2 (M C_5^2 + M(M-1)(C_5^2 - d_{min}^2))$ ,  $A = 4 \left( 2 L_\rho C_1 C_3 C_4 + C_4 |\phi(0)| \right) M$ , and  $C_5 = L_\rho C_1 C_3 + \phi(0)$ .



The proof of Theorem 1 relies on Lemma 3 in Publication I which upper-bounds the generalization error using the Rademacher complexity [21, 102] and the supremum of the loss class. The key idea is to upper-bound the different quantities in Lemma 3 in Publication I using  $d_{min}$ .

The upper-bound for the generalization gap, presented in Theorem 1, offers a pivotal insight into the role of feature diversity in generalization. The bound is a decreasing function with respect to  $d_{min}$  scaling as  $\sim (C_5^2 - d_{min}^2)/\sqrt{N}$ . Remarkably, an increase in  $d_{min}$ , indicative of a greater diversity in learned features, results in a correspondingly lower generalization error bound. This inverse dependency underscores the significance of learning distinct and diverse features in the context of neural networks.

We note that the bound in Theorem 1 converges to zero as the number of training samples  $N$  goes to infinity. From this perspective, it is non-vacuous. Furthermore, it is essential to clarify that here we are not claiming a tighter bound for the generalization error of neural networks in the universal sense, similar to previous works [103, 104, 105]. Rather, our principal contribution lies in the derivation of a generalization bound depending on the diversity of learned features, quantified by  $d_{min}$ . Notably, this work represents a pioneering effort in conducting such theoretical analysis using the average L2-distance between units within the hidden layer as  $d_{min}$ , offering a novel perspective highlighting the effect of diversity.

In Publication I, we show that the result of Theorem 1 can be extended to derive rigorous bounds for the classification task (Theorems 2 and 3 in Publication I), for general multi-layer networks (Theorem 4 in Publication I), and multi-dimensional output (Theorems 5 and 6 in Publication I).

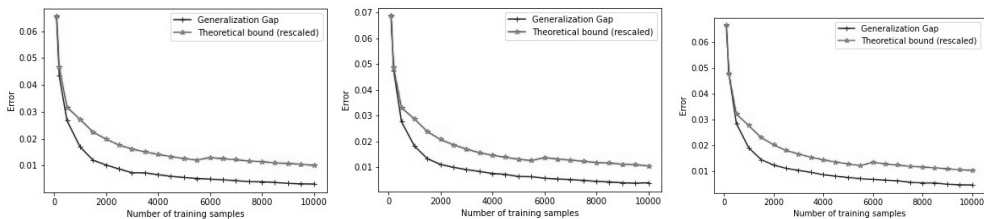
## Empirical Validation

We empirically validate our findings in Theorem 1, which suggests that generalization error scales as  $\sim (C_5^2 - d_{min}^2)/\sqrt{N}$ . We trained a two-layer neural network on the MNIST dataset [106], which is formed of grayscale images of size  $28 \times 28$  pixels. The inputs are vectorized to form 784-dimensional vectors. The dataset has a total of 50,000 training samples and 10,000 test images. For the intermediate layer in the neural network we use ReLU activation func-

tion. The models are trained for 100 epochs using Stochastic Gradient Descent (SGD) with a learning rate of 0.1 and a batch size of 256. The generalization gap, defined as the difference between the test error and the train error, was compared with the theoretical bound  $(C_5^2 - d_{min}^2)/\sqrt{N}$ , for different training set sizes. The quantity, i.e.,  $d_{min}$ , can be estimated, using the minimum average  $L_2$  distance over the training data, as follows:

$$\hat{d}_{min} = \min_{x \in S} \frac{1}{2M(M-1)} \sum_{n \neq m}^M (\phi_n(\mathbf{x}) - \phi_m(\mathbf{x}))^2. \quad (3.6)$$

Different sizes of the hidden layer, specifically 128, 256, and 512, were considered in our experiments. The averaged results from 5 random seeds are presented in Figure 3.2, demonstrating a consistent and strong correlation (correlation  $> 0.9939$ ) between the theoretical bound and the observed generalization error across various training sizes. This shows that our bound, based on diversity, is able to capture the behaviour of the generalization error.



**Figure 3.2** Generalization gap, i.e., train error - test error, and the theoretical bound, i.e.,  $(C_5^2 - d_{min}^2)/\sqrt{N}$ , as a function of the number of training samples on MNIST dataset for neural networks with intermediate layer sizes from left to right: 128 (correlation=0.9948), 256 (correlation=0.9939), and 512 (correlation=0.9953). The theoretical term has been scaled in the same range as the generalization gap. All results are averaged over 5 random seeds. [Publication I]

### 3.1.3 Discussion

In this section, we have demonstrated the influence of feature diversity on the generalization of neural networks. We quantified diversity through the average  $L_2$  distance computed between the hidden-layer features. Notably, we have introduced novel generalization bounds that are dependent on diversity for different settings in supervised learning. These derived bounds exhibit an inverse

relationship with the diversity term, elucidating that learning diverse features can reduce the generalization gap and mitigate overfitting. Furthermore, we have extended our analyses to encompass deeper networks and diverse loss functions. These findings address the Research Question 1 showing that it is possible to derive rigorous generalization bounds highlighting the role of feature diversity.

The key limitations of the analysis conducted here are the following: (i) The  $L_2$  distance is not scale-invariant, i.e., it is sensitive to the  $L_2$ -norm of the feature vector  $\Phi$ . This is captured by the dependency on  $(C_5^2 - d_{min}^2)$  in the bound of Theorem 1 and not  $d_{min}$  solely, as the variable  $C_5$  captures the feature norm. Using scale-invariant measures, e.g., correlation or Mutual Information, can be a potentially superior approach for modeling feature diversity and present an intriguing direction for future research. (ii) In the main diversity assumption, Assumption 1,  $d_{min}$  is highly sensitive to the input  $\mathbf{x}$ . Specifically, it is sufficient that there exists one input  $\mathbf{x}$  that yields an extremely small ( $\simeq 0$ ) average  $L_2$  distance between the features, for  $d_{min}$  to be vacuous. In our analysis of the EBM, in Section 3.3.1, we show how to relax the diversity assumption and ensure that  $d_{min} > 0$ .

## 3.2 Feature Diversity in Neural Networks: Algorithms

The dissertation’s contribution presented in this section provides an answer to Research Question 2, which concerns proposing practical feature diversity-promoting strategies to improve the performance of neural networks in multiple contexts. In the following, we will describe the main results that were proposed in Publication II, Publication III, and Publication IV.

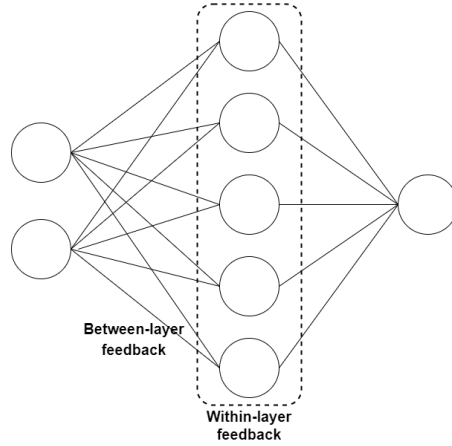
### 3.2.1 WLD-Reg: A Data-Dependent Within-Layer Diversity Regularizer

Here, we present the main approach, namely Within-Layer Diversity Regularizer (WLD-Reg), presented in Publication II. In Section 3.2.1.1, we highlight the

details of the proposed algorithm and discuss its mechanism. In Section 3.2.1.2, we discuss the key empirical performance achieved by WLD-Reg.

### 3.2.1.1 Methodology

We propose new algorithms that are designed explicitly to promote feature diversity and reduce redundancy within the feature layer. In this context, the feature layer corresponds to the final intermediate layer in a neural network. The objective is to ensure that each unit/neuron within this layer captures a distinct pattern contributing complementary information to the overall feature representation.



**Figure 3.3** Illustration of ‘within-layer’ feedback and the ‘between-layer’ feedback

Typically, during the training of a neural network model, units at a particular layer receive feedback from the subsequent layer, as illustrated in Figure 3.3. This can be referred to as ‘**between-layer**’ feedback. We propose an augmentation to this conventional feedback mechanism by introducing an additional ‘**within-layer**’ feedback to encourage diversity. This is achieved using new data-dependent regularizers at the feature layer that explicitly reduces redundancy within this layer. The first step is modeling similarity between two units  $\phi_n(\cdot)$  and  $\phi_m(\cdot)$ . For this end, we use the average radial basis function (RBF) over the available data  $\{\mathbf{x}\}_{j=1}^N$ :

$$s_{nm} := \frac{1}{N} \sum_{j=1}^N \exp(-\gamma \|\phi_n(\mathbf{x}_j) - \phi_m(\mathbf{x}_j)\|^2), \quad (3.7)$$

where  $\gamma$  is a hyper-parameter. The key advantage of the RBF distance is its ability to capture non-linear relationships [107]. The similarity  $s_{nm}$  can be computed either over the entire dataset or on a batch-wise basis. In essence, if two units  $n$  and  $m$  have similar outputs for many samples, their corresponding similarity  $s_{nm}$  will be elevated. Conversely, if their outputs differ on average, their similarity  $s_{nm}$  is small, making them "diverse".

Subsequently, based on the pairwise similarities  $s_{nm}$ , we propose three variations for deriving the overall layer similarity  $J$  encompassing all units' redundancy within the feature layer:

- **Direct:**  $J := \sum_{n \neq m} s_{nm}$ . In this variant, the global layer similarity is directly modeled as the sum of pairwise similarities between the different units. The minimization of this sum encourages each unit to acquire distinct and complimentary information.
- **Det:**  $J := -\det(\mathbf{S})$ , where  $\mathbf{S}$  is a similarity matrix defined as  $\mathbf{S}_{nm} = s_{nm}$ . This variant is motivated by the Determinantal Point Process (DPP) [97, 108]. The determinant of  $\mathbf{S}$  measures the global diversity of the set of features. Geometrically,  $\det(\mathbf{S})$  is the volume of the parallelepiped formed by the vectors within  $\mathbf{S}$  [97]. A larger volume indicates more "diversity". So, maximizing  $\det(\cdot)$  (minimizing  $-\det(\cdot)$ ) enforces diversity in the learned features.
- **Logdet:**  $J := -\log\det(\mathbf{S} + \epsilon \mathbf{I})$  (defined with the additional  $\epsilon \mathbf{I}$  for positive definiteness). Similar to the Det variant, Logdet serves the same motivation. However, Logdet is used instead of Det as it is a convex function over the positive definite matrix space.

We note here that the first proposed variant, denoted as "Direct", shares similarities with DeCov [33], specifically capturing only the pairwise similarity among components. Notably, this variant cannot model higher-order "diversity." In contrast, the remaining two variants, namely "Det" and "Logdet", adopt an approach that considers global similarity, enabling a more holistic measure of redundancy.

The final form of the proposed regularizer WLD-Reg is:

$$\hat{L}_{WLD-Reg} := \lambda_1 J + \lambda_2 \sum_{i=1}^N \|\Phi(\mathbf{x}_i)\|_2^2 \quad (3.8)$$

The primary term of WLD-Reg,  $J$  as defined with the three aforementioned variants, penalizes the similarity between the units, compelling units within the feature layer to learn distinct patterns. whereas the second term ensures scale invariance. For instance, being based on the RBF distance,  $s_{mn}$  is not scale-invariant. Trivially, it can be minimized by scaling all the activations of the feature layer with a high factor. This does not affect the performance of the model, as the model can easily rescale the high activations to normal values by learning small weights in the subsequent layer. The second term in equation 3.8 mitigates this problem by penalizing solutions with high feature norms.  $\lambda_1$  and  $\lambda_2$  are two hyperparameters controlling the contribution of each term to the loss.

The different variants of WLD-Reg introduce new data-dependant regularizers that leverage the concept of feature diversity aiming to enhance the performance of neural networks via redundancy reduction. It can be incorporated in a plug-and-play manner on top of any fully connected layer in a neural network to improve performance. The full details of the algorithm WLD-Reg are presented in Algorithm 1.

### 3.2.1.2 Empirical Results

#### Image Classification

To validate the effectiveness of WLD-Reg regularizers, we experiment with three different datasets, namely CIFAR10 [109], CIFAR100 [109], and ImageNet [110]. CIFAR10/100 are formed of  $32 \times 32$  rgb images. The training set is composed of 50000 samples and the test set is formed of 10,000 test samples. CIFAR10 has a total of 10 classes, while CIFAR100 is composed of 100 distinct classes. In our experiment, we split the original training set (50,000) into two sets: we use the first 40,000 images as the main training samples and the last 10,000 as validation samples for hyperparameters optimization. The ImageNet contains

---

**Algorithm 1** WLD-Reg approach [Publication II]

---

**Model:** Given a neural network  $f(\cdot)$  with a feature representation  $\phi(\cdot)$ , i.e., last intermediate layer.

**Input:** Training Data:  $\{\mathbf{x}_i, y_i\}_{i=1}^N$

**Parameters:**  $\lambda_1$  and  $\lambda_2$  in equation 3.8

- 1: **for** every mini-batch:  $\{\mathbf{x}_i, y_i\}_{i=1}^m \in \{\mathbf{x}_i, y_i\}_{i=1}^N$  **do**
  - 2:   Forward pass the inputs  $\{\mathbf{x}_i\}_{i=1}^m$  into the model to obtain the outputs  $\{f(\mathbf{x}_i)\}_{i=1}^m$  and the feature representations  $\{\Phi(\mathbf{x}_i)\}_{i=1}^m$
  - 3:   Compute the standard loss  $\hat{L}(f)$ .
  - 4:   Compute the extra WLD-Reg loss  $\hat{L}_{WLD-Reg}$  (equation 3.8).
  - 5:   Compute the total loss as the sum of the standard loss and WLD-Reg
  - 6:   Compute the gradient of the total loss and use it to update the weights of  $f$ .
  - 7: **end for**
  - 8: **return** Return  $f$ .
- 

1000 classes, with 1.28 million training samples and 50 thousand validation images.

For the CIFAR10/100 datasets, we experiment with two state-of-the-art CNNs, ResNeXt-29-08-16 [51] and ResNet50 [4]. For the ImageNet dataset, we experiment with four different models: ResNet50 [4], Wide-ResNet50 [50], ResNeXt50 [51], and ResNet101 [4]. The full models descriptions and experimental details are presented in Section 4.1 of Publication II.

In Tables 3.1 and 3.2, we report the error rates on the three datasets of our approach as well as the standard networks, i.e., training without a diversity regularizer and training with DeCov [33]. It is noteworthy that incorporating a diversity strategy (Decov or our approach) consistently enhances the results across all the different models and datasets compared to the standard approach.

In Table 3.1, for example with ResNet50, the three variants of our proposed approach significantly reduce test errors compared to the standard approach on both CIFAR10 and CIFAR100 datasets, showing improvements ranging from 0.51% to 0.63% for CIFAR10 and 1.25% to 1.44% for CIFAR100. Notably, the Direct variant and the Logdet variant yield superior performance for ResNeXt and ResNet models on CIFAR10, while the Logdet variant performs best for both models on CIFAR100. For instance, with ResNeXt on CIFAR10, the

**Table 3.1** Classification errors of the different methods on CIFAR10 and CIFAR100. Results are averaged over three random seeds [Publication II].

method	CIFAR10	CIFAR100
ResNeXt-29-08-16		
Standard	$6.93 \pm 0.10$	$26.73 \pm 0.10$
Decov [33]	$6.82 \pm 0.15$	$26.70 \pm 0.10$
WLD-Reg (Direct)	<b><math>6.28 \pm 0.11</math></b>	$26.20 \pm 0.18$
WLD-Reg (Det)	$6.51 \pm 0.16$	$26.35 \pm 0.23$
WLD-Reg (Logdet)	$6.38 \pm 0.08$	<b><math>25.88 \pm 0.21</math></b>
ResNet50		
Standard	$8.28 \pm 0.41$	$33.39 \pm 0.42$
Decov [33]	$8.03 \pm 0.11$	$32.26 \pm 0.22$
WLD-Reg (Direct)	$7.77 \pm 0.09$	$32.09 \pm 0.11$
WLD-Reg (Det)	$7.75 \pm 0.12$	$32.14 \pm 0.28$
WLD-Reg (Logdet)	<b><math>7.65 \pm 0.10</math></b>	<b><math>31.99 \pm 0.05</math></b>

**Table 3.2** Classification errors of different models with different diversity strategies on ImageNet dataset [Publication II].

	ResNet50	Wide-ResNet50	ResNeXt50	ResNet101
Standard	23.84	22.42	22.70	22.33
DeCov [33]	23.62	22.68	22.57	22.31
WLD-Reg (Direct)	<b>23.24</b>	21.95	<b>22.25</b>	22.14
WLD-Reg (Det)	23.34	<b>21.75</b>	22.44	<b>21.87</b>
WLD-Reg (Logdet)	23.32	21.96	22.40	22.04

Direct variant yields a 0.65% improvement over the standard approach and a 0.54% improvement over DeCov. Overall, our three proposed variants consistently outperform both DeCov and the standard approach across all testing configurations.

On the ImageNet dataset, as shown in Table 3.2, feature diversity (our approach and DeCov) reduces test errors and outperforms the standard approach. Our three variants of WLD-Reg consistently outperform DeCov, with the Di-



rect variant showing optimal performance for ResNet50 and ResNeXt50 and the Det variant yielding the lowest error rates for Wide-ResNet50 and ResNet101.

## Feature Diversity Reduces Overfitting

The aforementioned results on the three datasets show that using feature diversity-based approaches consistently boosts the performance of CNNs. This corroborates the theoretical findings of Section 3.1, which show that feature diversity can improve generalization and mitigate overfitting. To further highlight this effect empirically, we report the generalization gap, i.e., training error - test error, of the different models on ImageNet in Table 3.3.

**Table 3.3** The generalization gap, i.e., training error - test error, of different approaches on ImageNet dataset. \* denotes WLD-Reg variants [Publication II].

	Standard	DeCov	Direct*	Det*	Logdet*
ResNet50	2.87	2.70	<b>1.15</b>	1.23	1.21
Wide-ResNet50	6.33	6.34	4.44	<b>4.34</b>	4.58
ResNeXt50	5.99	5.85	<b>4.41</b>	4.59	4.48
ResNet101	4.64	4.61	3.68	<b>3.38</b>	3.71

As can be seen in Table 3.3, our diversity-promoting strategies reduce the generalization gap. For instance, the Logdet variant reduces the gap by more than 1.5% for all the models, except for ResNet101, where the gain is 0.93%. This further validates the importance of feature diversity on the generalization of neural networks. The convergence of the theoretical underpinnings and practical validations holds pivotal significance in furthering our understanding of generalization and, consequently, for the ongoing advancement of deep learning towards enhanced efficiency and adaptability across diverse datasets and real-world scenarios.

## Feature Diversity Helps in the Presence of Label Noise

To further substantiate the efficacy of employing a feature diversity strategy beyond standard classification, we assess the resilience of our methodology in the context of label noise. In this scenario, conventional neural networks tend

to overfit to noisy samples, thereby diminishing their generalization capacity to the test set [111, 112]. Introducing measures to enforce feature diversity can be useful in obtaining robust and more meaningful representations, mitigating the adverse impact of noise. To empirically demonstrate this idea, we report results with supplementary experiments on the CIFAR10 and CIFAR100 datasets with label noise, specifically at rates of 20% and 40%. The results are presented in Table 3.4.

**Table 3.4** Classification errors of ResNet50 using different diversity strategies on CIFAR10 and CIFAR100 datasets with different label noise ratios. Results are averaged over three random seeds [Publication II].

Method	20% label noise	
	CIFAR10	CIFAR100
Standard	$14.38 \pm 0.29$	$45.11 \pm 0.52$
DeCov [33]	$13.75 \pm 0.19$	$41.93 \pm 0.40$
WLD-Reg (Direct)	$13.31 \pm 0.40$	$40.10 \pm 0.31$
WLD-Reg (Det)	$13.21 \pm 0.21$	$40.35 \pm 0.31$
WLD-Reg (Logdet)	<b><math>13.01 \pm 0.40</math></b>	<b><math>39.97 \pm 0.19</math></b>
Method	40% label noise	
	CIFAR10	CIFAR100
Standard	$19.40 \pm 0.80$	$48.81 \pm 0.57$
DeCov [33]	$17.60 \pm 0.66$	$48.23 \pm 0.48$
WLD-Reg (Direct)	<b><math>16.96 \pm 0.32</math></b>	$46.73 \pm 0.23$
WLD-Reg (Det)	$17.49 \pm 0.04$	$46.93 \pm 0.62$
WLD-Reg (Logdet)	$17.24 \pm 0.31$	<b><math>46.52 \pm 0.22</math></b>

By comparing the results in Table 3.4 with results in Table 3.1, we note that the performance gap between the standard approach and diversity-promoting approaches (Decov and our WLD-Reg variants) becomes more noticeable. Notably, the Logdet variant of WLD-Reg demonstrates a substantial improvement, yielding improvements of  $\sim 2\%$  on both datasets with 40% noise.

## WLD-Reg Helps in Transfer Learning

In addition to the results presented in Publication II, in this dissertation, we also explore other learning contexts where feature diversity can be beneficial. For instance, transfer learning relies on the pre-training step to learn a set of transferable features. Learning diverse features in the pre-training phase enables the neural network to capture a broad spectrum of patterns and features from the pre-training data, yielding a more robust and rich representation. This richness in feature learning enhances the model’s ability to generalize effectively to new tasks. To evaluate the effect of feature diversity in the transfer learning context, we conduct an additional experiment.

We pre-train different models on ImageNet with different diversity strategies and then fine-tune these models to CIFAR10 and CIFAR100, with the different diversity strategies. The models are fine-tuned for 20 epochs using Adam optimizer [2] with a learning rate equal to 0.0001 and standard data augmentation is applied. The original images of CIFAR are preprocessed and resized to (96, 96, 3) to be adequate for ResNet50 trained on ImageNet. The results are reported in Table 3.5. As can be seen, employing a diversity strategy helps in the transfer learning context and leads to consistently lower error rates. For example, the Logdet variant of our approach leads to 0.94% and 1.27% gains on CIFAR10 and CIFAR100, respectively.

**Table 3.5** Transfer learning performance on CIFAR10 and CIFAR100 of ResNet50 models pre-trained on ImageNet with the different diversity approaches.

	$\hookrightarrow$ CIFAR10	$\hookrightarrow$ CIFAR100
Standard	6.14	22.99
DeCov	5.92	21.91
WLD-Reg (Direct)	5.89	<b>21.48</b>
WLD-Reg (Det)	5.51	22.01
WLD-Reg (Logdet)	<b>5.20</b>	21.72

### 3.2.1.3 Discussion

We have presented a diversity-inducing methodology, namely WLD-Reg and showed through extensive empirical evaluation that WLD-Reg consistently boost the performance of CNNs in multiple learning scenarios. Beyond CNN models, in Table 4 in Publication II, we show that WLD-Reg can be used to improve the generalization of modern attention-free multi-layer perceptron (MLP)-based models for image classification [113, 114, 115], which are known to exhibit high overfitting and require regularization. The findings presented in this section contribute insights to address Research Question 2, showing that methodologies based on feature diversity can improve performance and mitigate overfitting in deep learning models.

We note that WLD-Reg requires additional computations compared to the standard approach, i.e., computing  $\hat{L}_{WLD-Reg}$  in equation 3.8. However, in practice, this corresponds to a small additional time cost during the training. For instance, the Direct, Det, and Logdet variants cost only 0.29%, 0.39%, and 0.49% extra training time for ResNet50 on the ImageNet dataset.

Weight-based diversity strategies, as shown in Table 1 in Publication II, can degrade the performance of neural networks in some cases. On the contrary, as shown through the results in Section 3.2.1.2, WLD-Reg yields consistently an improved performance in multiple tasks and datasets. Furthermore, it is shown in [116], that WLD-Reg can help CNNs in the context of one-class classification [117, 118, 119, 120]. Another advantage compared to the weight-based diversity approaches is that they need to be applied on top of all layers which can result in high computational costs for deep models, whereas WLD-Reg applied only on top of the last hidden layer is able to achieve consistent performance boost. The main limitation of our proposed approach is the fact that it is compatible only with flat feature representations. Future work includes extending it for different topologies, e.g., convolutional output maps and recurrent representations.

## 3.2.2 Diversity in BoF Pooling

In this section, we present the main contribution of Publication III, where a diversity regularizer is proposed to boost the performance of BoF pooling-based CNNs.

### 3.2.2.1 Methodology

BoF pooling is an advanced aggregation technique that has been proposed as an alternative to standard global average and max pooling techniques in CNN in order to construct powerful models with a low computational cost. The key element of BoF is the dictionary used to compile the histogram representation. In this dissertation, we argue that diversity within this dictionary can help to achieve good performance. For instance, a diverse dictionary ensures that the quantized representations, i.e., the histograms, encompass various features and textures present in the images. This leads to a richer and more comprehensive description of the input and, consequently, enhances the generalization capability of BoF-based CNN models. In this dissertation, we test this idea by developing a simple regularizer that aims to reduce the redundancy with the dictionary items.

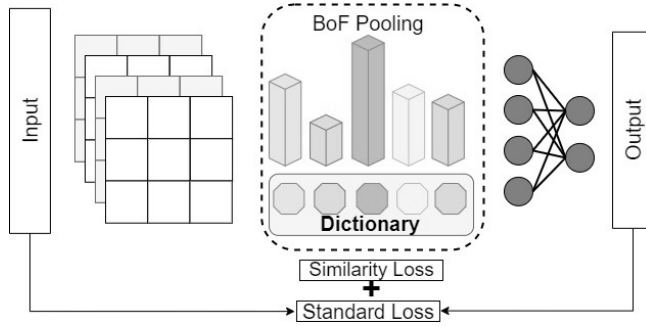
Given a CNN model containing a BoF pooling layer with a dictionary  $\{\mathbf{c}_1, \dots, \mathbf{c}_K\}$  of size  $K$ , the similarity between two elements  $\mathbf{c}_i$  and  $\mathbf{c}_j$  of this dictionary can be measured with the squared correlation:

$$SIM(\mathbf{c}_i, \mathbf{c}_j) = \left( corr(\mathbf{c}_i, \mathbf{c}_j) \right)^2. \quad (3.9)$$

Next, the total redundancy loss can be obtained as the sum of the pairwise similarities as follows:

$$\boxed{Similarity\ Loss = \sum_{i \neq j} \left( corr(\mathbf{c}_i, \mathbf{c}_j) \right)^2}. \quad (3.10)$$

As illustrated in Figure 3.4, the proposed regularizer can be integrated in any BoF-based CNN training to augment the original loss. A hyper-parameter  $\beta$  can be used to control its contribution to the total loss. As the proposed regularizer



**Figure 3.4** An illustration on how the BoF-based CNN model loss is computed using our approach. The standard loss can be least squares or cross entropy and the similarity loss corresponds to equation 3.10 [Publication III].

depends only on the dictionary items, only their corresponding gradients change during optimization. For instance, each item  $\mathbf{c}_i$  receives an extra feedback equal to

$$\beta \frac{\partial \sum_{i \neq j} SIM(\mathbf{c}_i, \mathbf{c}_j)}{\partial \mathbf{c}_i}.$$

### 3.2.2.2 Empirical Results

We report competitive results of the different pooling strategies, namely global max pooling (GMP) [2], global average pooling (GAP) [2], BoF [58, 59], and BoF augmented with our proposed regularizer. The detailed experimental setup is available in Publication III. In Table 3.6 and Table 3.7, we present the average error rates and standard deviations corresponding to different filter sizes on the MNIST and fashionMNIST datasets, respectively.

A noteworthy observation is that both variations of BoF consistently outperform standard pooling approaches, namely GMP and GAP. For instance, with 16 filters, GMP and GAP exhibit error rates of 3.63% and 4.67% on MNIST, respectively, while standard BoF and our BoF variant achieve lower error rates of 1.03% and 1.00%, respectively, for the same configuration. On the fashionMNIST dataset, our variant of BoF, which penalizes the correlations between dictionary items, consistently boosts the performance in all the settings. For instance, with a 128-filter model on fashionMNIST, our approach achieves an error rate of only 8.77% compared to the 9.02% error rate attained by standard BoF.

**Table 3.6** Average error rates and standard deviation of different approaches for different number of filters in the last convolutional layer on the MNIST dataset. Results are averaged over 5 random seeds. The top results for each approach are in bold and the best global result is underlined [Publication III].

method	16 filters	32 filters	64 filters	128 filters
GMP	$3.63 \pm 0.31$	$1.97 \pm 0.20$	$1.39 \pm 0.07$	<b><math>1.09 \pm 0.08</math></b>
GAP	$4.67 \pm 1.17$	$2.01 \pm 0.09$	$1.31 \pm 0.05$	<b><math>1.06 \pm 0.02</math></b>
BoF	$1.03 \pm 0.08$	<b><math>0.97 \pm 0.11</math></b>	$1.00 \pm 0.08$	$1.03 \pm 0.06$
BoF (ours)	$1.00 \pm 0.06$	$0.98 \pm 0.06$	<b><u><math>0.87 \pm 0.10</math></u></b>	$0.98 \pm 0.08$

**Table 3.7** Average error rates and standard deviation of different approaches for different numbers of filters in the last convolutional layer on the fashionMNIST dataset. Results are averaged over 5 random seeds. The top results for each approach are in bold and the best global result is underlined [Publication III].

method	16 filters	32 filters	64 filters	128 filters
GMP	$14.94 \pm 0.70$	$12.13 \pm 0.30$	$10.46 \pm 0.18$	<b><math>9.48 \pm 0.12</math></b>
GAP	$15.09 \pm 0.19$	$12.30 \pm 0.20$	$10.91 \pm 0.21$	<b><math>9.97 \pm 0.06</math></b>
BoF	$9.55 \pm 0.29$	$9.44 \pm 0.25$	$9.04 \pm 0.22$	<b><math>9.02 \pm 0.15</math></b>
BoF (ours)	$9.52 \pm 0.29$	$9.14 \pm 0.12$	$8.98 \pm 0.18$	<b><u><math>8.77 \pm 0.22</math></u></b>

To further show the effectiveness of our approach, we conduct an additional experiment with the CIFAR10 dataset. We evaluate the performance with three filter sizes, 32, 64, and 128 in the final convolutional layer. The results are presented in Table 3.8. Notably, our approach yields the best result, achieving a minimal error rate of 15.93% with 128 filters. This represents a noteworthy enhancement, surpassing the best results obtained by GMP, GAP, and the standard BoF by 2.87%, 1.68%, and 0.17%, respectively.

### 3.2.2.3 Discussion

In this section, we showed that BoF pooling can benefit from feature diversity. We developed a simple, yet effective, approach that reduces the redundancy along the features (dictionary items) and showed that it yields performance improvement. This provides insights into Research Question 2 showing that

**Table 3.8** Average error rates and standard deviation of different approaches for different number of filters in the last convolutional layer on the CIFAR10 dataset. Results are averaged over three random seeds. Top results for each approach are in bold and best global result is underlined [Publication III].

method	32 filters	64 filters	128 filters
GMP	$22.31 \pm 0.48$	$20.33 \pm 0.67$	<b><math>18.80 \pm 1.03</math></b>
GAP	$20.94 \pm 0.53$	$20.08 \pm 0.99$	<b><math>17.61 \pm 0.33</math></b>
BoF	$17.15 \pm 0.58$	$17.05 \pm 0.11$	<b><math>16.10 \pm 0.20</math></b>
BoF (ours)	$17.21 \pm 0.71$	$16.57 \pm 0.25$	<u><b><math>15.93 \pm 0.11</math></b></u>

feature diversity can be useful for neural networks beyond the standard CNN models.

The main limitation of this work is the limited empirical validation. However, we are confident these results will spark future research to develop advanced diversity-based methodologies and extensively validate their ability to boost the performance of BoF-based CNNs with large datasets.

### 3.2.3 Feature Diversity in Autoencoders

In addition to the results presented in Publication II and Publication III, in this dissertation, we also present the results in Publication IV, exploring feature diversity in the context of autoencoders.

#### 3.2.3.1 Methodology

We present an approach which leverages the concept of feature diversity to improve the performance of autoencoders. The key component of the autoencoders is the bottleneck layer, which typically has a low-dimensionality. By optimizing the autoencoder to learn to reconstruct the input, the model is forced to avoid redundancies and noise in the bottleneck.

In the context of autoencoders, the bottleneck layer, i.e., the output of the encoder is considered the feature layer. In this dissertation, we propose a new diversity-based regularizer that can be added on top of the bottleneck layer to explicitly minimize redundancy among features. Specifically, we propose to



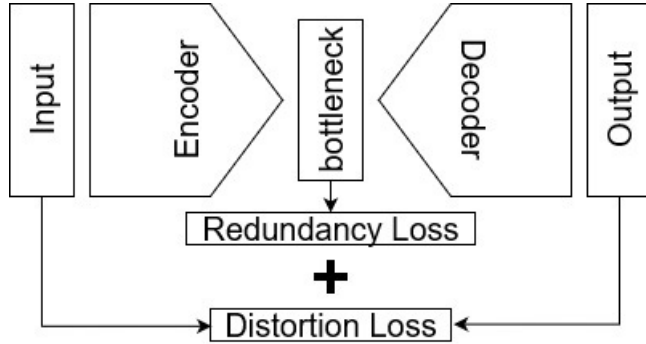
penalize high pairwise covariance between different features. Formally, given a training data set  $\{\mathbf{x}_i\}_{i=1}^N$  and an encoder  $\phi(\cdot) \in \mathbb{R}^D$ , the covariance between the  $i^{th}$  and  $j^{th}$  features,  $\phi_i$  and  $\phi_j$ , can be expressed as follows:

$$C(g_i, g_j) = \frac{1}{N} \sum_n \left( \phi_i(\mathbf{x}_n) - \mu_i \right) \left( \phi_j(\mathbf{x}_n) - \mu_j \right), \quad (3.11)$$

where  $\mu_i = \frac{1}{N} \sum_n \phi_i(\mathbf{x}_n)$  is the average output of the  $i^{th}$  unit. The objective is to minimize redundancy in the bottleneck representation, which corresponds to minimizing the pairwise covariance between distinct features. So, the final form of the proposed regularizer is as follows:

$$\boxed{Redundancy\ Loss = \sum_{i \neq j} \left( \frac{1}{N} \sum_n (g_i(\mathbf{x}_n) - \mu_i)(g_j(\mathbf{x}_n) - \mu_j) \right)} \quad (3.12)$$

As illustrated in Figure 3.5, the proposed regularizer can be seamlessly integrated into any autoencoder-based model to compliment the distortion loss and optimized in a batch-wise manner, making it adaptable to various learning strategies and network topologies. Furthermore, a hyper-parameter  $\alpha$  can be used to control its contribution to the final loss.



**Figure 3.5** An illustration of how the autoencoder loss is computed using our approach.

### 3.2.3.2 Empirical Results

#### Dimensionality Reduction

We assess the efficacy of the proposed methodology in the task of dimensionality reduction with autoencoders. We use three different datasets Madelon [121], ISOLET [122], and P53 Mutants [123]. The different characteristic of the datasets are reported in Table 3.9.

**Table 3.9** Statistics of the three datasets used in the dimensionality reduction experiments. # Dim: dimensionality of the data. # Train: number of training samples. # Test: number of test samples. d: projection dimension [Publication IV].

Dataset	# Dim	# Train	# Test	d
Madelon [121]	500	2000	1800	10
ISOLET [122]	617	6238	1559	10
P53 Mutants [123]	5408	21811	9348	50

In order to evaluate the quality of the bottleneck features obtained by the different autoencoders, we apply  $K$ -Nearest Neighbor ( $K$ -NN) classifier on top of them and report the classification accuracy. The results for  $K = 3$  and  $K = 5$  are presented in Table 3.10.

We note that the proposed approach exhibits improved performance for the different values of  $\alpha$ . For instance, with  $\alpha = 0.005$ , the proposed regularizer yields improvement of 4% for ( $K = 3$ )-NN and 3% for ( $K = 5$ )-NN for Madelon dataset. Similar trends are observed for the ISOLET and P53 Mutants datasets. Particularly, for ISOLET, the proposed approach achieves the highest accuracy of 78.96% for ( $K = 3$ )-NN and 79.83% for ( $K = 5$ )-NN with  $\alpha = 0.01$ . Overall, the results on the three datasets underscore the effectiveness of the proposed approach in improving the quality of the features of the bottleneck to achieve strong performance.

To further investigate the impact of dimensionality of the bottleneck on the performance of our approach, we plot the ( $K = 3$ )-NN accuracy as a function of the dimension of the bottleneck size  $d$  on ISOLET dataset in Figure 3.6. The results indicate that learning diverse features at the bottleneck side consistently improves the results. Notably, the improvement is higher for smaller values of  $d$ , as for smaller dimensions it is more crucial to learn rich and diverse features

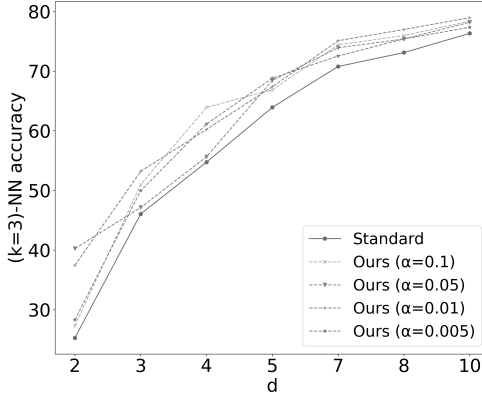
**Table 3.10** Classification accuracy of Nearest Neighbor classifier applied on the bottleneck representations (average and standard deviation over 10 repetitions) [Publication IV].

	Madelon	
	( $K = 3$ )-NN	( $K = 5$ )-NN
Standard	69.33% $\pm$ 2.71	71.32% $\pm$ 2.82
Ours ( $\alpha = 0.1$ )	72.51% $\pm$ 1.73	74.08% $\pm$ 1.63
Ours ( $\alpha = 0.05$ )	73.52% $\pm$ 1.91	74.53% $\pm$ 1.49
Ours ( $\alpha = 0.01$ )	72.65% $\pm$ 2.21	74.50% $\pm$ 1.87
Ours ( $\alpha = 0.005$ )	<b>73.82% <math>\pm</math> 1.83</b>	<b>74.78% <math>\pm</math> 1.78</b>
	ISOLET	
	( $K = 3$ )-NN	( $K = 5$ )-NN
Standard	76.32% $\pm$ 1.85	77.70% $\pm$ 1.60
Ours ( $\alpha = 0.1$ )	78.35% $\pm$ 0.46	79.82% $\pm$ 0.47
Ours ( $\alpha = 0.05$ )	78.18% $\pm$ 0.40	79.43% $\pm$ 0.44
Ours ( $\alpha = 0.01$ )	<b>78.96% <math>\pm</math> 0.56</b>	<b>79.83% <math>\pm</math> 0.54</b>
Ours ( $\alpha = 0.005$ )	77.34% $\pm$ 0.66	79.29% $\pm$ 0.66
	P53 Mutants	
	( $K = 3$ )-NN	( $K = 5$ )-NN
Standard	56.42 % $\pm$ 0.60	54.99% $\pm$ 0.48
Ours ( $\alpha = 0.1$ )	<b>57.88% <math>\pm</math> 0.46</b>	<b>56.18% <math>\pm</math> 0.59</b>
Ours ( $\alpha = 0.05$ )	56.17% $\pm$ 0.46	55.39% $\pm$ 1.09
Ours ( $\alpha = 0.01$ )	57.22% $\pm$ 0.50	55.65 % $\pm$ 0.46
Ours ( $\alpha = 0.005$ )	56.83% $\pm$ 0.41	55.92% $\pm$ 0.46

to be able to address the task at hand.

## Image Compression

Next, we assess the efficacy of the proposed methodology in the task of image compression with autoencoders on CIFAR10 dataset. The input original images are flattened to generate inputs with size  $32 \times 32 \times 3 = 3072$ . The autoencoder architecture used is presented in Table 3.11. We experiment with different



**Figure 3.6** Average ( $K = 3$ )-NN accuracy as a function of the dimension of the bottleneck size  $d$  on Isolet dataset. Results are averaged over 10 random seeds [Publication IV].

sizes of bottleneck, i.e.,  $d \in \{128, 256\}$ . The training is conducted with Adam optimizer for 50 epochs with a  $1e^{-2}$  learning rate and a batch size of 128.

**Table 3.11** Autoencoder topology used for CIFAR10. \* denotes the bottleneck representation [Publication IV].

Layer	Output shape
Input	[3072]
Linear	[512]
ReLU activation	[512]
Linear	[256]
ReLU activation	[256]
Linear	[d]
ReLU activation*	[d]
Linear	[256]
ReLU activation	[256]
Linear	[512]
ReLU activation	[512]
Linear	[3072]

We report different performance metrics, namely root-mean-square error (RMSE), peak signal-to-noise ratio (PSNR), and structural similarity index

(SSIM) over 10 random seeds. Given two images  $I_1$  and  $I_2$ , the three metrics can be computed as follows:

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n [I_1(i, j) - I_2(i, j)]^2}, \quad (3.13)$$

$$\text{PSNR} = 10 \cdot \log_{10} \left( \frac{\text{Max}_I^2}{\text{MSE}} \right), \quad (3.14)$$

$$\text{SSIM}(I_1, I_2) = \frac{(2\mu_{I_1}\mu_{I_2})(2\sigma_{I_1I_2})}{(\mu_{I_1}^2 + \mu_{I_2}^2)(\sigma_{I_1}^2 + \sigma_{I_2}^2)}, \quad (3.15)$$

where  $\text{Max}_I$  is the maximum value of the images,  $(\mu_{I_1}, \mu_{I_2})$  are the mean of the images,  $(\sigma_{I_1}, \sigma_{I_2})$  are the variances, and  $\sigma_{I_1I_2}$  is the covariance.

The results for different bottleneck sizes are reported in Table 3.12. Notably, our proposed approach, particularly when employing a regularization parameter of 0.0005, consistently surpasses the standard compression methodology across all compression ratios. For  $d = 256$ , the best performance is achieved by using  $\alpha = 0.0005$ , whereas for  $d = 128$ ,  $\alpha = 0.001$  leads to the best result. The results underscore the efficacy of our proposed approach in preserving image quality during compression, emphasizing the advantages of applying diversity-inducing approaches to improve the performance of autoencoders.

## Image Denoising

Next, we assess the efficacy of the proposed methodology in the task of image denoising with autoencoders on the CIFAR10 dataset. We construct a noisy variant of the dataset by adding a random noise from the normal distribution  $\beta \times \mathcal{N}(0, 1)$ , where  $\beta$  is the weight of the added noise. For the model topology, we use the same topology as in the compression task, i.e., Table 3.11. The results for different levels of noise  $\beta = 0.1$  and  $\beta = 0.2$  are presented in Table 3.13.

As shown in Table 3.13, for the noise level  $\beta = 0.1$ , our approach consistently outperforms the standard autoencoders. Specifically, by using our regularization technique with  $\alpha = 0.005$  and  $\alpha = 0.001$ , the autoencoder achieves lower RMSE (0.0940 and 0.0952, respectively), higher PSNR (0.6227 and 0.6129, respectively), and enhanced SSIM (21.0411 and 20.9325, respectively) compared

**Table 3.12** RMSE, PSNR, and SSIM on the image compression task with CIFAR10 dataset (average and standard deviation over 5 repetitions) [Publication IV].

	RMSE ↓	PSNR ↑	SSIM ↑
	3072 → 256		
Standard	0.0888 ± 0.0022	0.6601 ± 0.0068	21.5547 ± 0.2470
Ours (0.01)	0.0882 ± 0.0012	0.6637 ± 0.0064	21.6168 ± 0.1314
Ours (0.005)	0.0882 ± 0.0006	0.6628 ± 0.0060	21.6271 ± 0.0593
Ours (0.001)	0.0882 ± 0.0011	0.6613 ± 0.0068	21.6347 ± 0.1154
Ours (0.0005)	<b>0.0877 ± 0.0011</b>	<b>0.6642 ± 0.0078</b>	<b>21.6829 ± 0.1156</b>
Ours (0.0001)	0.0885 ± 0.0014	0.6610 ± 0.0060	21.5927 ± 0.1518
	3072 → 128		
Standard	0.0929 ± 0.0015	0.6151 ± 0.0100	21.2830 ± 0.1455
Ours (0.01)	0.0920 ± 0.0010	0.6210 ± 0.0078	21.3765 ± 0.0920
Ours (0.005)	0.0927 ± 0.0014	0.6144 ± 0.0089	21.3093 ± 0.1280
Ours (0.001)	<b>0.0917 ± 0.0009</b>	<b>0.6246 ± 0.0054</b>	<b>21.4150 ± 0.0888</b>
Ours (0.0005)	0.0923 ± 0.0016	0.6190 ± 0.0130	21.3436 ± 0.1527
Ours (0.0001)	0.0926 ± 0.0019	0.6182 ± 0.0072	21.3184 ± 0.2070

to the standard approach without diversity regularization. With higher noise level, i.e.,  $\beta = 0.2$ , our approach, particularly with a regularization parameter of 0.0001, also exhibits superior performance, showcasing improved scores across all metrics. These outcomes affirm the efficacy of our proposed approach in the denoising task, highlighting the role that feature diversity plays in this context.

### 3.2.3.3 Discussion

In this section, we showed that autoencoders can benefit from feature diversity. We showed through several experiments that reducing redundancy within the bottleneck representation of the autoencoder boosts its performance and its generalization to unseen data. This provides insights into Research Question 2 showing feature diversity can be useful for neural networks beyond the standard supervised learning.

The main limitation of the proposed approach is being based on covariance

**Table 3.13** RMSE, PSNR, and SSIM on the image denoising task with CIFAR10 dataset (average and standard deviation over 5 repetitions)

	RMSE ↓	PSNR ↑	SSIM ↑
	$\beta = 0.1$		
Standard	$0.0954 \pm 0.0019$	$0.6098 \pm 0.0121$	$20.9243 \pm 0.1734$
Ours (0.01)	$0.0948 \pm 0.0016$	$0.6172 \pm 0.0093$	$20.9703 \pm 0.1550$
Ours (0.005)	<b><math>0.0940 \pm 0.0010</math></b>	<b><math>0.6227 \pm 0.0056</math></b>	<b><math>21.0411 \pm 0.1021</math></b>
Ours (0.001)	$0.0952 \pm 0.0018$	$0.6129 \pm 0.0116$	$20.9325 \pm 0.1651$
Ours (0.0005)	$0.0943 \pm 0.0012$	$0.6190 \pm 0.0085$	$21.0238 \pm 0.1097$
Ours (0.0001)	$0.09489 \pm 0.0012$	$0.6157 \pm 0.0072$	$20.9644 \pm 0.1102$
	$\beta = 0.2$		
Standard	$0.1001 \pm 0.0012$	$0.5798 \pm 0.0081$	$20.4497 \pm 0.0972$
Ours (0.01)	$0.0996 \pm 0.0013$	$0.5846 \pm 0.0089$	$20.4900 \pm 0.1155$
Ours (0.005)	$0.1000 \pm 0.0015$	$0.5806 \pm 0.0104$	$20.4597 \pm 0.1118$
Ours (0.001)	$0.0999 \pm 0.0014$	$0.5824 \pm 0.0090$	$20.4626 \pm 0.1118$
Ours (0.0005)	$0.0997 \pm 0.0015$	$0.5814 \pm 0.0111$	$20.4881 \pm 0.1206$
Ours (0.0001)	<b><math>0.0992 \pm 0.0015</math></b>	<b><math>0.5884 \pm 0.0081</math></b>	<b><math>20.5186 \pm 0.1370</math></b>

as a measure of redundancy, which is sensitive to noise and scale. However, we are confident these results will spark future research to develop more advanced diversity-based methodologies that can further boost the performance of autoencoders and other unsupervised learning models.

### 3.3 Feature Diversity in Energy-Based Models

The dissertation’s contribution presented in this section provides an answer to Research Question 3, which concerns extending our findings to the energy-based learning framework. In the following, we will describe the main theoretical results of Publication V. Furthermore, in this dissertation, we propose and experiment with a methodology, inspired by the theoretical findings, that boosts the performance of EBMs.

### 3.3.1 Feature Diversity in Energy-Based Models: Theory

In this part of the dissertation, we present the main findings of Publication V, extending the theoretical findings of Section 3.1 to EBMs.

#### 3.3.1.1 Problem Formulation

There has been an interest in studying the theoretical generalization properties of EBMs [34, 124, 125]. In [124], a generalization bound based on Rademacher complexity is derived. Their main result is presented in Lemma 1.

**Lemma 1.** [124] *For a well-defined energy function  $E(h, \mathbf{x}, \mathbf{y})$  over hypothesis class  $\mathcal{H}$ , input set  $\mathcal{X}$  and output set  $\mathcal{Y}$ , the following holds for all  $h$  in  $\mathcal{H}$  with a probability of at least  $1 - \delta$*

$$\mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathbf{Z}}[E(h, \mathbf{x}, \mathbf{y})] \leq \frac{1}{m} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{S}} E(h, \mathbf{x}, \mathbf{y}) + 2\mathcal{R}_m(\mathcal{E}) + M\sqrt{\frac{\log(2/\delta)}{2m}}, \quad (3.16)$$

where  $m$  is the total training samples,  $\mathcal{E}$  is the energy function class defined as  $\mathcal{E} = \{E(h, \mathbf{x}, \mathbf{y}) | h \in \mathcal{H}\}$ ,  $\mathcal{R}_m(\mathcal{E})$  is its Rademacher complexity, and  $M$  is the upper-bound of  $\mathcal{E}$ .

The key insight of the result in Lemma 1 is that minimizing empirical energy over the training data, i.e.,  $\frac{1}{m} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{S}} E(h, \mathbf{x}, \mathbf{y})$  is not enough to guarantee the minimization of the true energy expectation, i.e.,  $\mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathbf{Z}}[E(h, \mathbf{x}, \mathbf{y})]$ , as the right-hand side of the bound contains other terms. Thus, it is crucial to understand and characterize the gap between these two quantities to boost the generalization power of EBM models to unseen data.

The inner model  $h = G_{\mathbf{W}}(\mathbf{x})$ , typically modeled with a neural network, plays a pivotal role in the performance of the EBM model and its generalization performance. Similar to Section 3.1, the inner model can be interpreted as a two-stage process

$$G_{\mathbf{W}}(\mathbf{x}) = \sum_i^M w_i \phi_i(\mathbf{x}), \quad (3.17)$$

where  $\{\phi_1(\cdot), \dots, \phi_M(\cdot)\}$  is the feature set. From this stand point,  $G_{\mathbf{W}}$  relies on the features.



In the scope of this dissertation, our hypothesis postulates the significance of acquiring a diverse set of features to ensure generalization. If each unit within  $\Phi(\mathbf{x})$  adeptly captures distinct and unique pattern aspects of the input data, the collective contribution of such units will yield a rich and robust representation of the data. Consequently, this feature diversity can boost the performance of the EBM model and supply it with superior generalization capabilities.

As mentioned in Section 3.1.3, one of the main limitations of using  $d_{min}$ , i.e., the lower bound of the average  $L_2$  distance to model diversity, is the sensitivity to data noise. For instance, if a ReLU-based neural network is used, several feature functions may yield a zero value for different inputs. From this perspective, defining diversity directly as a lower bound for the pair-wise diversity is unrealistic. To address this, in this part of the dissertation, we propose to relax the diversity assumption using a probabilistic lower-bound. This yields new diversity measure defined as follows:

**Definition 1** ( $(\vartheta - \tau)$ -diversity). *A set of feature functions,  $\{\phi_1(\cdot), \dots, \phi_D(\cdot)\}$  is called  $(\vartheta - \tau)$ -diverse, if there exists a constant  $\vartheta \in \mathbb{R}$ , such that for every input  $\mathbf{x}$  we have*

$$\frac{1}{2} \sum_{i \neq j}^D (\phi_i(\mathbf{x}) - \phi_j(\mathbf{x}))^2 \geq \vartheta^2 \quad (3.18)$$

*with a high probability  $\tau$ .*

Similar to the intuition in Section 3.1, if a pair of two feature maps,  $\phi_i(\cdot)$  and  $\phi_j(\cdot)$ , are diverse, they tend to produce distinct values for the same input with a high likelihood. As a result, their  $L_2$  distance is high and, cumulatively, the  $\vartheta$  of the whole feature set is large. From this perspective,  $\vartheta$  quantifies the diversity of the features' set.

### 3.3.1.2 Feature Diversity Improves the Generalization of EBMs

Here, we aim to study how feature diversity affects the generalization of EBMs and highlight its role. We consider different cases and derive rigorous generalization bounds, based on Lemma 1, on EBMs depending on  $(\vartheta - \tau)$ -diversity. The detailed proofs are available in Publication V.

## Regression with $E(h, \mathbf{x}, \mathbf{y}) = \frac{1}{2} \|G_{\mathbf{W}}(\mathbf{x}) - \mathbf{y}\|_2^2$

We first consider the task of regression. This problem can be solved with an EBM, as illustrated in Figure 2.1. In this case, several energy functions can be used [36, 37]. Here, we consider the case of  $E(h, \mathbf{x}, \mathbf{y}) = \frac{1}{2} \|G_{\mathbf{W}}(\mathbf{x}) - \mathbf{y}\|_2^2$ . The hypothesis class  $\mathcal{H} = \{h(\mathbf{x}) = G_{\mathbf{W}}(\mathbf{x}) = \sum_{i=1}^D w_i \phi_i(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) \mid \Phi \in \mathcal{F}, \forall \mathbf{x} : \|\Phi(\mathbf{x})\|_2 \leq A\}$ , and the output set  $\mathcal{Y} \subset \mathbb{R}$ . By bounding all the quantities in Lemma 1 using  $\vartheta$ , we obtain the first feature diversity-dependent bound for the EBM. The main result is presented in Theorem 2.

### Theorem 2: Generalization bound, regression

For the energy function  $E(h, \mathbf{x}, \mathbf{y}) = \frac{1}{2} \|G_{\mathbf{W}}(\mathbf{x}) - \mathbf{y}\|_2^2$ , if the feature set  $\{\phi_1(\cdot), \dots, \phi_D(\cdot)\}$  is  $(\vartheta - \tau)$ -diverse with a probability  $\tau$ , with a probability of at least  $(1 - \delta)\tau$ , the following holds for all  $h$  in  $\mathcal{H}$ :

$$\begin{aligned} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{Z}}[E(h, \mathbf{x}, \mathbf{y})] &\leq \frac{1}{m} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{S}} E(h, \mathbf{x}, \mathbf{y}) \\ &\quad + 4D \|\mathbf{w}\|_{\infty} (\|\mathbf{w}\|_{\infty} \sqrt{DA^2 - \vartheta^2} + B) \mathcal{R}_m(\mathcal{F}) \\ &\quad + \frac{1}{2} (\|\mathbf{w}\|_{\infty} \sqrt{DA^2 - \vartheta^2} + B)^2 \sqrt{\frac{\log(2/\delta)}{2m}}, \quad (3.19) \end{aligned}$$

where  $B$  is the upper-bound of  $\mathcal{Y}$ , i.e.,  $y \leq B, \forall y \in \mathcal{Y}$ .

The main insight of the bound in Theorem 2 is the dependency on feature diversity. Specifically, it exhibits an inverse proportionality to  $\vartheta$  and scales as  $\sqrt{DA^2 - \vartheta^2}$ . This suggests that reducing redundancy, characterized by an increase in  $\vartheta$ , results in a reduction of the gap between true and empirical energies, consequently boosting the performance of EBMs.

## Classification with $E(h, \mathbf{x}, \mathbf{y}) = -\mathbf{y} G_{\mathbf{W}}(\mathbf{x})$

Here, we consider the binary classification problem, as illustrated in Figure 2.1 (b). We use the same assumption as in regression for the inner model, i.e.,  $h(\mathbf{x}) = G_{\mathbf{W}}(\mathbf{x}) = \sum_{i=1}^D w_i \phi_i(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x})$ . In this setting, we consider the energy function  $E(h, \mathbf{x}, \mathbf{y}) = -\mathbf{y} G_{\mathbf{W}}(\mathbf{x})$  [34], over the input set  $\mathcal{X} \in \mathbb{R}^N$ ,

hypothesis class  $\mathcal{H} = \{h(\mathbf{x}) = G_{\mathbf{W}}(\mathbf{x}) = \sum_{i=1}^D w_i \phi_i(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) \mid \Phi \in \mathcal{F}, \forall \mathbf{x} : \|\Phi(\mathbf{x})\|_2 \leq A\}$ , and output set  $\mathcal{Y} \subset \mathbb{R}$ . The main result for this case is presented in Theorem 3.

Theorem 3 presents a rigorous bound for the generalization of EBM in the classification task. The bound highlights the effect of diversity on generalization. It shows that increasing diversity, i.e., increasing  $\vartheta$ , can yield better generalization. Compared to the regression case, i.e., Theorem 2, we note that for the classification, the diversity term, i.e.,  $\vartheta$ , appears only the last term of the bound, whereas for the regression task, increasing diversity leads to reducing last two terms, as they are both dependent on  $\vartheta$ .

### Theorem 3: Generalization bound, classification

For the energy function  $E(h, \mathbf{x}, \mathbf{y}) = -yG_{\mathbf{W}}(\mathbf{x})$ , if the feature set  $\{\phi_1(\cdot), \dots, \phi_D(\cdot)\}$  is  $(\vartheta - \tau)$ -diverse with a probability  $\tau$ , then with a probability of at least  $(1 - \delta)\tau$ , the following holds for all  $h$  in  $\mathcal{H}$ :

$$\begin{aligned} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{Z}}[E(h, \mathbf{x}, \mathbf{y})] &\leq \frac{1}{m} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{S}} E(h, \mathbf{x}, \mathbf{y}) \\ &\quad + 4D\|\mathbf{w}\|_{\infty} \mathcal{R}_m(\mathcal{F}) \quad + \|\mathbf{w}\|_{\infty} \sqrt{DA^2 - \vartheta^2} \sqrt{\frac{\log(2/\delta)}{2m}}. \end{aligned} \quad (3.20)$$

**Implicit regression with  $E(h, \mathbf{x}, \mathbf{y}) = \frac{1}{2} \|G_{\mathbf{W}}^{(1)}(\mathbf{x}) - G_{\mathbf{W}}^{(2)}(\mathbf{y})\|_2^2$**

Next, we consider the case of the implicit regression with an EBM (Figure 2.1 (c)). This constitutes a generalized formulation applicable to various problem domains, including metric learning, image denoising, object detection, as illustrated in [34], or semi-supervised learning [82, 126]. As illustrated in Figure 2.1 (c), this form of EBM features encompass two inner models,  $G_{\mathbf{W}}^1(\cdot)$  and  $G_{\mathbf{W}}^2(\cdot)$ , which can be either identical or distinct based on the specific problem being addressed. In this discussion, we consider the general scenario where the two models correspond to different combinations of diverse features, i.e.,  $G_{\mathbf{W}}^{(1)}(\mathbf{x}) = \sum_{i=1}^{D^{(1)}} w_i^{(1)} \phi_i^{(1)}(\mathbf{x})$  and  $G_{\mathbf{W}}^{(2)}(\mathbf{y}) = \sum_{i=1}^{D^{(2)}} w_i^{(2)} \phi_i^{(2)}(\mathbf{y})$ . Consequently, two distinct  $(\vartheta - \tau)$ -diversity terms are attributed to each set.

To sum up, we consider the energy function  $E(h, \mathbf{x}, \mathbf{y}) = \frac{1}{2} \|G_{\mathbf{W}}^{(1)}(\mathbf{x}) - G_{\mathbf{W}}^{(2)}(\mathbf{y})\|_2^2$ , over the input set  $\mathcal{X} \in \mathbb{R}^N$ , hypothesis class  $\mathcal{H} = \{h^{(1)}(\mathbf{x}) = G_{\mathbf{W}}^{(1)}(\mathbf{x}) = \sum_{i=1}^{D^{(1)}} w_i^{(1)} \phi_i^{(1)}(\mathbf{x}) = \mathbf{w}^{(1)T} \Phi^{(1)}(\mathbf{x}), h^{(2)}(\mathbf{x}) = G_{\mathbf{W}}^{(2)}(\mathbf{y}) = \sum_{i=1}^{D^{(2)}} w_i^{(2)} \phi_i^{(2)}(\mathbf{y}) = \mathbf{w}^{(2)T} \Phi^{(2)}(\mathbf{y}) \mid \Phi^{(1)} \in \mathcal{F}_1, \Phi^{(2)} \in \mathcal{F}_2, \forall \mathbf{x} : \|\Phi^{(1)}(\mathbf{x})\|_2 \leq A^{(1)}, \forall \mathbf{y} : \|\Phi^{(2)}(\mathbf{y})\|_2 \leq A^{(2)}\}$ , and the second input set  $\mathcal{Y} \subset \mathbb{R}^N$ . The result is presented in Theorem 4.

#### Theorem 4: Generalization bound, implicit regression

For the energy function  $E(h, \mathbf{x}, \mathbf{y}) = \frac{1}{2} \|G_{\mathbf{W}}^{(1)}(\mathbf{x}) - G_{\mathbf{W}}^{(2)}(\mathbf{y})\|_2^2$ , if the feature set  $\{\phi_1^{(1)}(\cdot), \dots, \phi_{D^{(1)}}^{(1)}(\cdot)\}$  is  $\vartheta^{(1)}$ -diverse with a probability  $\tau_1$  and the feature set  $\{\phi_1^{(2)}(\cdot), \dots, \phi_{D^{(2)}}^{(2)}(\cdot)\}$  is  $\vartheta^{(2)}$ -diverse with a probability  $\tau_2$ , then with a probability of at least  $(1 - \delta)\tau_1\tau_2$ , the following holds for all  $h$  in  $\mathcal{H}$ :

$$\begin{aligned} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{Z}}[E(h, \mathbf{x}, \mathbf{y})] &\leq \frac{1}{m} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{S}} E(h, \mathbf{x}, \mathbf{y}) \\ &\quad + 8(\sqrt{\mathcal{J}_1} + \sqrt{\mathcal{J}_2}) \left( D^{(1)} \|\mathbf{w}^{(1)}\|_{\infty} \mathcal{R}_m(\mathcal{F}_1) \right. \\ &\quad \left. + D^{(2)} \|\mathbf{w}^{(2)}\|_{\infty} \mathcal{R}_m(\mathcal{F}_2) \right) \\ &\quad + (\mathcal{J}_1 + \mathcal{J}_2) \sqrt{\frac{\log(2/\delta)}{2m}}, \end{aligned} \tag{3.21}$$

where  $\mathcal{J}_1 = \|\mathbf{w}^{(1)}\|_{\infty}^2 (D^{(1)} A^{(1)^2} - \vartheta^{(1)^2})$  and  $\mathcal{J}_2 = \|\mathbf{w}^{(2)}\|_{\infty}^2 (D^{(2)} A^{(2)^2} - \vartheta^{(2)^2})$ .

The upper-bound of the energy model is contingent on the diversity variable of both feature sets. Notably, the bound for implicit regression diminishes in direct proportion to  $\vartheta^2$ , which contrasts with cases such as classification, where the bound is proportional to  $\vartheta$ . This observation leads to the conclusion that diminishing redundancy enhances the generalization performance of EBMs within the context of implicit regression.

### 3.3.1.3 Discussion

In this section of the dissertation, we introduced the concept of  $(\vartheta - \tau)$ -diversity to characterize the feature diversity within the EBM models. Several rigorous generalization bounds were derived for different settings, i.e., regression, classification, and implicit regression. The results show that learning diverse feature is consistently beneficial to EBM models. Notably, our theoretical framework is agnostic toward both the choice of loss function and the specifics of the training strategy employed for parameter optimization in EBMs. This noteworthy characteristic provides more general theoretical guarantees that learning non-redundant feature is beneficial to EBM generalization. This provides an answer to Research Question 3, showing that feature diversity helps improve EBM generalization.

Similar to  $d_{min}$  in Section 3.1, the main limitation of  $(\vartheta - \tau)$ -diversity is sensitivity to the feature norm. Future research direction include using scale-invariant measures, e.g., correlation or Mutual Information, to mitigate this issue.

## 3.3.2 Feature Diversity in Energy-Based Models: Algorithms

Beyond the theoretical results in Publication V, in this dissertation, we also show how to translate these results into practical algorithms. Inspired by Section 3.2, we develop and test a practical approach that directly encourages the EBM to learn diverse features during the training process and we evaluate the proposed approach in several contexts and tasks.

### 3.3.2.1 Methodology

In the realm of deep learning, theoretical generalization bounds often lack direct practical implications due to their inherent looseness [10, 105]. Nonetheless, these bounds frequently serve as a guide for introducing regularizers to encourage desirable properties within the hypothesis class [28, 127, 128]. Building upon the theoretical insights from Section 3.3.1.2, we propose a direct strategy

to mitigate the acquisition of redundant features by introducing regularization during model training, which is inversely proportional to the  $(\vartheta - \tau)$ -diversity of the features. Given an EBM with a learnable feature set  $\{\phi_1(\cdot), \dots, \phi_D(\cdot)\}$  and a training set  $S$ , we propose to augment the original training loss  $L$  as follows:

$$L_{aug} = L - \beta \sum_{\mathbf{x} \in S} \sum_{i \neq j}^D (\phi_i(\mathbf{x}) - \phi_j(\mathbf{x}))^2, \quad (3.22)$$

where  $\beta$  is a hyperparameter that controls the contribution of the regularizer to the total loss. The approach presented in equation 3.22 can be incorporated in a plug-and-play manner into any EBM approach to ensure the models learn a rich representation composed of non-redundant features.

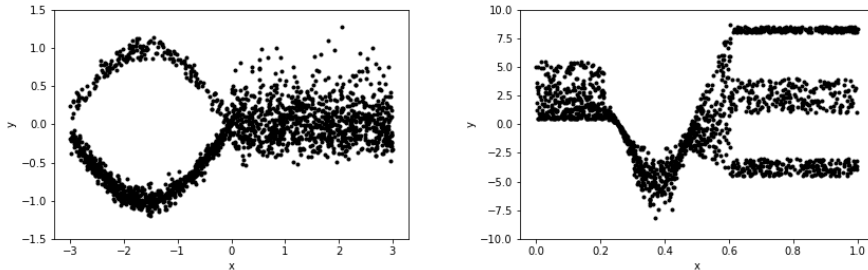
### 3.3.2.2 Empirical Results

#### Regression Task

EBMs have gained significant attention for their efficacy in addressing regression tasks, as shown in [36, 37, 38]. Here, we demonstrate that learning diverse features contributes to superior generalization. To validate the proposed regularizer introduced in equation 3.22, we conduct experiments on two 1-D regression tasks. The first task involves the dataset introduced in [37], comprising 2,000 training examples, as illustrated in the left side of Figure 3.7. For the second task, we evaluate our approach on the dataset in [36]. We utilize the dataset proposed in [129] following the same methodology as [36]. This dataset incorporates 1,900 test examples and 1,700 training examples, with the training data visualized in the right panel of Figure 3.7.

Various loss functions have been proposed to train EBMs for regression [38, 130, 131]. In alignment with the methodology presented in [37], we use the noise contrastive estimation (NCE) loss [131], defined by the noise distribution  $q(y) = \frac{1}{2} \sum_{j=1}^2 \mathcal{N}(y; y_i, \sigma_j^2 \mathbf{I})$ , where  $\sigma_1$  and  $\sigma_2$  serve as hyperparameters. Following the suggestions in [36, 37], we set  $\sigma_2 = 8\sigma_1$  in all experiments. To evaluate our approach, we augment the NCE loss using equation 3.22 to penalize feature redundancy.

The experimental setup mirrors that of [36, 37]. The inner model consists



**Figure 3.7** Visualization of the training data for the 1-D regression tasks: The dataset [37] on the left and the dataset [129] on the right [Publication V].

of two fully connected layers with ReLU activations and its final output is used to compute our proposed regularizer. The input  $y$  is fed to one fully-connected layer with 10 neurons and a ReLU activation function. Furthermore, the concatenated outputs of both models pass through a network composed of four hidden layers. All hidden layers have 10 units except the final output, which has one hidden unit, i.e., the estimated energy. The model undergoes end-to-end training for 75 epochs using Adam optimizer [2] and a learning rate of 0.001. Consistent with [36, 37], the batch size is set to 32, and the number of samples  $M$  is always fixed at  $M = 1024$ .

In Table 3.14, we present the results of the EBM trained with NCE and the EBM augmented with our proposed regularizer for 1-D regression tasks on two datasets [37, 129]. The reported metrics include the approximate KL divergence for the first dataset and the approximate Negative Log-Likelihood (NLL) for the second dataset, considering three different values of  $\sigma_1 \in \{0.05, 0.1, 0.2\}$ .

For the first dataset [37], our approach consistently outperforms the baseline EBM across all values of  $\sigma_1$ . Specifically, our method achieves a notable reduction in KL divergence, indicating enhanced model performance. The gains are particularly pronounced when  $\sigma_1$  is set to 0.05 and 0.1, demonstrating the efficacy of feature diversity strategy in improving the model’s representation of the underlying data distribution.

Similarly, for the second dataset [129], our approach consistently exhibits superior performance compared to the EBM without regularization. The reduction in NLL values across all values of  $\sigma_1$  signifies the efficacy of our regularizer in improving the model’s ability to capture the data distribution. Notably,

**Table 3.14** Results of the EBM trained with NCE (EBM) and the EBM trained with NCE augmented with our regularizer (ours) for the 1-D regression tasks. We report the approximate KL divergence for the first dataset [37], and the approximate NLL for the second dataset [129]. For each dataset, we report the results for three different values of  $\sigma_1$ .

Approach	Dataset [37]		
	$\sigma_1 = 0.05$	$\sigma_1 = 0.1$	$\sigma_1 = 0.2$
EBM	0.0445	0.0420	0.0374
ours ( $\beta = 1e^{-11}$ )	<b>0.0398</b>	0.0345	0.0357
ours ( $\beta = 1e^{-12}$ )	0.0409	0.0380	<b>0.0343</b>
ours ( $\beta = 1e^{-13}$ )	0.0410	<b>0.0332</b>	0.0377

Approach	Dataset [129]		
	$\sigma_1 = 0.05$	$\sigma_1 = 0.1$	$\sigma_1 = 0.2$
EBM	2.7776	2.5650	1.9876
ours ( $\beta = 1e^{-11}$ )	2.6187	2.4414	<b>1.8072</b>
ours ( $\beta = 1e^{-12}$ )	<b>2.5846</b>	<b>2.3685</b>	1.8880
ours ( $\beta = 1e^{-13}$ )	2.7483	2.5420	1.9303

the most substantial improvement is observed when  $\sigma_1$  is set to 0.2, showcasing the adaptability and effectiveness of our approach under varying noise levels. Overall, the results on both datasets underscore the consistent gains achieved by our approach in enhancing the performance of EBMs for 1-D regression tasks across different datasets and noise levels.

In addition to conducting regression experiments on the two toy datasets, we assess the efficacy of our methodology in a more challenging regression dataset involving image-based age estimation. Specifically, we employ the UTKFace Age Estimation dataset [132], which comprises 14,760 facial images along with corresponding age labels. The objective is to predict individuals' ages based on their images. In line with the methodology of [36], we allocate 80% of the dataset for training and the remaining 20% for testing. We adhere to the experimental setup detailed in [36] and adopt their EBM as a baseline. To assess the effectiveness of our approach, we enhance their loss function with our proposed regularizer. The results are presented in Table 3.15.



**Table 3.15** Results in terms of approximate NLL for the EBM age estimation experiments. The results are reported as the mean/SEM over these runs.

Approach	NLL
EBM [36]	$4.12 \pm 0.07$
ours ( $\beta = 1e^{-11}$ )	$4.04 \pm 0.10$
ours ( $\beta = 1e^{-12}$ )	$4.03 \pm 0.04$
ours ( $\beta = 1e^{-13}$ )	$3.99 \pm 0.15$

As shown by results in Table 3.15, introducing our regularizer to standard EBM leads to notable improvements. This confirms that feature diversity is helpful also for large datasets. Specifically, with  $\beta = 1e^{-13}$ , our method achieves a NLL of  $3.99 \pm 0.15$  compared to  $4.12 \pm 0.07$  of the standard EBM. These results suggest that deploying a diversity regularization approach enriches the feature representation of the model and consistently enhances the generalization capabilities of the EBM.

## Image Generation

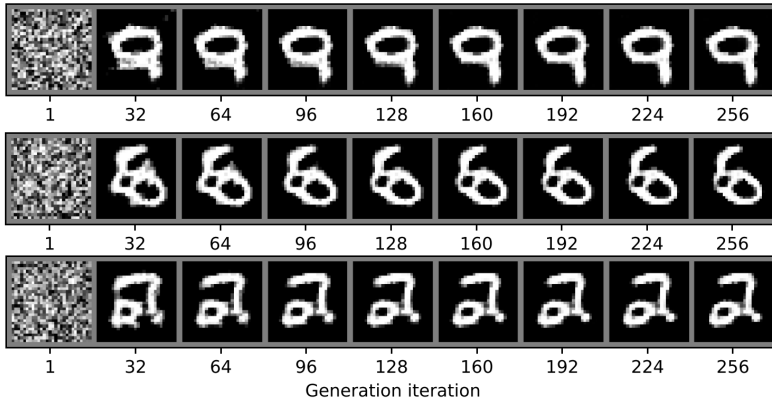
Recent attention has been directed toward leveraging EBMs for image and text generation tasks [8, 44, 46]. In this dissertation, we evaluate the proposed regularizer on image generation using the MNIST digits [44]. The EBM architecture is presented in Table 3.16. We use the same training protocol as in [44, 133], employing Langevin dynamics Markov chain Monte Carlo (MCMC) and a sampling buffer to accelerate training. All models were trained for 60 epochs using Adam optimizer with learning rate of  $1e^{-4}$  and a batch size of 128.

Our approach, i.e., augmenting the contrastive divergence loss using equation 3.22, is designed to discourage redundancy in the acquired features in the last intermediate layer. In Table 3.17, we report the Fréchet Inception Distance (FID) score [134] and the NLL loss for varying values of  $\beta$ . The results demonstrate that penalizing the similarity of learned features leads to improved FID and NLL scores. The optimal performance is obtained with  $\beta = 1e^{-13}$ , showcasing a notable over 10% enhancement in FID compared to the original EBM.

In Figure 3.8, we present some qualitative results for our approach. We

**Table 3.16** Simple CNN model used in the example. \* refers to the features' layer.

Layer	Output shape
Input	[1,28,28]
Cov (16 $5 \times 5$ )	[16,16,16]
Swish activation	[16,16,16]
Cov (32 $3 \times 3$ )	[32,8,8]
Swish activation	[32,8,8]
Cov (64 $3 \times 3$ )	[64,4,4]
Swish activation	[64,4,4]
Cov (64 $3 \times 3$ )	[64,2,2]
Swish activation	[64,2,2]
Flatten	[256]
Linear	[64]
Swish activation*	[64]
Linear	[1]



**Figure 3.8** Qualitative results of our approach ( $\beta = 1e^{-13}$ ) : Few intermediate samples of the MCMC sampling (Langevin Dynamics).

plot intermediate samples of the MCMC sampling (Langevin Dynamics) for  $\beta = 1e^{-13}$ . Initiating from random noise, MCMC obtains reasonable figures after only 64 steps. The digits get clearer and more realistic over the iterations.

**Table 3.17** Table of FID scores and NLL loss of different approaches for generations of MNIST images. Each experiment was performed three times with different random seeds, the results are reported as the mean/SEM over these runs.

Approach	FID	NLL loss
EBM	$0.0109 \pm 0.0004$	$0.7112 \pm 0.0190$
ours ( $\beta = 1e^{-11}$ )	$0.0107 \pm 0.0004$	$0.7109 \pm 0.0111$
ours ( $\beta = 1e^{-12}$ )	$0.0104 \pm 0.0003$	<b><math>0.7105 \pm 0.0112</math></b>
ours ( $\beta = 1e^{-13}$ )	<b><math>0.0099 \pm 0.0006</math></b>	$0.7108 \pm 0.0111$

## Continual Learning

We substantiate the efficacy of our approach in a more challenging task, specifically, the Continual Learning (CL) problem. CL addresses the challenge of catastrophic forgetting in deep learning models [135, 136, 137]. The primary objective of CL is to sequentially learn multiple tasks while retaining knowledge acquired from previous tasks. Consequently, a continual learner is anticipated to generalize well in new tasks without compromising the understanding of previously learned tasks. Notably, [76] proposed an EBM-based approach for CL that led to superior results on this task.

In this experimental setup, we use the models of [76] as baseline and use the exact same experimental protocol. Specifically, we evaluate the different models on the class-incremental learning task with CIFAR10 and CIFAR100 datasets, under both the boundary-aware and boundary-agnostic settings [76]. The former characterizes a scenario where there is an explicit separation between consecutive tasks during learning. The latter refers to the situation where data distributions undergo gradual changes without a predefined notion of task boundaries. The results are reported in Table 3.18.

As demonstrated in Table 3.18, promoting feature diversity substantively enhances the performance of the EBM, resulting in consistently superior accuracy across both datasets in both settings. Figure 3.9 shows the accumulated classification accuracy, computed as an average across the tasks at each step of the CL task. Over the course of five tasks, our methodology consistently attains higher classification accuracy compared to the standard EBM, across both boundary-aware and boundary-agnostic configurations.

**Table 3.18** Evaluation of class-incremental learning on both the boundary-aware and boundary-agnostic setting on CIFAR10 and CIFAR100 datasets. Each experiment was performed ten times with different random seeds, the results are reported as the mean/SEM over these runs.

Method	Boundary-aware	
	CIFAR10	CIFAR100
EBM	$39.15 \pm 0.86\%$	$29.02 \pm 0.24\%$
ours ( $\beta = 1e^{-11}$ )	$39.61 \pm 0.81\%$	$29.15 \pm 0.27\%$
ours ( $\beta = 1e^{-12}$ )	<b><math>40.64 \pm 0.79\%</math></b>	<b><math>29.38 \pm 0.21\%</math></b>
ours ( $\beta = 1e^{-13}$ )	$40.15 \pm 0.87\%$	$29.28 \pm 0.28\%$

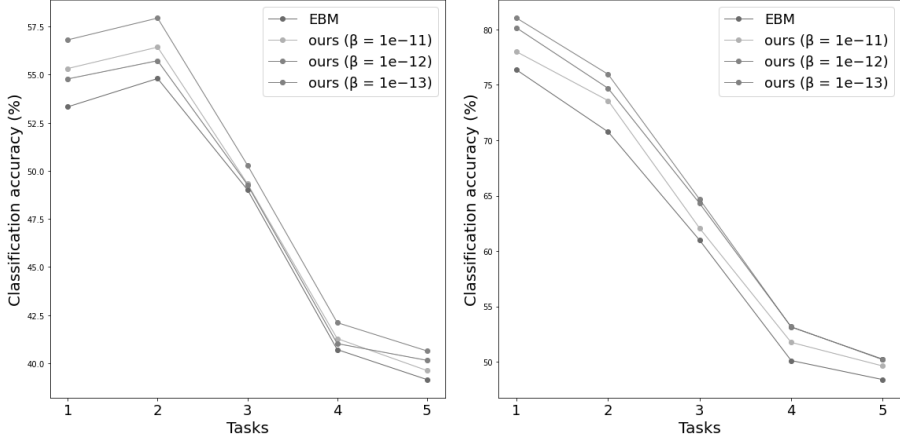
  

Method	Boundary-agnostic	
	CIFAR10	CIFAR100
EBM	$48.40 \pm 0.80\%$	$34.78 \pm 0.26\%$
ours ( $\beta = 1e^{-11}$ )	$49.63 \pm 0.90\%$	$34.86 \pm 0.30\%$
ours ( $\beta = 1e^{-12}$ )	<b><math>50.25 \pm 0.63\%</math></b>	<b><math>35.20 \pm 0.23\%</math></b>
ours ( $\beta = 1e^{-13}$ )	$50.20 \pm 0.94\%$	$35.03 \pm 0.21\%$

### 3.3.2.3 Discussion

In this section of the dissertation, inspired by the theoretical findings, a regularizer that explicitly penalizes similarities within the features set of EBM, was proposed. Empirical evaluation over multiple tasks and datasets consistently corroborates the theoretical findings showing that learning diverse features yields superior performance. This provides insights into Research Question 3 from an empirical prospective, showing that feature diversity methodologies can help boost the performance of EBMs.

We note that the regularizer is simply an illustrated example showing how our theory can be used in practice to inspire different regularization. Thus, when formulating the approach, the aim was to construct it as close as possible to the  $(\vartheta - \tau)$ -diversity definition. To this end, the exact definition of diversity was used directly as the regularizer (equation 3.22). It has two sums: the first encompasses the entire batch, while the second spans all pairs of units within the layers. This yields a cumulative count of  $ND^2$  terms, where N denotes

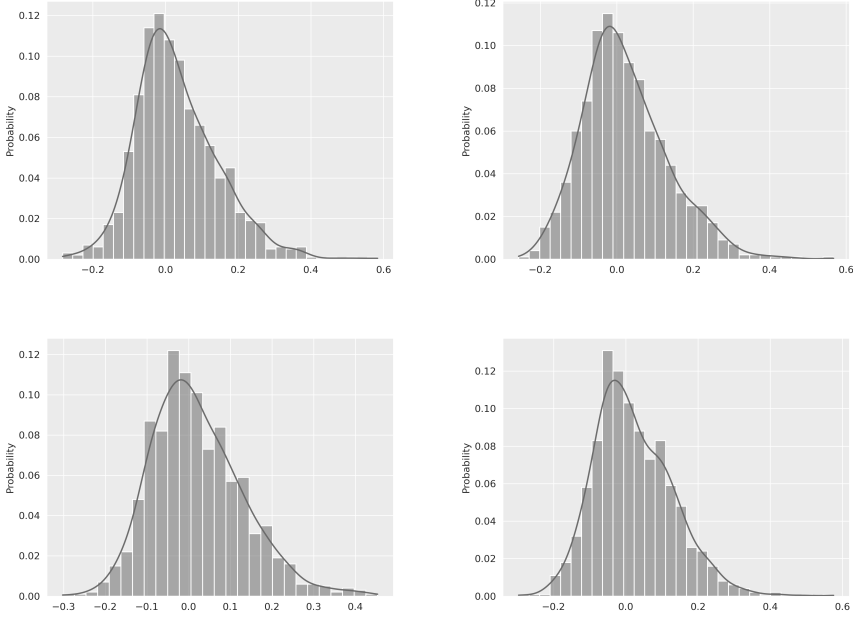


**Figure 3.9** Test classification accuracy vs number of observed tasks on CIFAR10 using the boundary-aware (left) and boundary-agnostic (right) setting. The results are averaged over ten random seeds.

the batch size and  $D$  signifies the number of units within the layer. Notably, this approach results in empirically significant magnitudes for the second term, approximately on the order of  $1e^9$ . Consequently, it is imperative to maintain a small value for the hyperparameter  $\beta$  to prevent the loss function from being dominated by the second term. Our empirical investigations reveal that the interval  $[1e^{-11}, 1e^{-13}]$  is a stable range for  $\beta$ . Additionally, it is pertinent to acknowledge that the incorporation of our proposed regularizer introduces a marginal temporal overhead during training, amounting to less than 1% of the total training time.

### 3.4 Feature Diversity vs Class-wise Overfitting

The preceding sections of this dissertation have exclusively investigated the influence of feature diversity on the generalization of neural networks. The empirical and practical examinations conducted thus far affirm that feature diversity contributes substantively to mitigating overfitting, thereby resulting in a diminished generalization gap. Nevertheless, recent observations underscore a nuance in the impact of overfitting, indicating disparate effects on different classes/categories within the learned task.



**Figure 3.10** Class-generalization errors of the different classes in ImageNet of different approaches: ResNet50 with no diversity regularizer (top left), ResNet50 with WLD-Reg (Direct) (top right), ResNet50 with WLD-Reg (Det) (bottom left), and ResNet50 with WLD-Reg (Logdet) (bottom right).

### 3.4.1 Feature Diversity and Class-wise Generalization

In [14, 47], it has been observed that various regularization techniques, while exhibiting improvements in standard average generalization and mitigating overall overfitting, inadvertently amplify the disparities in generalization across the different classes. Consequently, for a more comprehensive exploration of the impact of feature diversity on generalization, in this dissertation, we conduct an empirical study of the effect of feature diversity, mainly, WLD-Reg introduced in Section 3.2.1 and Publication II on the class-wise generalization performance of neural networks. Specifically, we investigate how different approaches affect the distribution of the class-generalization errors of ResNet50 trained ImageNet, composed of 1000 class. The class-generalization gaps’ histograms corresponding to the different approaches are presented in Figure 3.10.

As can be seen in Figure 3.10 for the different approaches, overfitting

**Table 3.19** Different quantitative measures of the effect of different approaches on the class-generalization error of the model.

	Variance	Worst 25%	Worst 10%	Worst 5%
Standard	1.27%	18.53%	25.82%	30.87%
WLD-Reg (Direct)	1.25%	17.45%	24.78%	29.32%
WLD-Reg (Det)	1.21%	16.93%	23.94%	28.80%
WLD-Reg (Logdet)	<b>1.18%</b>	<b>16.77%</b>	<b>23.31%</b>	<b>27.97%</b>

does not uniformly impact all classes. Rather, the models exhibit a nuanced class-generalization performance, demonstrating tendencies to underfit certain classes, effectively learn others, and manifest varying degrees of overfitting across other classes. Employing a feature diversity-promoting approach, namely WLD-Reg, alternates the distribution of the class-generalization gaps, with the Det and Logdet variants demonstrating more clear shifts. To further assess the impact of WLD-Reg on class-wise generalization performance of the models, we report various quantitative measures assessing the impact of different approaches. The metrics include the variance of class-generalization gaps, the mean of the worst 25%, 10%, and 5% of the class-generalization gaps of the different models. The results are reported in Table 3.19.

Table 3.19 shows that, in comparison to the standard approach, using WLD-Reg exhibits notable advantages. Specifically, both the direct and the Det variants approach result in reductions across all metrics with the Logdet approach demonstrating even more pronounced improvements. For instance, in terms of variance, our Logdet approach achieves a reduction to 1.18% compared to the standard 1.27%. These results underscore the efficacy of our approaches in enhancing the class-generalization performance of the model, particularly in mitigating disparities within the worst-performing subsets.

### 3.4.2 Class-wise Generalization: an Information-Theoretic Perspective

Here, we present the main findings of Publication VI.

**Notations:** We use upper-case letters to denote random variables, e.g.,  $\mathbf{Z}$ ,

and lower-case letters to denote the realization of random variables.  $\mathbb{E}_{\mathbf{Z} \sim P}$  denotes the expectation of  $\mathbf{Z}$  over a distribution  $P$ . Consider a pair of random variables  $\mathbf{W}$  and  $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$  with joint distribution  $P_{\mathbf{W}, \mathbf{Z}}$ . Let  $\overline{\mathbf{W}}$  be an independent copy of  $\mathbf{W}$  and  $\overline{\mathbf{Z}} = (\overline{\mathbf{X}}, \overline{\mathbf{Y}})$  be an independent copy of  $\mathbf{Z}$ , such that  $P_{\overline{\mathbf{W}}, \overline{\mathbf{Z}}} = P_{\mathbf{W}} \otimes P_{\mathbf{Z}}$ . For random variables  $\mathbf{X}$ ,  $\mathbf{Y}$  and  $\mathbf{Z}$ ,  $I(\mathbf{X}; \mathbf{Y}) \triangleq D(P_{\mathbf{X}, \mathbf{Y}} \| P_{\mathbf{X}} \otimes P_{\mathbf{Y}})$  denotes the mutual information (MI), and  $I_z(\mathbf{X}; \mathbf{Y}) \triangleq D(P_{\mathbf{X}, \mathbf{Y} | \mathbf{Z}=z} \| P_{\mathbf{X} | \mathbf{Z}=z} \otimes P_{\mathbf{Y} | \mathbf{Z}=z})$  denotes disintegrated conditional mutual information (CMI), and  $\mathbb{E}_{\mathbf{Z}}[I_{\mathbf{Z}}(\mathbf{X}; \mathbf{Y})] = I(\mathbf{X}; \mathbf{Y} | \mathbf{Z})$  is the standard CMI. We will also use the notation  $\mathbf{X}, \mathbf{Y} | z$  to simplify  $\mathbf{X}, \mathbf{Y} | \mathbf{Z} = z$  when it is clear from the context.

### 3.4.2.1 Problem Formulation

In order to study the class-wise generalization, we rely on the conditional mutual information (CMI) framework [15], which has been shown to yield tight generalization bounds [138, 139].

In the standard CMI setting [15], we assume that the dataset is formed of  $n$  super-samples  $\mathbf{Z}_{[2n]} = (\mathbf{Z}_1^\pm, \dots, \mathbf{Z}_n^\pm) \in \mathcal{Z}^{2n}$  i.i.d. generated from  $P_{\mathbf{Z}}$ . The training samples are selected using  $n$  independent Rademacher random variables  $\mathbf{U} = (\mathbf{U}_1, \dots, \mathbf{U}_n) \in \{-1, 1\}^n$ . Each  $\mathbf{U}_i$  selects a sample  $\mathbf{Z}_i^{\mathbf{U}_i}$  from the pair  $\mathbf{Z}_i^\pm$  to be used in training, and the remaining one  $\mathbf{Z}_i^{-\mathbf{U}_i}$  is assigned to the testing dataset. Let  $\mathbf{S} = (\mathbf{Z}_1^{\mathbf{U}_1}, \mathbf{Z}_2^{\mathbf{U}_2}, \dots, \mathbf{Z}_n^{\mathbf{U}_n})$  denote the training dataset. Furthermore, for a specific class  $y \in \mathcal{Y}$ , let  $n^y = nP(\mathbf{Y} = y)$ , the number of supersamples  $n$  scaled with the probability of class  $y$ .

Our main goal is to study the generalization error of a specific class  $y$ . To this end, we define the class-generalization error in the CMI setting as follows:

**Definition 2.** (*class-generalization error*) For any  $y \in \mathcal{Y}$ , the class-generalization error is defined as

$$\overline{\text{gen}}_y \triangleq \mathbb{E}_{\mathbf{Z}_{[2n]}} \left[ \frac{1}{n^y} \sum_{i=1}^n \mathbb{E}_{\mathbf{U}_i, \mathbf{W} | \mathbf{Z}_{[2n]}} \left[ \mathbb{1}_{\{Y_i^{-\mathbf{U}_i} = y\}} \ell(\mathbf{W}, \mathbf{Z}_i^{-\mathbf{U}_i}) - \mathbb{1}_{\{Y_i^{\mathbf{U}_i} = y\}} \ell(\mathbf{W}, \mathbf{Z}_i^{\mathbf{U}_i}) \right] \right], \quad (3.23)$$

where  $\mathbb{1}_{\{a=b\}}$  is the indicator function, returning 1 when  $a = b$  and zero otherwise.



Definition 2 quantifies the expected generalization error between the training set and the test set with respect to a particular class  $y$ . In contrast to the conventional generalization error definition in the CMI setting [15, 138], we underscore two main distinctions: (i) our class-generalization error uses indicator functions to exclusively account for samples belonging to the specified class  $y$ , and (ii) our generalization error is normalized by  $n^y$ , in contrasts with the standard practice of averaging over  $n$  samples in the conventional super-sample setting.

### 3.4.2.2 Main Results

The following theorem provides a bound for the class-generalization error using the disintegrated conditional mutual information between  $\mathbf{W}$  and the selection variable  $\mathbf{U}_i$  conditioned on super-sample  $\mathbf{Z}_{[2n]}$ .

#### **Theorem 5: class-CMI**

Assume that the loss  $\ell(w, x, y) \in [0, 1]$  is bounded, then the class-generalization error for class  $y$  in Definition 2 can be bounded as

$$|\overline{\text{gen}}_y| \leq \mathbb{E}_{\mathbf{Z}_{[2n]}} \left[ \frac{1}{n^y} \sum_{i=1}^n \sqrt{2 \max(\mathbb{1}_{\{\mathbf{Y}_i^- = y\}}, \mathbb{1}_{\{\mathbf{Y}_i^+ = y\}}) I_{\mathbf{Z}_{[2n]}}(\mathbf{W}; \mathbf{U}_i)} \right].$$

The detailed proof is available in Publication VI. The key idea is selecting the function in Donsker-Varadhan’s variational representation to correspond to the definition of class-generalization error. Next, using the fact that for a fixed realization  $z_{[2n]}$ ,  $\mathbb{1}_{\{y^{\mathbf{U}_i} = y\}} \ell(\mathbf{W}, z_i^{\mathbf{U}_i}) - \mathbb{1}_{\{y^{-\mathbf{U}_i} = y\}} \ell(\mathbf{W}, z_i^{-\mathbf{U}_i}) = \mathbf{U}_i(\mathbb{1}_{\{y_i^- = y\}} \ell(\mathbf{W}, z_i^-) - \mathbb{1}_{\{y_i^+ = y\}} \ell(\mathbf{W}, z_i^+))$  with Hoeffding’s Lemma [140] yields the final result.

Theorem 5 establishes an upper bound for the class-specific generalization error explicitly dependent on the learned parameters  $\mathbf{W}$ . The result suggest that the extent of information revealed by the random selection  $\mathbf{U}_i$  about the parameters  $\mathbf{W}$  is pivotal in determining the class-specific generalization error, particularly when one of the two samples,  $z_i^\pm$ , belongs to the class  $y$ . Previous works [141, 142, 143] has established connections between overfitting and the memorization of weights. Theorem 5 shows that this observation extends to

our context, where if the model parameters  $\mathbf{W}$  effectively memorize the random selection  $\mathbf{U}$ , the CMI and hence the class-generalization error will be large.

Although the bound presented in Theorem 5 is finite given the binary nature of  $\mathbf{U}_i$ , practical evaluation of  $I_{\mathbf{Z}_{[2n]}}(\mathbf{W}; \mathbf{U}_i)$  poses empirical challenges, particularly when dealing with high-dimensional  $\mathbf{W}$  as encountered in deep neural networks. A proposed strategy to address this challenge is to shift the focus from the model weights  $\mathbf{W}$  to the model predictions  $f_{\mathbf{W}}(\mathbf{X}_i^{\pm})$ , as proposed in [144]. We use a similar strategy in our context and derive a rigorous bound of the class-generalization error using the disintegrated CMI between the model prediction and the random selection, i.e.,  $I_{\mathbf{Z}_{[2n]}}(f_{\mathbf{W}}(\mathbf{X}_i^{\pm}); \mathbf{U}_i)$ . The main result is present in Theorem 6.

#### **Theorem 6: class-f-CMI**

Assume that the loss  $\ell(\hat{y}, y) \in [0, 1]$  is bounded, then the class-generalization error for class  $y$  in Definition 2 can be bounded as

$$|\overline{\text{gen}}_y| \leq \mathbb{E}_{\mathbf{Z}_{[2n]}} \left[ \frac{1}{ny} \sum_{i=1}^n \sqrt{2 \max(\mathbb{1}_{\{\mathbf{Y}_i^- = y\}}, \mathbb{1}_{\{\mathbf{Y}_i^+ = y\}}) I_{\mathbf{Z}_{[2n]}}(f_{\mathbf{W}}(\mathbf{X}_i^{\pm}); \mathbf{U}_i)} \right].$$

We note two key advantages of the class-f-CMI in Theorem 6 bound compared to the bound in Theorem 5: (i) It is typically easy to estimate as it involves two low-dimensional random variables. (ii) Similar to [144], it does not require access to the model parameters  $\mathbf{W}$  but solely relies on the model output  $f(\cdot)$ . This characteristic renders it well-suited for non-parametric methods and black-box algorithms.

The main limitation of the two previous bounds in Theorems 5 and 6 is the dependency on  $\max(\mathbb{1}_{\{\mathbf{Y}_i^- = y\}}, \mathbb{1}_{\{\mathbf{Y}_i^+ = y\}})$ . This term can potentially make the bounds loose. Specifically, in the scenarios where one sample in the pair  $(Z_i^-, Z_i^+)$  belongs to class  $y$  and the other do not, this term is non-zero and the information from both samples of the pair contributes to the bound ( $I_{\mathbf{Z}_{[2n]}}(f_{\mathbf{W}}(\mathbf{X}_i^{\pm}); \mathbf{U}_i)$ ). Consequently, samples from other classes ( $\neq y$ ) can still affect these bounds.

To overcome this limitation, we consider a new random variable  $\Delta_y \mathbf{L}_i$  based on the indicator function and the loss, i.e.,

$$\Delta_y \mathbf{L}_i \triangleq \mathbb{1}_{\{y_i^- = y\}} \ell(f_{\mathbf{W}}(\mathbf{X}_i)^-, y_i^-) - \mathbb{1}_{\{y_i^+ = y\}} \ell(f_{\mathbf{W}}(\mathbf{X}_i)^+, y_i^+). \quad (3.24)$$

$\Delta_y \mathbf{L}_i$  can be interpreted as a weighted sum of class-dependent losses. Theorem 7 provides a bound depending on the CMI between this newly introduced variable and the random selection.

**Theorem 7: class- $\Delta_y L$ -CMI**

Assume that the loss  $\ell(\hat{y}, y) \in [0, 1]$ , then the class-generalization error of class  $y$  defined in 2 can be bounded as

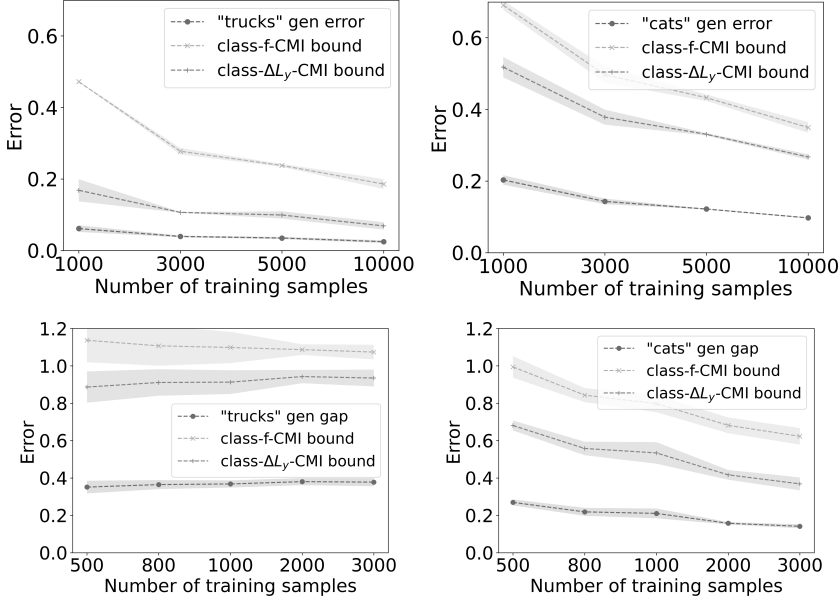
$$|\overline{\text{gen}}_y| \leq \mathbb{E}_{\mathbf{Z}_{[2n]}} \left[ \frac{1}{n^y} \sum_{i=1}^n \sqrt{2I_{\mathbf{Z}_{[2n]}}(\Delta_y \mathbf{L}_i; \mathbf{U}_i)} \right]. \quad (3.25)$$

Moreover, the  $\Delta_y L$ -CMI bound is always tighter than the class- $f$ -CMI bound in Theorem 6, and the latter is always tighter than the class-CMI bound in Theorem 5.

The key advantages of the bound in Theorem 7 are: (i) It is easier to estimate as  $\Delta_y \mathbf{L}_i$  is a one-dimensional scalar, whereas  $f_{\mathbf{W}}(\mathbf{X}_i^\pm)$  is two-dimensional. (ii) It is tighter than bounds in Theorems 6 and 5. Intuitively, the difference between two weighted loss values,  $\Delta_y \mathbf{L}_i$ , reveals considerably less information about the selection process  $\mathbf{U}_i$  compared to the pair  $f_{\mathbf{W}}(\mathbf{X}_i^\pm)$ .

Here, we perform empirical experiments to assess the efficacy of our class-generalization error bounds. Specifically, we evaluate the bounds in Theorems 6 and 7. As previously mentioned, the bounds outlined here are notably easy to estimate in practice. We follow the same experimental settings in [144], i.e., we fine-tune a ResNet-50 [4] on the CIFAR10 dataset [109] (Pretrained on ImageNet [145]). Additionally, we extend our experimentation to a more challenging dataset—specifically, a noisy variant of CIFAR10 with 5% label noise.

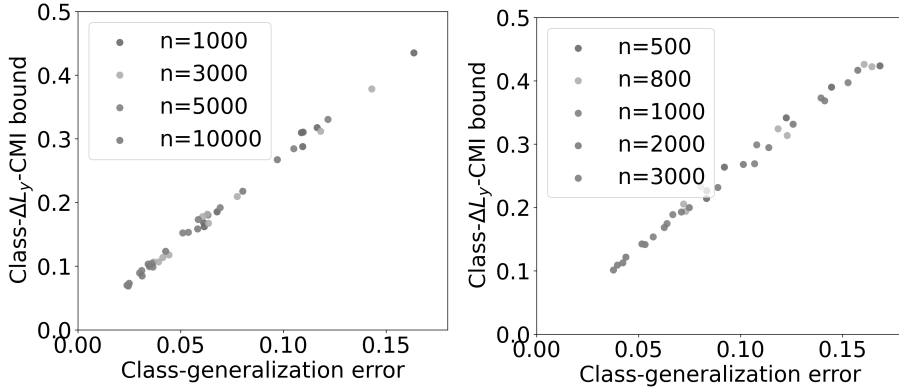
The class-generalization error of two classes “trucks” and “cats”, along with the bounds in Theorems 6 and 7 are presented in the first two columns of Figure 3.11 (first row for CIFAR10 and second row for noisy CIFAR10). As can be seen, both bounds are able to capture the behaviour of the class-generalization



**Figure 3.11** Experimental results of class-generalization error and our bounds in Theorems 6 and 7 for the class of “trucks” (left) and “cats” (right) in CIFAR10 (top) and noisy CIFAR10 (bottom), as we increase the number of training samples [Publication VI].

error with the class- $\Delta_y L$ -CMI bound being consistently tighter. For instance, as we increase the number of training samples, the “trucks” class generalization error in CIFAR10 decreases at a low rate, whereas in the noisy CIFAR10, it increases. The “cats” class in CIFAR10 has a large slope at the start and then an incremental decrease. The class- $\Delta_y L$ -CMI, as shown in Figure 3.11, precisely predicts the behaviour of class-generalization error. All these complex behaviours of class-generalization errors are successfully captured by the class- $\Delta_y L$ -CMI bound. Notably, this bound exhibits a proportional relationship with the true class-generalization error, wherein an elevated class- $\Delta_y L$ -CMI bound corresponds to a heightened class-generalization error.

To underscore the ability of our bound in predicting the behaviour of the true class-generalization error, we plot in Figure 3.12 a scatter plot illustrating the relationship between distinct class-generalization errors and their respective class- $\Delta_y L$ -CMI bound values for various classes in CIFAR10 (left) and Noisy CIFAR10 (right), under varying sample sizes. Notably, a linear correlation is observed between our bound and the class-generalization error, demonstrating



**Figure 3.12** The scatter plots between the bound in Theorem 7 and the class-generalization error of the different classes for CIFAR10 (left) and noisy CIFAR10 (right) [Publication VI].

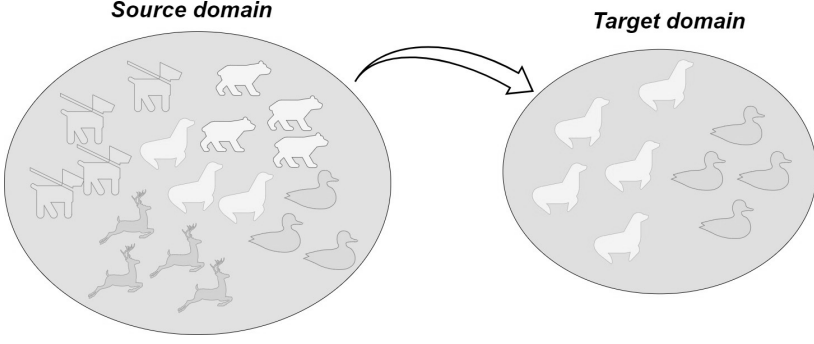
its efficacy in predicting the behaviour of the latter. This implies that one can potentially use our bound for predicting the relative generalization performance across different classes.

### 3.4.2.3 Extra Applications

In addition to allowing the study of class-generalization errors, the theoretical tools developed in this section offer the opportunity to glean theoretical insights across various applications. In this part, we delve into several potential use cases for the tools developed.

#### The Subtask Problem

The subtask problem is a specific instance of distribution shift in supervised learning. The training data generated from the source domain  $P_{\mathbf{X},\mathbf{Y}}$  encompasses multiple classes or labels, whereas the test data associated with target domain  $Q_{\mathbf{X},\mathbf{Y}}$  is limited to a specific subset of the classes encountered during the training phase. An example of the subtask problem is illustrated in Figure 3.13. The motivation behind this problem arises in scenarios where a large model has undergone training on an extensive array of classes, possibly thousands. However, the model is subsequently employed in a target environment where only a limited subset of the classes, observed during the training phase,



**Figure 3.13** Illustration of the Subtask problem. The source domain is composed of 5 different classes and the target domain is composed of data from two particular classes encountered during training in the source domain.

is present.

By tackling the problem as a standard domain adaptation task, the generalization error of the subtask problem,  $\overline{\text{gen}}_{Q, E_P}$ , can be bounded as follows:

$$|\overline{\text{gen}}_{Q, E_P}| \leq \sqrt{2\sigma^2 D(Q\|P)} + \sqrt{2\sigma^2 I(\mathbf{W}; \mathbf{S})}, \quad (3.26)$$

where  $\mathbf{S}$  is the training data and  $E_P$  is the empirical risk on the source domain. The bound in equation 3.26, similar to classic domain adaptation bounds [146, 147], is based on the KL divergence  $D(Q_{\mathbf{X}, \mathbf{Y}}\|P_{\mathbf{X}, \mathbf{Y}})$  and does not utilize the fact that the target task is encapsulated within the source task.

Deriving tight bounds for the subtask problem is straightforward using our class-wise generalization bounds. In essence, using Jensen's inequality, we can show that

$$|\overline{\text{gen}}_{Q, E_Q}| = |\mathbb{E}_{\mathbf{Y} \sim Q_{\mathbf{Y}}} [\overline{\text{gen}}_{\mathbf{Y}}]| \leq \mathbb{E}_{\mathbf{Y} \sim Q_{\mathbf{Y}}} [|\overline{\text{gen}}_{\mathbf{Y}}|], \quad (3.27)$$

where  $E_Q$  is the empirical risk relative to the target domain and  $\overline{\text{gen}}_{\mathbf{Y}}$  is the class-generalization error of class  $\mathbf{Y}$ . Equation 3.27 shows that it is possible to bound the subtask task error using the expectation over the class-generalization error. For instance, using Theorem 7, we can bound the subtask generalization error bound using  $\Delta_{\mathbf{Y}} \mathbf{L}_i; \mathbf{U}_i$ ). The main result is presented in Theorem 8.

The bound in Theorem 8 is discrepancy-independent. It involves only a quantity analog to the second term with the mutual information in equation 3.26. The first term, in the latter, depends on some measure that quantifies

the discrepancy between the target and domain distributions. This demonstrates the tightness of our bound. We note that similarly, it is possible to extend the results of Theorems 5 and 6 to derive CMI and f-CMI bounds for the subtask problem.

**Theorem 8: subtask- $\Delta L_y$ -CMI**

Assume that the loss  $\ell(w, x, y) \in [0, 1]$  is bounded, Then the subtask generalization error can be bounded as

$$|\overline{\text{gen}}_{Q, E_Q}| \leq \mathbb{E}_{\mathbf{Y} \sim Q_{\mathbf{Y}}} \left[ \mathbb{E}_{\mathbf{Z}_{[2n]}} \left[ \frac{1}{n^{\mathbf{Y}}} \sum_{i=1}^n \sqrt{2I_{\mathbf{Z}_{[2n]}}(\Delta_{\mathbf{Y}} \mathbf{L}_i; \mathbf{U}_i)} \right] \right].$$

## Generalization Certificates With Sensitive Attributes

A primary concern impeding the deployment of machine learning models in high-stake applications revolves around the potential biases associated with sensitive attributes, including but not limited to gender and skin color [148, 149]. Consequently, it is imperative not only to mitigate sensitivity to such attributes but also to provide guarantees regarding the fairness of the models [150, 151]. An integral facet of fairness entails ensuring that the machine learning model exhibits equitable generalization performance across various minority groups characterized by different sensitive attributes [149, 152].

By refining the definition of our class-generalization error, we demonstrate the applicability of the theoretical tools developed in this dissertation to establish rigorous bounds for attribute-generalization errors. Let us consider a random variable  $\mathbf{T} \in \mathcal{T}$  representing a sensitive feature. The focus is on investigating the model’s generalization performance for the sub-population characterized by the attribute  $\mathbf{T} = t$ . Drawing inspiration from our class-generalization framework, we introduce the attribute-generalization error with the following definition:

**Definition 3.** (*attribute-generalization error*) Given  $t \in \mathcal{T}$ , the attribute-

generalization error is defined as follows:

$$\overline{\text{gen}}_t \triangleq \mathbb{E}_{\mathbf{Z}_{[2n]}} \left[ \frac{1}{n^t} \sum_{i=1}^n \mathbb{E}_{\mathbf{U}_i, \mathbf{W} | \mathbf{Z}_{[2n]}} \left[ \mathbb{1}_{\{\mathbf{T}_i^{-U_i}=t\}} \ell(\mathbf{W}, \mathbf{Z}_i^{-U_i}) - \mathbb{1}_{\{\mathbf{T}_i^{U_i}=t\}} \ell(\mathbf{W}, \mathbf{Z}_i^{U_i}) \right] \right], \quad (3.28)$$

Let  $\Delta_t \mathbf{L}_i \triangleq \mathbb{1}_{\{t_i^- = t\}} \ell(f_{\mathbf{W}}(\mathbf{X}_i)^-, y_i^-) - \mathbb{1}_{\{t_i^+ = t\}} \ell(f_{\mathbf{W}}(\mathbf{X}_i)^+, y_i^+)$ . By exchanging  $\Delta_y \mathbf{L}_i$  and  $\mathbf{Y}$  with  $\mathbf{T}$  and  $\Delta_t \mathbf{L}_i$  in Theorem 7, respectively, we can show the following CMI bound for the attribute-generalization error.

**Theorem 9: subtask- $\Delta$   $L_y$ -CMI**

Assume that the loss  $\ell(\hat{y}, y) \in [0, 1]$ , then the attribute-generalization error of the sub-population  $\mathbf{T} = t$ , can be bounded as follows:

$$|\overline{\text{gen}}_t| \leq \mathbb{E}_{\mathbf{Z}_{[2n]}} \left[ \frac{1}{n^t} \sum_{i=1}^n \sqrt{2I_{\mathbf{Z}_{[2n]}}(\Delta_t \mathbf{L}_i; \mathbf{U}_i)} \right]. \quad (3.29)$$

Similarly, we can extend the results of Theorems 5 and 6 to this case. Using the attribute generalization, we can show that the standard expected generalization error can be bounded as follows:

**Corollary 1.** Assume that the loss  $\ell(\hat{y}, y) \in [0, 1]$ , then

$$|\overline{\text{gen}}(P_{\mathbf{X}, \mathbf{Y}}, P_{\mathbf{W} | \mathbf{S}})| \leq \mathbb{E}_{\mathbf{Y}} \left[ \mathbb{E}_{\mathbf{Z}_{[2n]}} \left[ \frac{1}{n^{\mathbf{T}}} \sum_{i=1}^n \sqrt{2I_{\mathbf{Z}_{[2n]}}(\Delta_{\mathbf{T}} \mathbf{L}_i; \mathbf{U}_i)} \right] \right]. \quad (3.30)$$

The result of Corollary 1 proves that the average generalization error is upper-bounded by the expectation over the attribute-wise generalization. This observation suggests that one potential way to enhance the overall generalization is by reducing the generalization across each individual population.

### 3.4.3 Discussion

In this section, our contributions are two folds: (i) we have analyzed how the feature diversity regularizer in Publication II affects the class-wise generalization error distribution. These findings contribute insights to address the Research



Question 4 showing that feature diversity techniques, specifically WLD-Reg, is able to mitigate the disparities in class-generalization performance of the model. (ii) We have tackled the deep learning generalization puzzle concerning the pronounced nuance in the impact of overfitting on different classes/categories within the learned task. We close the literature gap by introducing and exploring the concept of “class-generalization error” using information-theoretic tools. These findings contribute insights to address the Research Question 5 providing the first theoretical analysis of this generalization puzzle. We also empirically strengthened the findings with supporting experiments validating the efficiency of the proposed bounds in capturing the behaviour of class-wise generalization.

The developed theoretical tools in Section 3.4.2 are versatile and useful beyond studying class-generalization error. In fact, they allow to study the sub-task problem and providing guarantees in the presence of sensitive attributes, as shown in Section 3.4.2.3. Furthermore, beyond the discussed extra applications in this dissertation, as shown in Publication VI, the developed tools can be used to efficiently capture the behaviour of generalization in terms of recall and specificity and derive label-dependent standard generalization bounds.

It should be noted that, in Section 3.4.2 of this dissertation, we have mainly focused on the CMI setting to derive our bounds for class-generalization error. However, in Publication VI Section 2.1, we show that it is possible to address this problem in MI setting [153, 154] also.



## 4 CONCLUSIONS

In this comprehensive exploration, the dissertation has focused on the critical role of feature diversity within the hidden layers of neural networks, contributing novel insights to the understanding of their generalization capabilities. The dissertation started by addressing the fundamental question of deriving rigorous generalization bounds that highlight the influence of feature diversity on overfitting. The theoretical contributions show feature diversity can help improve the generalization of neural networks.

Moving beyond theory, the dissertation introduced practical methodologies, specifically a family of regularizers, aimed at promoting feature diversity within neural network layers, in multiple contexts. Empirical results substantiated the efficacy of these strategies, demonstrating improved performance through richer representations and reduced overfitting.

The third research question extended the theoretical and empirical findings to the energy-based learning paradigm. EBMs provide a unified framework for various learning tasks, and the dissertation explored the impact of feature diversity on their generalization ability. This extension broadened the applicability of feature diversity strategies beyond standard supervised learning contexts. Extensive experimental results showed that reducing redundancy within the features of EBMs leads to improved performance across multiple tasks.

The fourth research question focused on the phenomenon of class-wise generalization disparities observed in deep learning models. The dissertation analyzed how feature-diversity approaches affect class-wise generalization performance, showing that feature diversity-based regularization can mitigate this effect and reduce the variance of class-generalization errors.

The final research question aimed to develop a comprehensive theory of class-wise generalization. Existing generalization theories often provide holistic views, but fail to capture the nuanced behavior observed in deep learning

models. The dissertation addressed this gap by introducing a rigorous theoretical framework specifically designed to study and understand class-wise generalization, contributing novel insights to deep learning theory. Furthermore, the dissertation showed how the developed theoretical tools are useful beyond studying this phenomena and can be applied to study other tasks.

In summary, this dissertation has provided a holistic examination of feature diversity’s pivotal role in generalization. The main contributions span theoretical advancements, methodological innovations, and empirical validation and insights. The establishment of feature diversity as a key factor in neural network generalization, the development of practical strategies for promoting feature diversity, and the extension of these findings to energy-based learning settings demonstrate the versatility and impact of leveraging feature diversity. The combination of the theoretical understanding and algorithmic development collectively contribute to a deeper understanding of neural networks’ generalization and offer pathways for mitigating overfitting and hence developing more efficient models. The overarching aim of the dissertation was to unravel the dynamics of feature diversity and provide a nuanced perspective on its effect on generalization. The insights gained from this exploration are expected to inspire future research in developing more advanced feature diversity-promoting strategies and regularizers to improve the performance of neural networks in multiple contexts. For example, similar to WLD-Reg Det and Logdet variants, regularizers using global diversity measures can be used in the context of EBM’s and autoencoders. Additionally, the analysis of class-wise generalization disparities and the development of a theoretical framework for studying this phenomenon contribute substantially to the understanding of deep learning generalization. Future works in this line of research include deriving tighter bounds for class-generalization error and developing practical techniques to mitigate this problem in deep learning models.

Besides the limitations discussed within the main body of the dissertation, the scope of the theoretical analysis considered here is restricted to the standard generalization. For a more comprehensive study of feature diversity, future research direction can include for example theoretically studying its effect on adversarial generalization [22, 155]. Intuitively, learning diverse features reduces the dependency on a single pattern to make decisions and hence can improve

the adversarial robustness of the model. Another noteworthy limitation of the methodologies developed within this dissertation is the limited compatibility with flat feature representations. Future work can include studying diversity within other topologies, e.g., convolutional output maps and recurrent representations. Furthermore, as mentioned in Section 3.1.3, future research can also include studying feature diversity using scale-invariant measures such as correlation and mutual information.



# REFERENCES

- [1] M. Mohri, A. Rostamizadeh and A. Talwalkar. *Foundations of machine learning*. MIT press, 2018.
- [2] I. Goodfellow, Y. Bengio and A. Courville. *Deep learning*. MIT press, 2016.
- [3] A. Krizhevsky, I. Sutskever and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* 25 (2012), 1097–1105.
- [4] K. He, X. Zhang, S. Ren and J. Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*. 2016, 770–778.
- [5] D. Bank, N. Koenigstein and R. Giryes. Autoencoders. *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook* (2023), 353–374.
- [6] W. Wang, Y. Huang, Y. Wang and L. Wang. Generalized autoencoder: A neural network framework for dimensionality reduction. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2014, 490–497.
- [7] J. Xie, S.-C. Zhu and Y. Nian Wu. Synthesizing dynamic patterns by spatial-temporal generative convnet. *IEEE Conference on Computer Vision and Pattern Recognition*. 2017, 7093–7101.
- [8] M. Khalifa, H. Elsahar and M. Dymetman. A Distributional Approach to Controlled Text Generation. *International Conference on Learning Representations* (2021).
- [9] J. Zhao, M. Mathieu and Y. LeCun. Energy-based generative adversarial network. *International Conference on Learning Representation* (2017).

- [10] C. Zhang, S. Bengio, M. Hardt, B. Recht and O. Vinyals. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations*. 2017.
- [11] H. Xu and S. Mannor. Robustness and generalization. *Machine learning* 86 (2012), 391–423.
- [12] A. Harutyunyan and G. Neu. Towards a theory of generalization and regularization in deep learning. *arXiv preprint arXiv:2103.02548* (2021).
- [13] R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino and S. Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in Neural Information Processing Systems* 31 (2018).
- [14] P. Kirichenko, R. Balestrieri, M. Ibrahim, S. R. Vedantam, H. Firooz and A. G. Wilson. Understanding the class-specific effects of data augmentations. *ICLR Workshop on Pitfalls of limited data and computation for Trustworthy ML*. 2023.
- [15] T. Steinke and L. Zakynthinou. Reasoning about generalization via conditional mutual information. *Conference on Learning Theory*. PMLR. 2020, 3437–3452.
- [16] Z. Allen-Zhu, Y. Li and Y. Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in Neural Information Processing Systems* 32 (2019).
- [17] Z.-H. Zhou. Why over-parameterization of deep neural networks does not overfit?: *Science China Information Sciences* 64 (2021), 1–3.
- [18] M. Hardt, B. Recht and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. *International Machine Learning*. PMLR. 2016, 1225–1234.
- [19] J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti and D. M. Roy. Information-theoretic generalization bounds for SGLD via data-dependent estimates. *Advances in Neural Information Processing Systems* 32 (2019).
- [20] M. Haghifam, Y. Xie, G. Valiant and P. Valiant. Generalization bounds for large-scale learning. *arXiv preprint arXiv:2106.12889* (2021).



- [21] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* (2002), 463–482.
- [22] D. Yin, R. Kannan and P. Bartlett. Rademacher complexity for adversarially robust generalization. *International Conference on Machine Learning*. PMLR. 2019, 7085–7094.
- [23] L. V. Truong. On rademacher complexity-based generalization bounds for deep learning. *arXiv preprint arXiv:2208.04284* (2022).
- [24] E. D. Sontag et al. VC dimension of neural networks. *NATO ASI Series F Computer and Systems Sciences* 168 (1998), 69–96.
- [25] P. L. Bartlett and W. Maass. Vapnik-Chervonenkis and uniform dimension bounds. *Computational Complexity* 12.1-2 (2003), 140–174.
- [26] Z. Gong, P. Zhong and W. Hu. Diversity in Machine Learning. *IEEE Access* 7 (2019), 64323–64350.
- [27] B. Xie, Y. Liang and L. Song. Diverse neural network learns true target functions. *Artificial Intelligence and Statistics*. 2017, 1216–1224.
- [28] P. Xie, Y. Deng and E. Xing. On the generalization error bounds of neural networks under diversity-inducing mutual angular regularization. *arXiv preprint arXiv:1511.07110* (2015).
- [29] P. Xie, A. Singh and E. P. Xing. Uncorrelation and evenness: a new diversity-promoting regularizer. *International Conference on Machine Learning*. 2017, 3811–3820.
- [30] N. Bansal, X. Chen and Z. Wang. Can we gain more from orthogonality regularizations in training deep cnns?: *Advances in Neural Information Processing Systems* (2018).
- [31] B. O. Ayinde, T. Inanc and J. M. Zurada. Regularizing deep neural networks by enhancing diversity in feature extraction. *IEEE Transactions on Neural Networks and Learning Systems* 30.9 (2019), 2650–2661.
- [32] Y. Yu, Y.-F. Li and Z.-H. Zhou. Diversity regularized machine. *International Joint Conference on Artificial Intelligence*. 2011.

- [33] M. Cogswell, F. Ahmed, R. B. Girshick, L. Zitnick and D. Batra. Reducing Overfitting in Deep Networks by Decorrelating Representations. *International Conference on Learning Representations*. 2016.
- [34] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato and F. Huang. A tutorial on energy-based learning. *Predicting Structured Data 1* (2006).
- [35] Y. Fang and M. Liu. A Unified Energy-based Framework for Learning to Rank. *ACM International Conference on the Theory of Information Retrieval*. 2016.
- [36] F. K. Gustafsson, M. Danelljan and T. B. Schön. Learning Proposals for Practical Energy-Based Regression. *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2022.
- [37] F. K. Gustafsson, M. Danelljan, R. Timofte and T. B. Schön. How to Train Your Energy-Based Model for Regression. *Proceedings of the British Machine Vision Conference (BMVC)*. 2020.
- [38] F. K. Gustafsson, M. Danelljan, G. Bhat and T. B. Schön. Energy-Based Models for Deep Probabilistic Regression. *Proceedings of the European Conference on Computer Vision (ECCV)*. 2020.
- [39] A. Bakhtin, Y. Deng, S. Gross, M. Ott, M. Ranzato and A. Szlam. Residual Energy-Based Models for Text. *Journal of Machine Learning Research* 22.40 (2021), 1–41.
- [40] Y. Zhao, J. Xie and P. Li. Learning energy-based generative models via coarse-to-fine expanding and sampling. *International Conference on Learning Representations*. 2020.
- [41] Y. Xu, J. Xie, T. Zhao, C. Baker, Y. Zhao and Y. N. Wu. Energy-based continuous inverse optimal control. *IEEE Transactions on Neural Networks and Learning Systems* (2022).
- [42] S. C. Zhu and D. Mumford. Grade: Gibbs reaction and diffusion equations. *Sixth International Conference on Computer Vision*. IEEE. 1998, 847–854.

- [43] T. Che, R. Zhang, J. Sohl-Dickstein, H. Larochelle, L. Paull, Y. Cao and Y. Bengio. Your gan is secretly an energy-based model and you should use discriminator driven latent sampling. *arXiv preprint arXiv:2003.06060* (2020).
- [44] Y. Du and I. Mordatch. Implicit Generation and Generalization in Energy-Based Models. *Advances in Neural Information Processing Systems*. 2019.
- [45] Z. Dai, A. Almahairi, P. Bachman, E. Hovy and A. Courville. Calibrating energy-based generative adversarial networks. *arXiv preprint arXiv:1702.01691* (2017).
- [46] Y. Du and I. Mordatch. Implicit Generation and Generalization in Energy-Based Models. *Proceedings of Machine Learning Research* 130 (2021), 2071–2080.
- [47] R. Balestriero, L. Bottou and Y. LeCun. The Effects of Regularization and Data Augmentation are Class Dependent. *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc., 2022, 37878–37891.
- [48] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2015).
- [49] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger. Densely connected convolutional networks. *IEEE Conference on Computer Vision and Pattern Recognition* (2017), 4700–4708.
- [50] S. Zagoruyko and N. Komodakis. Wide Residual Networks. *Proceedings of the British Machine Vision Conference (BMVC)*. 2016.
- [51] S. Xie, R. Girshick, P. Dollár, Z. Tu and K. He. Aggregated residual transformations for deep neural networks. *IEEE Conference on Computer Vision and Pattern Recognition*. 2017, 1492–1500.
- [52] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong and Q. He. A comprehensive survey on transfer learning. *Proceedings of the IEEE* 109 (2020), 43–76.

- [53] B. Neyshabur, H. Sedghi and C. Zhang. What is being transferred in transfer learning?: *Advances in Neural Information Processing Systems* 33 (2020), 512–523.
- [54] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee and F. Makedon. A survey on contrastive self-supervised learning. *Technologies* 9.1 (2020), 2.
- [55] A. Antoniou, A. Storkey and H. Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340* (2017).
- [56] H. Lee, H. Lee and J. Kim. DropMix: Reducing Class Dependency in Mixed Sample Data Augmentation. *arXiv preprint arXiv:2307.09136* (2023).
- [57] Y. N. D. Hongyi Zhang Moustapha Cisse and D. Lopez-Paz. mixup: Beyond Empirical Risk Minimization. *International Conference on Learning Representations* (2018).
- [58] N. Passalis and A. Tefas. Neural bag-of-features learning. *Pattern Recognition* 64 (2017), 277–294.
- [59] N. Passalis and A. Tefas. Learning bag-of-features pooling for deep convolutional neural networks. *IEEE International Conference on Computer Vision*. 2017, 5755–5763.
- [60] F. Laakom, N. Passalis, J. Raitoharju, J. Nikkanen, A. Tefas, A. Iosifidis and M. Gabbouj. Bag of color features for color constancy. *IEEE Transactions on Image Processing* 29 (2020), 7722–7734.
- [61] N. Passalis, J. Raitoharju, A. Tefas and M. Gabbouj. Efficient adaptive inference for deep convolutional neural networks using hierarchical early exits. *Pattern Recognition* 105 (2020), 107346.
- [62] K. Chumachenko, A. Iosifidis and M. Gabbouj. Self-Attention Neural Bag-of-Features. *2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)*. 2022, 1–6.
- [63] N. Passalis and A. Tefas. Learning deep representations with probabilistic knowledge transfer. *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, 268–284.

- [64] D. T. Tran, N. Passalis, A. Tefas, M. Gabbouj and A. Iosifidis. Attention-Based Neural Bag-of-Features Learning for Sequence Data. *IEEE Access* 10 (2022), 45542–45552.
- [65] P. Baldi. Autoencoders, unsupervised learning, and deep architectures. *Proceedings of ICML workshop on Unsupervised and Transfer Learning*. JMLR Workshop and Conference Proceedings. 2012, 37–49.
- [66] F. Zhuang, X. Cheng, P. Luo, S. J. Pan and Q. He. Supervised representation learning: Transfer learning with deep autoencoders. *Twenty-Fourth International Joint Conference on Artificial Intelligence*. 2015.
- [67] J. Lu, N. Verma and N. K. Jha. Convolutional autoencoder-based transfer learning for multi-task image inferences. *IEEE Transactions on Emerging Topics in Computing* 10.2 (2021), 1045–1057.
- [68] F. Laakom, J. Raitoharju, A. Iosifidis, J. Nikkanen and M. Gabbouj. Color constancy convolutional autoencoder. *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. 2019, 1085–1090.
- [69] S. Petscharnig, M. Lux and S. Chatzichristofis. Dimensionality reduction for image features using deep learning and autoencoders. *Proceedings of the 15th international workshop on content-based multimedia indexing*. 2017, 1–6.
- [70] Q. Fournier and D. Aloise. Empirical comparison between autoencoders and traditional dimensionality reduction methods. *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. IEEE. 2019, 211–214.
- [71] A. Creswell and A. A. Bharath. Denoising adversarial autoencoders. *IEEE Transactions on Neural Networks and Learning Systems* 30.4 (2018), 968–984.
- [72] L. Gondara. Medical image denoising using convolutional denoising autoencoders. *IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. IEEE. 2016, 241–246.
- [73] Z. Chen, C. K. Yeo, B. S. Lee and C. T. Lau. Autoencoder-based network anomaly detection. *2018 Wireless telecommunications symposium (WTS)*. IEEE. 2018, 1–5.

- [74] C. Zhou and R. C. Paffenroth. Anomaly detection with robust deep autoencoders. *23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017, 665–674.
- [75] M. Sakurada and T. Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. *Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis*. 2014, 4–11.
- [76] S. Li, Y. Du, G. M. van de Ven, A. Torralba and I. Mordatch. Energy-Based Models for Continual Learning. *arXiv preprint arXiv:2011.12216* (2020).
- [77] G. Pang, C. Shen and A. van den Hengel. Deep anomaly detection with deviation networks. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, 353–362.
- [78] Y. Du, J. Meier, J. Ma, R. Fergus and A. Rives. Energy-based models for atomic-resolution protein conformations. *International Conference on Learning Representations*. 2020.
- [79] R. Boney, J. Kannala and A. Ilin. Regularizing model-based planning with energy-based models. *Conference on Robot Learning*. 2020, 182–191.
- [80] K. Sharma, B. Singh, E. Herman, R. Regine, S. S. Rajest and V. P. Mishra. Maximum information measure policies in reinforcement learning with deep energy-based model. *International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*. 2021, 19–24.
- [81] L. Gan, D. Nurbakova, L. Laporte and S. Calabretto. Enhancing Recommendation Diversity using Determinantal Point Processes on Knowledge Graphs. *Conference on Research and Development in Information Retrieval*. 2020, 2001–2004.
- [82] J. Zbontar, L. Jing, I. Misra, Y. LeCun and S. Deny. Barlow twins: Self-supervised learning via redundancy reduction. *International Conference on Machine Learning* (2021).
- [83] N. Li, Y. Yu and Z.-H. Zhou. Diversity regularized ensemble pruning. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 2012, 330–345.

- [84] M. Derezhinski, D. Calandriello and M. Valko. Exact sampling of determinantal point processes with sublinear time preprocessing. *Advances in Neural Information Processing Systems*. 2019, 11546–11558.
- [85] E. Bıyık, K. Wang, N. Anari and D. Sadigh. Batch active learning using determinantal point processes. *arXiv preprint arXiv:1906.07975* (2019).
- [86] M. Gartrell, V.-E. Brunel, E. Dohmatob and S. Krichene. Learning non-symmetric determinantal point processes. *Advances in Neural Information Processing Systems*. 2019, 6718–6728.
- [87] K. Yang, V. Gkatzelis and J. Stoyanovich. Balanced Ranking with Diversity Constraints. *International Joint Conference on Artificial Intelligence*, 6035–6042.
- [88] Y. Kondo and K. Yamauchi. A dynamic pruning strategy for incremental learning on a budget. *International Conference on Neural Information Processing*. Springer. 2014, 295–303.
- [89] Y. He, P. Liu, Z. Wang, Z. Hu and Y. Yang. Filter pruning via geometric median for deep convolutional neural networks acceleration. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, 4340–4349.
- [90] S. Lee, B. Heo, J.-W. Ha and B. C. Song. Filter Pruning and Re-Initialization via Latent Space Clustering. *IEEE Access* 8 (2020), 189587–189597.
- [91] P. Singh, V. K. Verma, P. Rai and V. Namboodiri. Leveraging filter correlations for deep model compression. *The IEEE Winter Conference on Applications of Computer Vision*. 2020, 835–844.
- [92] Y. Bao, H. Jiang, L. Dai and C. Liu. Incoherent training of deep neural networks to de-correlate bottleneck features for speech recognition. *International Conference on Acoustics, Speech and Signal Processing*. 2013, 6980–6984.
- [93] P. Xie, Y. Deng and E. Xing. Diversifying restricted boltzmann machine for document modeling. *International Conference on Knowledge Discovery and Data Mining*. 2015, 1315–1324.

- [94] P. Xie, J. Zhu and E. Xing. Diversity-promoting bayesian learning of latent variable models. *International Conference on Machine Learning*. 2016, 59–68.
- [95] J. Malkin and J. Bilmes. Multi-layer ratio semi-definite classifiers. *International Conference on Acoustics, Speech and Signal Processing*. 2009, 4465–4468.
- [96] J. Malkin and J. Bilmes. Ratio semi-definite classifiers. *International Conference on Acoustics, Speech and Signal Processing*. 2008, 4113–4116.
- [97] A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083* (2012).
- [98] J. T. Kwok and R. P. Adams. Priors for diversity in generative latent variable models. *Advances in Neural Information Processing Systems*. 2012, 2996–3004.
- [99] A. R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning* (1994), 115–133.
- [100] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory* (1993), 930–945.
- [101] K. Zhai and H. Wang. Adaptive dropout with rademacher complexity regularization. *International Conference on Learning Representations*. 2018.
- [102] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
- [103] G. K. Dziugaite and D. M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008* (2017).
- [104] Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan and S. Bengio. Fantastic generalization measures and where to find them. *International Conference on Learning Representations* (2019).
- [105] B. Neyshabur, S. Bhojanapalli, D. McAllester and N. Srebro. Exploring generalization in deep learning. *Advances in Neural Information Processing Systems* 30 (2017).



- [106] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86.11 (1998), 2278–2324.
- [107] B. Schölkopf. The kernel trick for distances. *Advances in Neural Information Processing Systems* 13 (2000).
- [108] A. Kulesza and B. Taskar. Structured determinantal point processes. *Advances in Neural Information Processing Systems*. 2010, 1171–1179.
- [109] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master’s thesis, Department of Computer Science, University of Toronto* (2009).
- [110] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115.3 (2015), 211–252.
- [111] G. Algan and I. Ulusoy. Label noise types and their effects on deep learning. *arXiv preprint arXiv:2003.10471* (2020).
- [112] C. Dong, L. Liu and J. Shang. Label Noise in Adversarial Training: A Novel Perspective to Study Robust Overfitting. *Advances in Neural Information Processing Systems* 35 (2022), 17556–17567.
- [113] I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic and A. Dosovitskiy. MLP-Mixer: An all-MLP Architecture for Vision. *arXiv preprint arXiv:2105.01601* (2021).
- [114] H. Liu, Z. Dai, D. So and Q. V. Le. Pay attention to mlps. *Advances in Neural Information Processing Systems* 34 (2021), 9204–9215.
- [115] J. Lee-Thorp, J. Ainslie, I. Eckstein and S. Ontanon. FNet: Mixing Tokens with Fourier Transforms. *arXiv preprint arXiv:2105.03824* (2021).
- [116] F. Laakom, F. Sohrab, J. Raitoharju, A. Iosifidis and M. Gabbouj. Convolutional autoencoder-based multimodal one-class classification. *arXiv preprint arXiv:2309.14090* (2023).

- [117] S. S. Khan and M. G. Madden. One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review* 29.3 (2014), 345–374.
- [118] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller and M. Kloft. Deep one-class classification. *International Conference on Machine Learning*. PMLR. 2018, 4393–4402.
- [119] F. Sohrab, J. Raitoharju, M. Gabbouj and A. Iosifidis. Subspace support vector data description. *24th International Conference on Pattern Recognition (ICPR)*. IEEE. 2018, 722–727.
- [120] F. Sohrab and J. Raitoharju. Boosting rare benthic macroinvertebrates taxa identification with one-class classification. *IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. 2020, 928–933.
- [121] I. Guyon. *Madelon*. UCI Machine Learning Repository. 2008.
- [122] R. Cole and M. Fanty. *ISOLET*. UCI Machine Learning Repository. 1994.
- [123] R. Lathrop. *p53 Mutants*. UCI Machine Learning Repository. 2010.
- [124] X. Zhang. Pac-learning for energy-based models. PhD thesis. Citeseer, 2013.
- [125] R. Kumar, S. Ozair, A. Goyal, A. Courville and Y. Bengio. Maximum entropy generators for energy-based models. *arXiv preprint arXiv:1901.08508* (2019).
- [126] A. Bardes, J. Ponce and Y. LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. *arXiv preprint arXiv:2105.04906* (2021).
- [127] J. Li, Y. Liu, R. Yin and W. Wang. Multi-Class Learning using Unlabeled Samples: Theory and Algorithm. *International Joint Conference on Artificial Intelligence*. 2019.
- [128] K. Kawaguchi, L. P. Kaelbling and Y. Bengio. Generalization in deep learning. *arXiv preprint arXiv:1710.05468* (2017).
- [129] A. Brando, J. A. Rodriguez, J. Vitria and A. Rubio Muñoz. Modelling heterogeneous distributions with an uncountable mixture of asymmetric laplacians. *Advances in Neural Information Processing Systems* 32 (2019).

- [130] A. Hyvärinen and P. Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research* 6.4 (2005).
- [131] M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. *International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings. 2010, 297–304.
- [132] Z. Zhang, Y. Song and H. Qi. Age progression/regression by conditional adversarial autoencoder. *IEEE Conference on Computer Vision and Pattern Recognition*. 2017, 5810–5818.
- [133] *Deep Energy-Based Generative Models*. [https://uvadlc-notebooks.readthedocs.io/en/latest/tutorial\\_notebooks/tutorial8/Deep\\_Energy\\_Models.html](https://uvadlc-notebooks.readthedocs.io/en/latest/tutorial_notebooks/tutorial8/Deep_Energy_Models.html). Accessed: 2022-05-01.
- [134] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems* 30 (2017).
- [135] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan and S. Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks* (2019).
- [136] Z. Li and D. Hoiem. Learning without forgetting. *IEEE TPAMI* (2017).
- [137] T. Shibata, G. Irie, D. Ikami and Y. Mitsuzumi. Learning with Selective Forgetting. *International Joint Conference on Artificial Intelligence*. 2021.
- [138] R. Zhou, C. Tian and T. Liu. Individually conditional individual mutual information bound on generalization error. *IEEE Transactions on Information Theory* 68.5 (2022), 3304–3316.
- [139] Z. Wang and Y. Mao. Tighter Information-Theoretic Generalization Bounds from Supersamples. *arXiv preprint arXiv:2302.02432* (2023).
- [140] W. Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding* (1994), 409–426.
- [141] C. Zhang, S. Bengio, M. Hardt, B. Recht and O. Vinyals. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations*. 2016.

- [142] D. Arpit, S. Jastrzkebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio et al. A closer look at memorization in deep networks. *International Conference on Machine Learning*. PMLR. 2017, 233–242.
- [143] S. Chatterjee. Learning and memorization. *International Conference on Machine Learning*. PMLR. 2018, 755–763.
- [144] H. Harutyunyan, M. Raginsky, G. Ver Steeg and A. Galstyan. Information-theoretic generalization bounds for black-box learning algorithms. *Advances in Neural Information Processing Systems* 34 (2021), 24670–24682.
- [145] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*. 2009, 248–255.
- [146] X. Wu, J. H. Manton, U. Aickelin and J. Zhu. Information-theoretic analysis for transfer learning. *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2020, 2819–2824.
- [147] Z. Wang and Y. Mao. Information-Theoretic Analysis of Unsupervised Domain Adaptation. *arXiv preprint arXiv:2210.00706* (2022).
- [148] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman and A. Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54.6 (2021), 1–35.
- [149] S. Barocas, M. Hardt and A. Narayanan. Fairness in machine learning. *Nips tutorial 1* (2017), 2017.
- [150] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik and H. Wallach. Improving fairness in machine learning systems: What do industry practitioners need?: *CHI Conference on Human Factors in Computing Systems*. 2019, 1–16.
- [151] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado and M. H. Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine* 169.12 (2018), 866–872.
- [152] R. Williamson and A. Menon. Fairness risk measures. *International Conference on Machine Learning*. PMLR. 2019, 6786–6797.

- [153] H. Xu and M. Raginsky. Information-theoretic tools for the analysis of learning algorithms. *IEEE Transactions on Information Theory* 63.7 (2017), 4217–4233.
- [154] Y. Bu, S. Zou and V. V. Veeravalli. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory* 1.1 (2020), 121–130.
- [155] P. L. Bartlett, A. Gupta, J. Ma and H. Xu. Nearly-tight bounds for adversarial robustness via rademacher complexity. *arXiv preprint arXiv:1906.03220* (2019).



# PUBLICATIONS





# PUBLICATION

I

## **Learning distinct features helps, provably**

F. Laakom, J. Raitoharju, A. Iosifidis and M. Gabbouj


*Joint European Conference on Machine Learning and Knowledge Discovery in  
Databases2023*, 206–222

DOI: 10.1007/978-3-031-43415-0\_13

© 2023 . Reproduced with permission from Springer Nature



# Learning distinct features helps, provably

Firas Laakom<sup>1</sup>[0000–0001–7436–5692] , Jenni Raitoharju<sup>2,3</sup>[0000–0003–4631–9298],  
Alexandros Iosifidis<sup>4</sup>[0000–0003–4807–1345], and  
Moncef Gabbouj<sup>1</sup>[0000–0002–9788–2323]

<sup>1</sup> Faculty of Information Technology and Communication Sciences  
Tampere University, Finland.

<sup>2</sup> Faculty of Information Technology, University of Jyväskylä, Finland

<sup>3</sup> Programme for Environmental Information, Finnish Environment Institute,  
Jyväskylä, Finland

<sup>4</sup> Department of Electrical and Computer Engineering  
Aarhus University, Denmark

**Abstract.** We study the diversity of the features learned by a two-layer neural network trained with the least squares loss. We measure the diversity by the average  $L_2$ -distance between the hidden-layer features and theoretically investigate how learning non-redundant distinct features affects the performance of the network. To do so, we derive novel generalization bounds depending on feature diversity based on Rademacher complexity for such networks. Our analysis proves that more distinct features at the network’s units within the hidden layer lead to better generalization. We also show how to extend our results to deeper networks and different losses.

**Keywords:** Neural Networks · Generalization Theory · Feature Diversity

## 1 Introduction

Neural networks are a powerful class of non-linear function approximators that have been successfully used to tackle a wide range of problems. They have enabled breakthroughs in many tasks, such as image classification [31], speech recognition [20], and anomaly detection [16]. However, neural networks are often over-parameterized, i.e., have more parameters than the data they are trained on. As a result, they tend to overfit to the training samples and not generalize well on unseen examples [18]. Avoiding overfitting has been extensively studied [14, 15, 43, 45, 47] and various approaches and strategies have been proposed, such as data augmentation [18, 64], regularization [1, 8, 32], and Dropout [21, 38, 39], to close the gap between the empirical loss and the expected loss.

Formally, the output of a neural network consisting of  $P$  layers can be defined as follows:

$$f(\mathbf{x}; \mathbf{W}) = \rho^P(\mathbf{W}^P(\rho^{P-1}(\cdots \rho^2(\mathbf{W}^2 \rho^1(\mathbf{W}^1 \mathbf{x})))), \quad (1)$$

where  $\rho^i(\cdot)$  is the element-wise activation function, e.g., *ReLU* or *Sigmoid*, of the  $i^{th}$  layer and  $\mathbf{W} = \{\mathbf{W}^1, \dots, \mathbf{W}^P\}$  are the weights of the network with the

superscript denoting the layer. By defining  $\Phi(\cdot) = \rho^{P-1}(\cdots \rho^2(\mathbf{W}^2 \rho^1(\mathbf{W}^1 \cdot)))$ , the output of neural network becomes

$$f(\mathbf{x}; \mathbf{W}) = \rho^P(\mathbf{W}^P \Phi(\mathbf{x})), \quad (2)$$

where  $\Phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \dots, \phi_M(\mathbf{x})]$  is the  $M$ -dimensional feature representation of the input  $\mathbf{x}$ . This way neural networks can be interpreted as a two-stage process, with the first stage being representation learning, i.e., learning  $\Phi(\cdot)$ , followed by the final prediction layer. Both parts are jointly optimized.

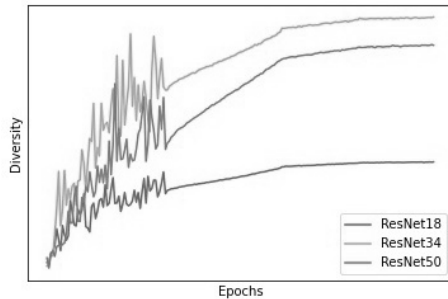
Learning a rich and diverse set of features, i.e., the first stage, is critical for achieving top performance [3, 10, 34]. Studying the different properties of the learned features is an active field of research [11, 13, 29]. For example, [13] showed theoretically that learning a good feature representation can be helpful in few-shot learning. In this paper, we focus on the diversity of the features. This property has been empirically studied in [10, 36, 35] and has been shown to boost performance and reduce overfitting. However, no theoretical guarantees are provided. In this paper, we close this gap and we conduct a theoretical analysis of feature diversity. In particular, we propose to quantify the diversity of the feature set  $\{\phi_1(\cdot), \dots, \phi_M(\cdot)\}$  using the average pairwise  $L_2$ -distance between their outputs. Formally, given a dataset  $\{\mathbf{x}_i\}_{i=1}^N$ , we have

$$diversity = \frac{1}{N} \sum_{k=1}^N \frac{1}{2M(M-1)} \sum_{i \neq j}^M (\phi_i(\mathbf{x}_k) - \phi_j(\mathbf{x}_k))^2. \quad (3)$$

Intuitively, *diversity* measures how distinct the learned features are. If the mappings learned by two different units are redundant, then, given the same input, both units would yield similar output. This yields in low  $L_2$ -distance and as a result a low diversity. In contrast, if the mapping learned by each unit is distinct, the corresponding average distances to the outputs of the other units within the layer are high. Thus, this yields a high global diversity.

To confirm this intuition and further motivate the analysis of this attribute, we conduct empirical simulations. We track the diversity of the representation of the last hidden layer, as defined in equation 3, during the training of three different ResNet [19] models on CIFAR10 [30]. The results are reported in Figure 1. Indeed, diversity consistently increases during the training for all the models. This shows that, in order to solve the task at hand, neural networks learn distinct features.

**Our contributions:** In this paper, we theoretically investigate diversity in the neural network context and study how learning non-redundant features affects the performance of the model. We derive a bound for the generalization gap which is inversely proportional to the proposed diversity measure showing that learning distinct features helps. In our analysis, we focus on the simple neural network model with one-hidden layer trained with mean squared error. This configuration is simple, however, it has been shown to be convenient and insightful for the theoretical analysis [9, 12, 13]. Moreover, we show how to extend our theoretical analysis to different losses and different network architectures.



**Fig. 1.** Preliminary empirical results for additional motivation to theoretically understand feature diversity. The figure shows diversity versus the number of epochs for three different ResNet models trained on CIFAR10 dataset.

Our contributions can be summarized as follows:

- We analyze the effect the feature diversity on the generalization error bound of a neural network. The analysis is presented in Section 3. In Theorem 1, we derive an upper bound for the generalization gap which is inversely proportional to the diversity factor. Thus, we provide theoretical evidence that learning distinct features can help reduce the generalization error.
- We extend our analysis to different losses and general multi-layer networks. These results are presented in Theorems 2, 3, 4, 5, and 6.

**Outline of the paper:** The rest of the paper is organized as follows: Section 2 summarizes the preliminaries for our analysis. Section 3 presents our main theoretical results along with the proofs. Section 4 extends our results for different settings. Section 5 concludes the work with a discussion and several open problems.

## 2 PRELIMINARIES

Generalization theory [50, 28] focuses on the relation between the empirical loss defined as

$$\hat{L}(f) = \frac{1}{N} \sum_{i=1}^N l(f(\mathbf{x}_i; \mathbf{W}), y_i), \quad (4)$$

and the expected risk, for any  $f$  in the hypothesis class  $\mathcal{F}$ , defined as

$$L(f) = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{Q}}[l(f(\mathbf{x}), y)], \quad (5)$$

where  $\mathcal{Q}$  is the underlying distribution of the dataset and  $y_i$  the corresponding label of  $x_i$ . Let  $f^* = \arg \min_{f \in \mathcal{F}} L(f)$  be the expected risk minimizer and  $\hat{f} = \arg \min_{f \in \mathcal{F}} \hat{L}(f)$  be the empirical risk minimizer. We are interested in the estimation error, i.e.,  $L(f^*) - L(\hat{f})$ , defined as the gap in the loss between both

minimizers [6]. The estimation error represents how well an algorithm can learn. It usually depends on the complexity of the hypothesis class and the number of training samples [5, 63].

Several techniques have been proposed to quantify the generalization error, such as Probably Approximately Correctly (PAC) learning [50, 53], VC dimension [52], and the Rademacher complexity [50]. The Rademacher complexity has been widely used as it usually leads to a tighter generalization error bound than the other metrics [17, 45, 51]. The formal definition of the empirical Rademacher complexity is given as follows:

**Definition 1.** [7, 50] For a given dataset with  $N$  samples  $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$  generated by a distribution  $\mathcal{Q}$  and for a model space  $\mathcal{F} : \mathcal{X} \rightarrow \mathbb{R}$  with a single dimensional output, the empirical Rademacher complexity  $\mathcal{R}_N(\mathcal{F})$  of the set  $\mathcal{F}$  is defined as follows:

$$\mathcal{R}_N(\mathcal{F}) = \mathbb{E}_{\sigma} \left[ \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^N \sigma_i f(\mathbf{x}_i) \right], \quad (6)$$

where the variables  $\sigma = \{\sigma_1, \dots, \sigma_N\}$  are independent uniform random variables in  $\{-1, 1\}$ .

In this work, we rely on the Rademacher complexity to study diversity. We recall the following three lemmas related to the Rademacher complexity and the generalization error:

**Lemma 1.** [7] For  $\mathcal{F} \in \mathbb{R}^{\mathcal{X}}$ , assume that  $g : \mathbb{R} \rightarrow \mathbb{R}$  is a  $L_g$ -Lipschitz continuous function and  $\mathcal{A} = \{g \circ f : f \in \mathcal{F}\}$ . Then we have

$$\mathcal{R}_N(\mathcal{A}) \leq L_g \mathcal{R}_N(\mathcal{F}). \quad (7)$$

**Lemma 2.** [58] The Rademacher complexity  $\mathcal{R}_N(\mathcal{F})$  of the hypothesis class  $\mathcal{F} = \{f | f(\mathbf{x}) = \sum_{m=1}^M v_m \phi_m(\mathbf{x}) = \sum_{m=1}^M v_m \phi(\mathbf{w}_m^T \mathbf{x})\}$  can be upper-bounded as follows:

$$\mathcal{R}_N(\mathcal{F}) \leq \frac{2L_{\rho} C_{134} M}{\sqrt{N}} + \frac{C_4 |\phi(0)| M}{\sqrt{N}}, \quad (8)$$

where  $C_{134} = C_1 C_3 C_4$  and  $\phi(0)$  is the output of the activation function at the origin.

**Lemma 3.** [7] With a probability of at least  $1 - \delta$ ,

$$L(\hat{f}) - L(f^*) \leq 4\mathcal{R}_N(\mathcal{A}) + B \sqrt{\frac{2 \log(2/\delta)}{N}}, \quad (9)$$

where  $B \geq \sup_{\mathbf{x}, y} |l(f(\mathbf{x}), y)|$  and  $\mathcal{R}_N(\mathcal{A})$  is the Rademacher complexity of the loss set  $\mathcal{A}$ .

Lemma 3 upper-bounds the generalization error using the Rademacher complexity defined over the loss set and  $\sup_{x,y,f} |l(f(x), y)|$ . Our analysis aims at expressing this bound in terms of diversity, in order to understand how it affects the generalization.

In order to study the effect of diversity on the generalization, given a layer with  $M$  units  $\{\phi_1(\cdot), \dots, \phi_M(\cdot)\}$ , we make the following assumption:

**Assumption 1** *Given any input  $\mathbf{x}$ , we have*

$$\frac{1}{2M(M-1)} \sum_{i \neq j}^M (\phi_i(\mathbf{x}) - \phi_j(\mathbf{x}))^2 \geq d_{min}^2. \quad (10)$$

$d_{min}$  lower-bounds the average  $L_2$ -distance between the different units' activations within the same representation layer. Intuitively, if several neuron pairs  $i$  and  $j$  have similar outputs, the corresponding  $L_2$  distance is small. Thus, the lower bound  $d_{min}$  is also small and the units within this layer are considered redundant and "not diverse". Otherwise, if the average distance between the different pairs is large, their corresponding  $d_{min}$  is large and they are considered "diverse". By studying how the lower bound  $d_{min}$  affects the generalization of the model, we can analyze how the diversity theoretically affects the performance of neural networks. In the rest of the paper, we derive generalization bounds for neural networks using  $d_{min}$ .

### 3 Learning distinct features helps

In this section, we derive generalization bounds for neural networks depending on their diversity. Here, we consider a simple tow-layer neural network with a hidden layer composed of  $M$  neurons and one-dimensional output trained for a regression task. The full characterization of the setup can be summarized as follows:

- The activation function of the hidden layer,  $\rho(\cdot)$ , is a positive  $L_\rho$ -Lipschitz continuous function.
- The input vector  $\mathbf{x} \in \mathbb{R}^D$  satisfies  $\|\mathbf{x}\|_2 \leq C_1$  and the output scalar  $y \in \mathbb{R}$  satisfies  $|y| \leq C_2$ .
- The weight matrix  $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M] \in \mathcal{R}^{D \times M}$  connecting the input to the hidden layer satisfies  $\|\mathbf{w}_m\|_2 \leq C_3$ .
- The weight vector  $\mathbf{v} \in \mathbb{R}^M$  connecting the hidden-layer to the output satisfies  $\|\mathbf{v}\|_\infty \leq C_4$ .
- The hypothesis class is  $\mathcal{F} = \left\{ f | f(\mathbf{x}) = \sum_{m=1}^M v_m \phi_m(\mathbf{x}) = \sum_{m=1}^M v_m \rho(\mathbf{w}_m^T \mathbf{x}) \right\}$ .
- Loss function set is  $\mathcal{A} = \left\{ l | l(f(\mathbf{x}), y) = \frac{1}{2} |f(\mathbf{x}) - y|^2 \right\}$ .
- Given an input  $\mathbf{x}$ ,  $\frac{1}{2M(M-1)} \sum_{n \neq m}^M (\phi_n(\mathbf{x}) - \phi_m(\mathbf{x}))^2 \geq d_{min}^2$ .

Our main goal is to analyze the generalization error bound of the neural network and to see how its upper-bound is linked to the diversity of the different

units, expressed by  $d_{min}$ . The main result of the paper is presented in Theorem 1. Our proof consists of three steps: At first, we derive a novel bound for the hypothesis class  $\mathcal{F}$  depending on  $d_{min}$ . Then, we use this bound to derive bounds for the loss class  $\mathcal{A}$  and its Rademacher complexity  $\mathcal{R}_N(\mathcal{A})$ . Finally, we plug all the derived bounds in Lemma 3 to complete the proof of Theorem 1.

The first step of our analysis is presented in Lemma 4:

**Lemma 4.** *We have*

$$\sup_{\mathbf{x}, f \in \mathcal{F}} |f(\mathbf{x})| \leq \sqrt{\mathcal{J}}, \quad (11)$$

where  $\mathcal{J} = C_4^2(MC_5^2 + M(M-1)(C_5^2 - d_{min}^2))$  and  $C_5 = L_\rho C_1 C_3 + \phi(0)$ ,

*Proof.*

$$\begin{aligned} f^2(\mathbf{x}) &= \left( \sum_{m=1}^M v_m \phi_m(\mathbf{x}) \right)^2 \leq \left( \sum_{m=1}^M \|v\|_\infty \phi_m(\mathbf{x}) \right)^2 = \|v\|_\infty^2 \left( \sum_{m=1}^M \phi_m(\mathbf{x}) \right)^2 \\ &\leq C_4^2 \left( \sum_{m=1}^M \phi_m(\mathbf{x}) \right)^2 = C_4^2 \left( \sum_{m,n} \phi_m(\mathbf{x}) \phi_n(\mathbf{x}) \right) \\ &= C_4^2 \left( \sum_m \phi_m(\mathbf{x})^2 + \sum_{m \neq n} \phi_m(\mathbf{x}) \phi_n(\mathbf{x}) \right). \end{aligned} \quad (12)$$

We have  $\sup_{w, \mathbf{x}} \phi_m(\mathbf{x}) = \sup_{w, \mathbf{x}} \rho(\mathbf{w}^T \mathbf{x}) \leq \sup(L_\rho |\mathbf{w}^T \mathbf{x}| + \phi(0))$ , because  $\rho$  is  $L_\rho$ -Lipschitz. Thus,  $\|\phi\|_\infty \leq L_\rho C_1 C_3 + \phi(0) = C_5$ . For the first term in equation 12, we have  $\sum_m \phi_m(\mathbf{x})^2 < M(L_\rho C_1 C_3 + \phi(0))^2 = MC_5^2$ . The second term, using the identity  $\phi_m(\mathbf{x}) \phi_n(\mathbf{x}) = \frac{1}{2} (\phi_m(\mathbf{x})^2 + \phi_n(\mathbf{x})^2 - (\phi_m(\mathbf{x}) - \phi_n(\mathbf{x}))^2)$ , can be rewritten as

$$\sum_{m \neq n} \phi_m(\mathbf{x}) \phi_n(\mathbf{x}) = \frac{1}{2} \left( \sum_{m \neq n} \phi_m(\mathbf{x})^2 + \phi_n(\mathbf{x})^2 - (\phi_m(\mathbf{x}) - \phi_n(\mathbf{x}))^2 \right). \quad (13)$$

In addition, we have  $\frac{1}{2} \sum_{m \neq n} (\phi_m(\mathbf{x}) - \phi_n(\mathbf{x}))^2 \geq M(M-1)d_{min}^2$ . Thus, we have:

$$\sum_{m \neq n} \phi_m(\mathbf{x}) \phi_n(\mathbf{x}) \leq \frac{1}{2} \sum_{m \neq n} (2C_5^2 - M(M-1)d_{min}^2) = M(M-1)(C_5^2 - d_{min}^2). \quad (14)$$

By putting everything back to equation 12, we have:

$$f^2(\mathbf{x}) \leq C_4^2 (MC_5^2 + M(M-1)(C_5^2 - d_{min}^2)) = \mathcal{J}. \quad (15)$$

Thus,  $\sup_{\mathbf{x}, f} |f(\mathbf{x})| \leq \sqrt{\sup_{\mathbf{x}, f} f(\mathbf{x})^2} \leq \sqrt{\mathcal{J}}$ .

Note that in Lemma 4, we have expressed the upper-bound of  $\sup_{\mathbf{x}, f} |f(\mathbf{x})|$  in terms of  $d_{min}$ . Using this bound, we can now find an upper-bound for  $\sup_{\mathbf{x}, f, y} |l(f(\mathbf{x}), y)|$  in the following lemma:



**Lemma 5.** *We have*

$$\sup_{\mathbf{x}, y, f} |l(f(\mathbf{x}), y)| \leq \frac{1}{2}(\sqrt{\mathcal{J}} + C_2)^2. \quad (16)$$

*Proof.* We have  $\sup_{\mathbf{x}, y, f} |f(\mathbf{x}) - y| \leq \sup_{\mathbf{x}, y, f} (|f(\mathbf{x})| + |y|) = \sqrt{\mathcal{J}} + C_2$ . Thus,  $\sup_{\mathbf{x}, y, f} |l(f(\mathbf{x}), y)| \leq \frac{1}{2}(\sqrt{\mathcal{J}} + C_2)^2$ .

Next, using the result of lemmas 1, 2, and 5, we can derive a bound for the Rademacher complexity of  $\mathcal{A}$ . We have, thus, expressed all the elements of Lemma 3 using the diversity term  $d_{min}$ . By plugging in the derived bounds in Lemmas 4, 5, we obtain Theorem 1.

**Theorem 1.** *With probability at least  $(1 - \delta)$ , we have*

$$L(\hat{f}) - L(f^*) \leq \left(\sqrt{\mathcal{J}} + C_2\right) \frac{A}{\sqrt{N}} + \frac{1}{2}(\sqrt{\mathcal{J}} + C_2)^2 \sqrt{\frac{2 \log(2/\delta)}{N}}, \quad (17)$$

where  $C_{134} = C_1 C_3 C_4$ ,  $\mathcal{J} = C_4^2 (M C_5^2 + M(M-1)(C_5^2 - d_{min}^2))$ ,  $A = 4 \left( 2L_\rho C_{134} + C_4 |\phi(0)| \right) M$ , and  $C_5 = L_\rho C_1 C_3 + \phi(0)$ .

*Proof.* Given that  $l(\cdot)$  is  $K$ -Lipschitz with a constant  $K = \sup_{\mathbf{x}, y, f} |f(\mathbf{x}) - y| \leq \sqrt{\mathcal{J}} + C_2$ , and using Lemma 1, we can show that  $\mathcal{R}_N(\mathcal{A}) \leq K \mathcal{R}_N(\mathcal{F}) \leq (\sqrt{\mathcal{J}} + C_2) \mathcal{R}_N(\mathcal{F})$ . For  $\mathcal{R}_N(\mathcal{F})$ , we use the bound found in Lemma 2. Using Lemmas 3 and 5, we have

$$L(\hat{f}) - L(f^*) \leq 4 \left( \sqrt{\mathcal{J}} + C_2 \right) \left( 2L_\rho C_{134} + C_4 |\phi(0)| \right) \frac{M}{\sqrt{N}} + \frac{1}{2} (\sqrt{\mathcal{J}} + C_2)^2 \sqrt{\frac{2 \log(2/\delta)}{N}}, \quad (18)$$

where  $C_{134} = C_1 C_3 C_4$ ,  $\mathcal{J} = C_4^2 (M C_5^2 + M(M-1)(C_5^2 - d_{min}^2))$ , and  $C_5 = L_\rho C_1 C_3 + \phi(0)$ . Thus, setting  $A = 4 \left( 2L_\rho C_{134} + C_4 |\phi(0)| \right) M$  completes the proof.

Theorem 1 provides an upper-bound for the generalization gap. We note that it is a decreasing function of  $d_{min}$ . Thus, this suggests that higher  $d_{min}$ , i.e., more diverse activations, yields a lower generalization error bound. This shows that learning distinct features helps in neural network context.

We note that the bound in Theorem 1 is non-vacuous in the sense that it converges to zero when the number of training samples  $N$  goes to infinity. Moreover, we note that in this paper we do not claim to reach a tighter generalization bound for neural networks in general [14, 24, 44, 48]. Our main claim is that we derive a generalization bound which depends on the diversity of learned features, as measured by  $d_{min}$ . To the best of our knowledge, this is the first work that performs such theoretical analysis based on the average  $L_2$ -distance between the units within the hidden layer.

## Connection to prior studies

Theoretical analysis of the properties of the features learned by neural network models is an active field of research. Feature representation has been theoretically studied in the context of few-shot learning in [13], where the advantage of learning a good representation in the case of scarce data was demonstrated. [2] showed the same in the context of imitation learning, demonstrating that it has sample complexity benefits for imitation learning. [55] developed similar findings for the self-supervised learning task. [42] derived novel bounds showing the statistical benefits of multitask representation learning in linear Markov Decision Processes. Opposite to the aforementioned works, the main focus of this paper is not on the large sample complexity problems. Instead, we focused on feature diversity in the learned representation and showed that learning distinct features leads to better generalization.

Another line of research related to our work is weight-diversity in neural networks [4, 33, 57, 58, 61]. Diversity in this context is defined based on dissimilarity between the weight component using, e.g., cosine distance and weight matrix covariance [59]. In [58], theoretical benefits of weight-diversity have been demonstrated. We note that, in our work, diversity is defined in a fundamentally different way. We do not consider dissimilarity between the parameters of the neural network. Our main scope is the feature representation and, to this end, diversity is defined based on the  $L_2$  distance between the feature maps directly and not the weights. Empirical analysis of the deep representation of neural networks has drawn attention lately [10, 11, 29, 36]. For example, [10, 36] showed empirically that learning decorrelated features reduces overfitting. However, theoretical understanding of the phenomena is lacking. Here, we close this gap by studying how feature diversity affects generalization.

## 4 Extensions

In this section, we show how to extend our theoretical analysis for classification, for general multi-layer networks, and for different losses.

### 4.1 Binary classification

Here, we extend our analysis of the effect of learning a diverse feature representation on the generalization error to the case of a binary classification task, i.e.,  $y \in \{-1, 1\}$ . Here, we consider the special cases of a hinge loss and a logistic loss. To derive diversity-dependent generalization bounds for these cases, similar to the proofs of Lemmas 7 and 8 in [58], we can show the following two lemmas:

**Lemma 6.** *Using the hinge loss, we have with probability at least  $(1 - \delta)$*

$$L(\hat{f}) - L(f^*) \leq 4 \left( 2L_\rho C_{134} + C_4 |\phi(0)| \right) \frac{M}{\sqrt{N}} + (1 + \sqrt{\mathcal{J}}) \sqrt{\frac{2 \log(2/\delta)}{N}}, \quad (19)$$

where  $C_{134} = C_1 C_3 C_4$ ,  $\mathcal{J} = C_4^2(MC_5^2 + M(M-1)(C_5^2 - d_{min}^2))$ , and  $C_5 = L_\rho C_1 C_3 + \phi(0)$ .

**Lemma 7.** *Using the logistic loss  $l(f(x), y) = \log(1 + e^{-yf(x)})$ , we have with probability at least  $(1 - \delta)$*

$$L(\hat{f}) - L(f^*) \leq \frac{4}{1 + e^{\sqrt{-\mathcal{J}}}} \left( 2L_\rho C_{134} + C_4 |\phi(0)| \right) \frac{M}{\sqrt{N}} + \log(1 + e^{\sqrt{\mathcal{J}}}) \sqrt{\frac{2 \log(2/\delta)}{N}}, \quad (20)$$

where  $C_{134} = C_1 C_3 C_4$ ,  $\mathcal{J} = C_4^2(MC_5^2 + M(M-1)(C_5^2 - d_{min}^2))$ , and  $C_5 = L_\rho C_1 C_3 + \phi(0)$ .

Using the above lemmas, we can now derive a diversity-dependant bound for the binary classification case. The extensions of Theorem 1 in the cases of a hinge loss and a logistic loss are presented in Theorems 2 and 3, respectively.

**Theorem 2.** *Using the hinge loss, with probability at least  $(1 - \delta)$ , we have*

$$L(\hat{f}) - L(f^*) \leq A/\sqrt{N} + (1 + \sqrt{\mathcal{J}}) \sqrt{\frac{2 \log(2/\delta)}{N}}, \quad (21)$$

where  $\mathcal{J} = C_4^2(MC_5^2 + M(M-1)(C_5^2 - d_{min}^2))$ ,  $A = 4 \left( 2L_\rho C_{134} + C_4 |\phi(0)| \right) M$ , and  $C_5 = L_\rho C_1 C_3 + \phi(0)$ .

**Theorem 3.** *Using the logistic loss  $l(f(x), y) = \log(1 + e^{-yf(x)})$ , with probability at least  $(1 - \delta)$ , we have*

$$L(\hat{f}) - L(f^*) \leq \frac{A}{(1 + e^{\sqrt{-\mathcal{J}}})\sqrt{N}} + \log(1 + e^{\sqrt{\mathcal{J}}}) \sqrt{\frac{2 \log(2/\delta)}{N}}, \quad (22)$$

where  $\mathcal{J} = C_4^2(MC_5^2 + M(M-1)(C_5^2 - d_{min}^2))$ ,  $A = 4 \left( 2L_\rho C_{134} + C_4 |\phi(0)| \right) M$ , and  $C_5 = L_\rho C_1 C_3 + \phi(0)$ .

As we can see, also for the binary classification task, the generalization bounds for the hinge and logistic losses are decreasing with respect to  $d_{min}$ . Thus, this shows that learning distinct features helps and can improve the generalization also in binary classification.

## 4.2 Multi-layer networks

Here, we extend our result for networks with  $P (> 1)$  hidden layers. We assume that the pair-wise distances between the activations within layer  $p$  are lower-bounded by  $d_{min}^{(p)}$ . In this case, the hypothesis class can be defined recursively. In addition, we assume that:  $\|\mathbf{W}^{(p)}\|_\infty \leq C_3^{(p)}$  for every  $\mathbf{W}^{(p)}$ , i.e., the weight matrix of the  $p$ -th layer. In this case, the main theorem is extended as follows:

**Theorem 4.** *With probability of at least  $(1 - \delta)$ , we have*

$$L(\hat{f}) - L(f^*) \leq (\sqrt{\mathcal{J}^P} + C_2) \frac{A}{\sqrt{N}} + \frac{1}{2} \left( \sqrt{\mathcal{J}^P} + C_2 \right)^2 \sqrt{\frac{2 \log(2/\delta)}{N}}, \quad (23)$$

where  $A = 4((2L_\rho)^P C_1 C_3^0 \prod_{p=0}^{P-1} \sqrt{M^{(p)}} C_3^{(p)} + |\phi(0)| \sum_{p=0}^{P-1} (2L_\rho)^{P-1-p} \prod_{j=p}^{P-1} \sqrt{M^j} C_3^j)$ , and  $\mathcal{J}^P$  is defined recursively using the following identities:  $\mathcal{J}^0 = C_3^0 C_1$  and  $\mathcal{J}^{(p)} = M^{(p)} C^{p2} (M^{p2} (L_\rho \mathcal{J}^{p-1} + \phi(0))^2 - M(M-1) d_{\min}^{(p)2})$ , for  $p = 1, \dots, P$ .

*Proof.* Lemma 5 in [58] provides an upper-bound for the hypothesis class. We denote by  $\mathbf{v}^{(p)}$  the outputs of the  $p^{\text{th}}$  hidden layer before applying the activation function:

$$\mathbf{v}^0 = [\mathbf{w}_1^{0T} \mathbf{x}, \dots, \mathbf{w}_{M^0}^{0T} \mathbf{x}], \quad (24)$$

$$\mathbf{v}^{(p)} = \left[ \sum_{j=1}^{M^{p-1}} w_{j,1}^{(p)} \phi(\mathbf{v}_j^{p-1}), \dots, \sum_{j=1}^{M^{p-1}} w_{j,M^{(p)}}^{(p)} \phi(\mathbf{v}_j^{p-1}) \right], \quad (25)$$

$$\mathbf{v}^{(p)} = \left[ \mathbf{w}_1^{(p)T} \boldsymbol{\phi}^{(p)}, \dots, \mathbf{w}_{M^{(p)}}^{(p)T} \boldsymbol{\phi}^{(p)} \right], \quad (26)$$

where  $\boldsymbol{\phi}^{(p)} = [\phi(v_1^{p-1}), \dots, \phi(v_{M^{p-1}}^{p-1})]$ . We have  $\|\mathbf{v}^{(p)}\|_2^2 = \sum_{m=1}^{M^{(p)}} (\mathbf{w}_m^{(p)T} \boldsymbol{\phi}^{(p)})^2$  and  $\mathbf{w}_m^{(p)T} \boldsymbol{\phi}^{(p)} \leq C_3^{(p)} \sum_n \phi_n^{(p)}$ . Thus,

$$\|\mathbf{v}^{(p)}\|_2^2 \leq \sum_{m=1}^{M^{(p)}} \left( C_3^{(p)} \sum_n \phi_n^{(p)} \right)^2 = M^{(p)} C_3^{p2} \left( \sum_n \phi_n^{(p)} \right)^2 = M^{(p)} C_3^{p2} \sum_{mn} \phi_m^{(p)} \phi_n^{(p)}. \quad (27)$$

We use the same decomposition trick of  $\phi_m^{(p)} \phi_n^{(p)}$  as in the proof of Lemma 2. We need to bound  $\sup_x \phi^{(p)}$ :

$$\sup_x \phi^{(p)} < \sup_x (L_\rho |\mathbf{v}^{p-1}| + \phi(0)) < L_\rho \|\mathbf{v}^{p-1}\|_2^2 + \phi(0). \quad (28)$$

Thus, we have

$$\|\mathbf{v}^{(p)}\|_2^2 \leq M^{(p)} C_3^{p2} (M^2 (L_\rho \|\mathbf{v}^{p-1}\|_2^2 + \phi(0))^2 - M(M-1) d_{\min}^2) = \mathcal{J}^P. \quad (29)$$

We found a recursive bound for  $\|\mathbf{v}^{(p)}\|_2^2$  and we note that for  $p = 0$  we have  $\|\mathbf{v}^0\|_2^2 \leq \|W^0\|_\infty C_1 \leq C_3^0 C_1 = \mathcal{J}^0$ . Thus,

$$\sup_{\mathbf{x}, f^P \in \mathcal{F}^P} |f(\mathbf{x})| = \sup_{\mathbf{x}, f^P \in \mathcal{F}^P} |\mathbf{v}^P| \leq \sqrt{\mathcal{J}^P}. \quad (30)$$

By replacing the variables in Lemma 3, we have

$$\begin{aligned} L(\hat{f}) - L(f^*) &\leq 4(\sqrt{\mathcal{J}^P} + C_2) \left( \frac{(2L_\rho)^P C_1 C_3^0}{\sqrt{N}} \prod_{p=0}^{P-1} \sqrt{M^{(p)}} C_3^{(p)} \right. \\ &\quad \left. + \frac{|\phi(0)|}{\sqrt{N}} \sum_{p=0}^{P-1} (2L_\rho)^{P-1-p} \prod_{j=p}^{P-1} \sqrt{M^j} C_3^j \right) + \frac{1}{2} \left( \sqrt{\mathcal{J}^P} + C_2 \right)^2 \sqrt{\frac{2 \log(2/\delta)}{N}}, \end{aligned}$$

Taking  $A = 4((2L_\rho)^P C_1 C_3^0 \prod_{p=0}^{P-1} \sqrt{M^{(p)}} C_3^{(p)} + |\phi(0)| \sum_{p=0}^{P-1} (2L_\rho)^{P-1-p} \prod_{j=p}^{P-1} \sqrt{M^j} C_3^j)$  completes the proof.

In Theorem 4, we see that  $\mathcal{J}^P$  is decreasing with respect to  $d_{min}^{(p)}$ . This extends our results to the multi-layer neural network case.

### 4.3 Multiple outputs

Finally, we consider the case of a neural network with a multi-dimensional output, i.e.,  $\mathbf{y} \in R^D$ . In this case, we can extend Theorem 1 with the following two theorems:

**Theorem 5.** *For a multivariate regression trained with the squared error, there exists a constant  $A$  such that, with probability at least  $(1 - \delta)$ , we have*

$$L(\hat{f}) - L(f^*) \leq (\sqrt{\mathcal{J}} + C_2) \frac{A}{\sqrt{N}} + \frac{D}{2} (\sqrt{\mathcal{J}} + C_2)^2 \sqrt{\frac{2 \log(2/\delta)}{N}} \quad (31)$$

where  $\mathcal{J} = C_4^2 (MC_5^2 + M(M-1)(C_5^2 - d_{min}^2))$ ,  $C_5 = L_\rho C_1 C_3 + \phi(0)$ , and  $A = 4D(2L_\rho C_{134} + C_4 |\phi(0)|)M$ .

*Proof.* The squared loss  $\frac{1}{2} \|f(\mathbf{x}) - \mathbf{y}\|_2^2$  can be decomposed into  $D$  terms  $\frac{1}{2} (f(\mathbf{x})_k - y_k)^2$ . Using Theorem 1, we can derive the bound for each term and, thus, we have:

$$L(\hat{f}) - L(f^*) \leq 4D(\sqrt{\mathcal{J}} + C_2) \left( 2L_\rho C_{134} + C_4 |\phi(0)| \right) \frac{M}{\sqrt{N}} + \frac{D}{2} (\sqrt{\mathcal{J}} + C_2)^2 \sqrt{\frac{2 \log(2/\delta)}{N}}, \quad (32)$$

where  $C_{134} = C_1 C_3 C_4$ ,  $\mathcal{J} = C_4^2 (MC_5^2 + M(M-1)(C_5^2 - d_{min}^2))$ , and  $C_5 = L_\rho C_1 C_3 + \phi(0)$ . Taking  $A = 4D(2L_\rho C_{134} + C_4 |\phi(0)|)M$  completes the proof.

**Theorem 6.** *For a multi-class classification task using the cross-entropy loss, there exists a constant  $A$  such that, with probability at least  $(1 - \delta)$ , we have*

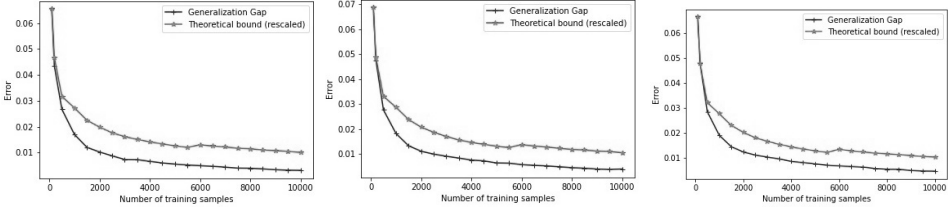
$$L(\hat{f}) - L(f^*) \leq \frac{A}{(D-1 + e^{-2\sqrt{\mathcal{J}}})\sqrt{N}} + \log \left( 1 + (D-1)e^{2\sqrt{\mathcal{J}}} \right) \sqrt{\frac{2 \log(2/\delta)}{N}} \quad (33)$$

where  $\mathcal{J} = C_4^2 (MC_5^2 + M(M-1)(C_5^2 - d_{min}^2))$  and  $C_5 = L_\rho C_1 C_3 + \phi(0)$ , and  $A = 4D(D-1)(2L_\rho C_{134} + C_4 |\phi(0)|)M$ .

*Proof.* Using Lemma 9 in [58], we have  $\sup_{f, \mathbf{x}, \mathbf{y}} l = \log(1 + (D-1)e^{2\sqrt{\mathcal{J}}})$  and  $l$  is  $\frac{D-1}{D-1+e^{-2\sqrt{\mathcal{J}}}}$ -Lipschitz. Thus, using the decomposition property of the Rademacher complexity, we have

$$\mathcal{R}_n(\mathcal{A}) \leq \frac{4D(D-1)}{D-1+e^{-2\sqrt{\mathcal{J}}}} (2L_\rho C_{134} + C_4 |\phi(0)|) \frac{M}{\sqrt{N}}. \quad (34)$$

Taking  $A = 4D(D-1)(2L_\rho C_{134} + C_4 |\phi(0)|)M$  completes the proof.



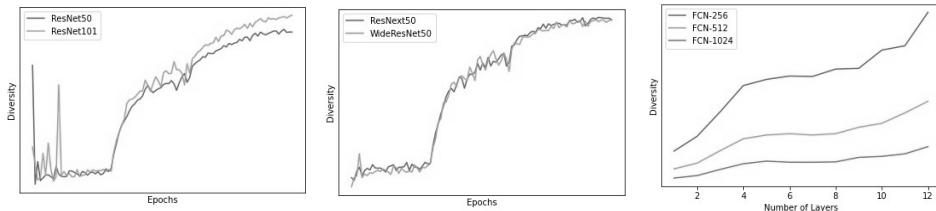
**Fig. 2.** Generalization gap, i.e., train error - test error, and the theoretical bound, i.e.,  $(C_5^2 - d_{min}^2)/\sqrt{N}$ , as a function of the number of training samples on MNIST dataset for neural networks with intermediate layer sizes from left to right: 128 (correlation=0.9948), 256 (correlation=0.9939), and 512 (correlation=0.9953). The theoretical term has been scaled in the same range as the generalization gap. All results are averaged over 5 random seeds.

Theorems 5 and 6 extend our result for the multi-dimensional regression and classification tasks, respectively. Both bounds are inversely proportional to the diversity factor  $d_{min}$ . We note that for the classification task the upper-bound is exponentially decreasing with respect to  $d_{min}$ . This shows that learning a diverse and rich feature representation yields a tighter generalization gap and, thus, theoretically guarantees a stronger generalization performance.

## 5 Discussion and open problems

In this paper, we showed how the diversity of the features learned by a two-layer neural network trained with the least-squares loss affects generalization. We quantified the diversity by the average  $L_2$ -distance between the hidden-layer features and we derived novel diversity-dependant generalization bounds based on Rademacher complexity for such models. The derived bounds are inversely-proportional to the diversity term, thus demonstrating that more distinct features within the hidden layer can lead to better generalization. We also showed how to extend our results to deeper networks and different losses.

The bound found in Theorem 1 suggests that the generalization gap, with respect to diversity, is inversely proportional to  $d_{min}$  and scales as  $\sim (C_5^2 - d_{min}^2)/\sqrt{N}$ . We validate this finding empirically in Figure 2. We train a two-layer neural network on the MNIST dataset for 100 epochs using SGD with a learning rate of 0.1 and batch size of 256. We show the generalization gap, i.e., test error - train error, and the theoretical bound, i.e.,  $(C_5^2 - d_{min}^2)/\sqrt{N}$ , for different training set sizes.  $d_{min}$  is the lower bound of diversity. Empirically, it can be estimated as the minimum feature diversity over the training data  $S$ :  $d_{min} = \min_{x \in S} \frac{1}{2M(M-1)} \sum_{n \neq m}^M (\phi_n(x) - \phi_m(x))^2$ . We experiment with different sizes of the hidden layer, namely 128, 256, and 512. The average results using 5 random seeds are reported for different training sizes in Figure 2 showing that the theoretical bound correlates consistently well (correlation > 0.9939) with the generalization error.



**Fig. 3.** From left to right: (a)-(b) Tracking the diversity during the training for different models on ImageNet. (c) Final diversity as a function of depth for different models on MNIST.

As shown in Figure 1, diversity increases for neural networks along the training phase. To further investigate this observation, we conduct additional experiments on ImageNet [49] dataset using 4 different state-of-the-art models: **ResNet50** and **ResNet101**, i.e., the standard ResNet model [19] with 50 layers and 101 layers, **ResNext50** [60], and **WideResNet50** [62] with 50 layers. All models are trained with SGD using standard training protocol [10, 22, 64]. We track the diversity, as defined in equation 3, of the features of the last intermediate layer. The results are shown in Figure 3 (a) and (b). As it can be seen, SGD without any explicit regularization implicitly optimizes diversity and converges toward regions with high features’ distinctness. These observations suggest the following conjecture:

*Conjecture 1.* Standard training with SGD implicitly optimizes the diversity of intermediate features.

Studying the fundamental properties of SGD is extremely important to understand generalization in deep learning [23, 25, 27, 54, 65]. Conjecture 1 suggests a new implicit bias for SGD, showing that it favors regions with high feature diversity.

Another research question related to diversity that is worth investigating is: *How does the network depth affect diversity?* In order to answer this question, we conduct an empirical experiment using MNIST dataset [37]. We use fully connected networks (FCNs) with ReLU activation and different depths (1 to 12). We experiment with three models with different widths, namely FCN-256, FCN-512, and FCN-1024, with 256, 512, and 1024 units per layer, respectively. We measure the final diversity of the last hidden layer for the different depths. The average results using 5 random seeds are reported in Figure 3 (c). Interestingly, in this experiment, increasing the depth consistently leads to learning more distinct features and higher diversity for the different models. However, by looking at Figure 1, we can see that having more parameters does not always lead to higher diversity. This suggests the following open question:

**Open Problem 1** *When does having more parameters/depth lead to higher diversity?*

Understanding the difference between shallow and deep models and why deeper models generalize better is one of the puzzles of deep learning [26, 40, 47]. The insights gained by studying Open Problem 1 can lead to a novel key advantage of depth: deeper models are able to learn a richer and more diverse set of features.

Another interesting line of research is adversarial robustness [40, 41, 46, 56]. Intuitively, learning distinct features can lead to a richer representation and, thus, more robust networks. However, the theoretical link is missing. This leads to the following open problem:

**Open Problem 2** *Can the theoretical tools proposed in this paper be used to prove the benefits of feature diversity for adversarial robustness?*

## Ethical consideration

This is a theoretical work and does not present any foreseeable societal consequences. The data used in this work comes from publicly accessible channels.

**Acknowledgements** This work has been supported by the NSF-Business Finland Center for Visual and Decision Informatics (CVDI) project AMALIA. The work of Jenni Raitoharju was funded by the Academy of Finland (project 324475). Alexandros Iosifidis acknowledges funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 957337.

## References

1. Arora, S., Cohen, N., Hu, W., Luo, Y.: Implicit regularization in deep matrix factorization. In: *Advances in Neural Information Processing Systems*. pp. 7413–7424 (2019)
2. Arora, S., Du, S., Kakade, S., Luo, Y., Saunshi, N.: Provable representation learning for imitation learning via bi-level optimization. In: *International Conference on Machine Learning*. PMLR (2020)
3. Arpit, D., Jastrzkebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A closer look at memorization in deep networks. In: *International Conference on Machine Learning*. pp. 233–242. PMLR (2017)
4. Bao, Y., Jiang, H., Dai, L., Liu, C.: Incoherent training of deep neural networks to de-correlate bottleneck features for speech recognition. In: *International Conference on Acoustics, Speech and Signal Processing*. pp. 6980–6984 (2013)
5. Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory* pp. 930–945 (1993)
6. Barron, A.R.: Approximation and estimation bounds for artificial neural networks. *Machine Learning* pp. 115–133 (1994)
7. Bartlett, P.L., Mendelson, S.: Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* pp. 463–482 (2002)



8. Bietti, A., Mialon, G., Chen, D., Mairal, J.: A kernel perspective for regularizing deep neural networks. In: International Conference on Machine Learning. pp. 664–674 (2019)
9. Bubeck, S., Sellke, M.: A universal law of robustness via isoperimetry. Neural Information Processing Systems (Neurips) (2021)
10. Cogswell, M., Ahmed, F., Girshick, R.B., Zitnick, L., Batra, D.: Reducing overfitting in deep networks by decorrelating representations. In: International Conference on Learning Representations (2016)
11. Deng, H., Ren, Q., Chen, X., Zhang, H., Ren, J., Zhang, Q.: Discovering and explaining the representation bottleneck of dnns. arXiv preprint arXiv:2111.06236 (2021)
12. Deng, Z., Zhang, L., Vodrahalli, K., Kawaguchi, K., Zou, J.: Adversarial training helps transfer learning via better representations. Neural Information Processing Systems (Neurips) (2021)
13. Du, S.S., Hu, W., Kakade, S.M., Lee, J.D., Lei, Q.: Few-shot learning via learning the representation, provably. International Conference on Learning Representations (2021)
14. Dziugaite, G.K., Roy, D.M.: Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. arXiv preprint arXiv:1703.11008 (2017)
15. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. arXiv preprint arXiv:2010.01412 (2020)
16. Golan, I., El-Yaniv, R.: Deep anomaly detection using geometric transformations. In: Advances in Neural Information Processing Systems. pp. 9758–9769 (2018)
17. Golowich, N., Rakhlin, A., Shamir, O.: Size-independent sample complexity of neural networks. In: Conference On Learning Theory. pp. 297–299 (2018)
18. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep learning. MIT Press (2016)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
20. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal processing magazine **29**(6), 82–97 (2012)
21. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012)
22. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
23. Ji, Z., Telgarsky, M.: The implicit bias of gradient descent on nonseparable data. In: Proceedings of the Thirty-Second Conference on Learning Theory. pp. 1772–1798 (2019)
24. Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., Bengio, S.: Fantastic generalization measures and where to find them. International Conference on Learning Representations (2019)
25. Kalimeris, D., Kaplun, G., Nakkiran, P., Edelman, B., Yang, T., Barak, B., Zhang, H.: Sgd on neural networks learns functions of increasing complexity. Neural Information Processing Systems **32**, 3496–3506 (2019)

26. Kawaguchi, K., Bengio, Y.: Depth with nonlinearity creates no bad local minima in resnets. *Neural Networks* **118**, 167–174 (2019)
27. Kawaguchi, K., Huang, J.: Gradient descent finds global minima for generalizable deep neural networks of practical sizes. In: 2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton). pp. 92–99. IEEE (2019)
28. Kawaguchi, K., Kaelbling, L.P., Bengio, Y.: Generalization in deep learning. arXiv preprint arXiv:1710.05468 (2017)
29. Kornblith, S., Chen, T., Lee, H., Norouzi, M.: Why do better loss functions lead to less transferable features? *Advances in Neural Information Processing Systems* **34** (2021)
30. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
31. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems* (2012)
32. Kukačka, J., Golkov, V., Cremers, D.: Regularization for deep learning: A taxonomy. arXiv preprint arXiv:1710.10686 (2017)
33. Kwok, J.T., Adams, R.P.: Priors for diversity in generative latent variable models. In: *Advances in Neural Information Processing Systems*. pp. 2996–3004 (2012)
34. Laakom, F., Raitoharju, J., Iosifidis, A., Gabbouj, M.: Efficient cnn with uncorrelated bag of features pooling. In: 2022 IEEE Symposium Series on Computational Intelligence (SSCI) (2022)
35. Laakom, F., Raitoharju, J., Iosifidis, A., Gabbouj, M.: Reducing redundancy in the bottleneck representation of the autoencoders. arXiv preprint arXiv:2202.04629 (2022)
36. Laakom, F., Raitoharju, J., Iosifidis, A., Gabbouj, M.: Wld-reg: A data-dependent within-layer diversity regularizer. the 37th AAAI Conference on Artificial Intelligence (2023)
37. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**(11), 2278–2324 (1998)
38. Lee, H.B., Nam, T., Yang, E., Hwang, S.J.: Meta dropout: Learning to perturb latent features for generalization. In: *International Conference on Learning Representations* (2019)
39. Li, Z., Gong, B., Yang, T.: Improved dropout for shallow and deep learning. In: *Advances in Neural Information Processing Systems*. pp. 2523–2531 (2016)
40. Liao, Q., Miranda, B., Rosasco, L., Banburski, A., Liang, R., Hidary, J., Poggio, T.: Generalization puzzles in deep networks. *International Conference on Learning Representations* (2020)
41. Mao, C., Gupta, A., Nitin, V., Ray, B., Song, S., Yang, J., Vondrick, C.: Multitask learning strengthens adversarial robustness. In: *European Conference on Computer Vision*. pp. 158–174. Springer (2020)
42. Maurer, A., Pontil, M., Romera-Paredes, B.: The benefit of multitask representation learning. *Journal of Machine Learning Research* (2016)
43. Nagarajan, V., Kolter, J.Z.: Uniform convergence may be unable to explain generalization in deep learning. In: *Advances in Neural Information Processing Systems* (2019)
44. Neyshabur, B., Bhojanapalli, S., McAllester, D., Srebro, N.: Exploring generalization in deep learning. *NIPS* (2017)
45. Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., Srebro, N.: The role of over-parametrization in generalization of neural networks. In: *International Conference on Learning Representations* (2018)

46. Pinot, R., Meunier, L., Araujo, A., Kashima, H., Yger, F., Gouy-Pailler, C., Atif, J.: Theoretical evidence for adversarial robustness through randomization. In: *Advances in Neural Information Processing Systems (Neurips)* (2019)
47. Poggio, T., Kawaguchi, K., Liao, Q., Miranda, B., Rosasco, L., Boix, X., Hidary, J., Mhaskar, H.: Theory of deep learning III: explaining the non-overfitting puzzle. *arXiv preprint arXiv:1801.00173* (2017)
48. Rodriguez-Galvez, B., Bassi, G., Thobaben, R., Skoglund, M.: Tighter expected generalization error bounds via wasserstein distance. *Advances in Neural Information Processing Systems* (2021)
49. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. *International journal of computer vision* (2015)
50. Shalev-Shwartz, S., Ben-David, S.: *Understanding machine learning: From theory to algorithms*. Cambridge university press (2014)
51. Sokolic, J., Giryes, R., Sapiro, G., Rodrigues, M.R.: *Lessons from the rademacher complexity for deep learning* (2016)
52. Sontag, E.D.: VC dimension of neural networks. *NATO ASI Series F Computer and Systems Sciences* pp. 69–96 (1998)
53. Valiant, L.: A theory of the learnable. *Commun. of the ACM* **27**(1), 134–1 (1984)
54. Volhejn, V., Lampert, C.: Does sgd implicitly optimize for smoothness? In: *DAGM German Conference on Pattern Recognition*. pp. 246–259. Springer (2020)
55. Wang, X., Chen, X., Du, S.S., Tian, Y.: Towards demystifying representation learning with non-contrastive self-supervision. *arXiv preprint arXiv:2110.04947* (2021)
56. Wu, B., Chen, J., Cai, D., He, X., Gu, Q.: Do wider neural networks really help adversarial robustness? *Advances in Neural Information Processing Systems* **34** (2021)
57. Xie, B., Liang, Y., Song, L.: Diverse neural network learns true target functions. In: *Artificial Intelligence and Statistics*. pp. 1216–1224 (2017)
58. Xie, P., Deng, Y., Xing, E.: On the generalization error bounds of neural networks under diversity-inducing mutual angular regularization. *arXiv preprint arXiv:1511.07110* (2015)
59. Xie, P., Singh, A., Xing, E.P.: Uncorrelation and evenness: a new diversity-promoting regularizer. In: *International Conference on Machine Learning*. pp. 3811–3820 (2017)
60. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 1492–1500 (2017)
61. Yu, Y., Li, Y.F., Zhou, Z.H.: Diversity regularized machine. In: *International Joint Conference on Artificial Intelligence* (2011)
62. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: *Proceedings of the British Machine Vision Conference (BMVC)* (2016)
63. Zhai, K., Wang, H.: Adaptive dropout with rademacher complexity regularization. In: *International Conference on Learning Representations* (2018)
64. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. *International Conference on Learning Representations 2018* (2018)
65. Zou, D., Wu, J., Gu, Q., Foster, D.P., Kakade, S., et al.: The benefits of implicit regularization from sgd in least squares problems. *Neural Information Processing Systems* **34** (2021)



# PUBLICATION

## II

**WLD-Reg: A Data-Dependent Within-Layer Diversity Regularizer**

F. Laakom, J. Raitoharju, A. Iosifidis and M. Gabbouj

*Proceedings of the AAAI Conference on Artificial Intelligence* 2023, 8421–8429

DOI: 10.1609/aaai.v37i7.26015

© 2023 . Association for the Advancement of Artificial Intelligence



# WLD-Reg: A Data-dependent Within-layer Diversity Regularizer

Firas Laakom<sup>1\*</sup> Jenni Raitoharju,<sup>2</sup> Alexandros Iosifidis,<sup>3</sup> Moncef Gabbouj<sup>1</sup>

<sup>1</sup> Faculty of Information Technology and Communication Sciences, Tampere University, Finland

<sup>2</sup> Faculty of Information Technology, University of Jyväskylä, Finland

<sup>3</sup> DIGIT, Department of Electrical and Computer Engineering, Aarhus University, Denmark

## Abstract

Neural networks are composed of multiple layers arranged in a hierarchical structure jointly trained with a gradient-based optimization, where the errors are back-propagated from the last layer back to the first one. At each optimization step, neurons at a given layer receive feedback from neurons belonging to higher layers of the hierarchy. In this paper, we propose to complement this traditional ‘between-layer’ feedback with additional ‘within-layer’ feedback to encourage the diversity of the activations within the same layer. To this end, we measure the pairwise similarity between the outputs of the neurons and use it to model the layer’s overall diversity. We present an extensive empirical study confirming that the proposed approach enhances the performance of several state-of-the-art neural network models in multiple tasks. The code is publicly available at <https://github.com/firas/AAAI-23-WLD-Reg>

## Introduction

Deep learning has been extensively used in the last decade to solve several tasks (Krizhevsky, Sutskever, and Hinton 2012; Golan and El-Yaniv 2018; Hinton et al. 2012a). A deep learning model, i.e., a neural network, is formed of a sequence of layers with parameters optimized during the training process using training data. Formally, an  $m$ -layer neural network model can be defined as follows:

$$f(\mathbf{x}; \mathbf{W}) = \phi^m(\mathbf{W}^m(\phi^{m-1}(\dots \phi^2(\mathbf{W}^2 \phi^1(\mathbf{W}^1 \mathbf{x}))))), \quad (1)$$

where  $\phi^i(\cdot)$  is the non-linear activation function of the  $i^{th}$  layer and  $\mathbf{W} = \{\mathbf{W}^1, \dots, \mathbf{W}^m\}$  are the model’s weights. Given a training data  $\{\mathbf{x}_i, y_i\}_{i=1}^N$ , the parameters of  $f(\mathbf{x}; \mathbf{W})$  are obtained by minimizing a loss  $\hat{L}(\cdot)$ :

$$\hat{L}(f) = \frac{1}{N} \sum_{i=1}^N l(f(\mathbf{x}_i; \mathbf{W}), y_i). \quad (2)$$

However, neural networks are often over-parameterized, i.e., have more parameters than data. As a result, they tend to overfit to the training samples and not generalize well on unseen examples (Goodfellow et al. 2016). While research on double descent (Advani, Saxe, and Sompolinsky

2020; Belkin et al. 2019; Nakkiran et al. 2020) shows that over-parameterization does not necessarily lead to overfitting, avoiding overfitting has been extensively studied (Dziugaite and Roy 2017; Foret et al. 2020; Nagarajan and Kolter 2019; Neyshabur et al. 2018; Poggio et al. 2017; Grari et al. 2021) and various approaches and strategies, such as data augmentation (Goodfellow et al. 2016; Zhang et al. 2018), regularization (Arora et al. 2019; Bietti et al. 2019; Kukačka, Golkov, and Cremers 2017; Ouali, Hudelot, and Tami 2021; Han and Guo 2021), and Dropout (Hinton et al. 2012b; Lee et al. 2019; Li, Gong, and Yang 2016; Wang et al. 2019), have been proposed to close the gap between the empirical loss and the expected loss.

Diversity of learners is widely known to be important in ensemble learning (Li, Yu, and Zhou 2012; Yu, Li, and Zhou 2011) and, particularly in deep learning context, diversity of information extracted by the network neurons has been recognized as a viable way to improve generalization (Xie, Liang, and Song 2017; Xie, Deng, and Xing 2015b). In most cases, these efforts have focused on making the set of weights more diverse (Yang, Gkatzelis, and Stoyanovich 2019; Malkin and Bilmes 2009). However, diversity of the activations has not received much attention. Here, we argue that due to the presence of non-linear activations, diverse weights do not guarantee diverse feature representation. Thus, we propose focusing on the diversity on top of feature mapping instead of the weights.

To the best of our knowledge, only (Cogswell et al. 2016; Laakom et al. 2021a) have considered diversity of the activations directly in the neural network context. The work in (Laakom et al. 2021a) studied theoretically how diversity affects generalization showing that it can reduce overfitting. The work in (Cogswell et al. 2016) proposed an additional loss term using cross-covariance of hidden activations, which encourages the neurons to learn diverse or non-redundant representations. The proposed approach, known as DeCov, was empirically proven to alleviate overfitting and to improve the generalization ability of neural network. However, modeling diversity as the sum of the pairwise cross-covariance, it is not scale-invariant and can lead to trivial solutions. Moreover, it can capture only the pairwise diversity between components and is unable to capture the “higher-order diversity”.

In this work, we propose a novel approach to encour-

\*This work was supported by NSF-Business Finland Center for Visual and Decision Informatics (CVDI) project AMALIA.

age activation diversity within the same layer. We propose complementing the ‘between-layer’ feedback with additional ‘within-layer’ feedback to penalize similarities between neurons on the same layer. Thus, we encourage each neuron to learn a distinctive representation and to enrich the data representation learned within each layer. We propose three variants for our approach that are based on different global diversity definitions.

Our contributions in this paper are as follows:

- We propose a new approach to encourage the ‘diversification’ of the layers’ output feature maps in neural networks. The proposed approach has three variants. The main intuition is that, by promoting the within-layer activation diversity, neurons within a layer learn distinct patterns and, thus, increase the overall capacity of the model.
- We show empirically that the proposed within-layer activation diversification boosts the performance of neural networks. Experimental results on several tasks show that the proposed approach outperforms competing methods.

### Within-layer Diversity Regularizer

In this section, we propose a novel diversification strategy, where we encourage neurons within a layer to activate in a mutually different manner, i.e., to capture different patterns. In this paper, we define as “feature layer” the last intermediate layer in a neural network. In the rest of the paper, we focus on this layer and propose a data-dependent regularizer which forces each unit within this layer to learn a distinct pattern and penalizes the similarities between the units. Intuitively, the proposed approach reduces the reliance of the model on a single pattern and, thus, can improve generalization.

We start by modeling the global similarity between two units. Let  $\phi_n(\mathbf{x}_j)$  and  $\phi_m(\mathbf{x}_j)$  be the outputs of the  $n^{th}$  and  $m^{th}$  unit in the feature layer for the same input sample  $\mathbf{x}_j$ . The similarity  $s_{nm}$  between the  $n^{th}$  and  $m^{th}$  neurons can be obtained as the average similarity measure of their outputs for  $N$  input samples. We use the radial basis function to express the similarity:

$$s_{nm} := \frac{1}{N} \sum_{j=1}^N \exp(-\gamma \|\phi_n(\mathbf{x}_j) - \phi_m(\mathbf{x}_j)\|^2), \quad (3)$$

where  $\gamma$  is a hyper-parameter. The similarity  $s_{nm}$  can be computed over the whole dataset or batch-wise. Intuitively, if two neurons  $n$  and  $m$  have similar outputs for many samples, their corresponding similarity  $s_{nm}$  will be high. Otherwise, their similarity  $s_{nm}$  is small and they are considered “diverse”.

Next, based on these pairwise similarities, we propose three variants for obtaining the overall similarity  $J$  of all the units within the feature layer:

- **Direct:**  $J := \sum_{n \neq m} s_{nm}$ . In this variant, we model the global layer similarity directly as the sum of the pairwise similarities between the neurons. By minimizing their sum, we encourage the neurons to learn different representations.

- **Det:**  $J := -\det(\mathbf{S})$ , where  $\mathbf{S}$  is a similarity matrix defined as  $S_{nm} = s_{nm}$ . This variant is inspired by the Determinantal Point Process (DPP) (Kulesza and Taskar 2010, 2012), as the determinant of  $\mathbf{S}$  measures the global diversity of the set. Geometrically,  $\det(\mathbf{S})$  is the volume of the parallelepiped formed by vectors in the feature space associated with  $s$ . Vectors that result in a larger volume are considered to be more “diverse”. Thus, maximizing  $\det(\cdot)$  (minimizing  $-\det(\cdot)$ ) encourages the diversity of the learned features.
- **Logdet:**  $J := -\log \det(\mathbf{S})^1$ . This variant has the same motivation as the second one. We use Logdet instead of Det as Logdet is a convex function over the positive definite matrix space.

It should be noted here that the first proposed variant, i.e., direct, similar to DeCov (Cogswell et al. 2016), captures only the pairwise similarity between components and is unable to capture the higher-order “diversity”, whereas the other two variants consider the global similarity and are able to measure diversity in a more global manner. Promoting diversity of activations within a layer can lead to tighter generalization bound and can theoretically decrease the gap between the empirical and the true risks (Laakom et al. 2021a).

The proposed global similarity measures  $J$  can be minimized by using them as an additional loss term. However, we note that the pair-wise similarity measure  $s_{nm}$ , expressed in equation 3, is not scale-invariant. In fact, it can be trivially minimized by making all activations of the feature layer high, i.e., by multiplying by a high scaling factor, which has no effect on the performance, since the model can rescale high activations to normal values simply by learning small weights on the next layer. To alleviate this problem, we propose an additional term, which penalizes high activation values. The total proposed additional loss is defined as follows:

$$\hat{L}_{WLD-Reg} := \lambda_1 J + \lambda_2 \sum_{i=1}^N \|\Phi(\mathbf{x}_i)\|_2^2, \quad (4)$$

where  $\Phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \dots, \phi_C(\mathbf{x})]$  is the feature vector,  $C$  is the number of units within the feature layer, and  $\lambda_1$  and  $\lambda_2$  are two hyper-parameters controlling the contribution of each term to the diversity loss. Intuitively, the first term of equation 4 penalizes the similarity between the units and promotes diversity, whereas the second term ensures the scale-invariance of the proposed regularizer.

The total loss function  $\hat{L}(f)$  defined in equation 2 is augmented as follows:

$$\begin{aligned} \hat{L}_{aug}(f) &:= \hat{L}(f) + \hat{L}_{WLD-Reg} \\ &= \hat{L}(f) + \lambda_1 J + \lambda_2 \sum_{i=1}^N \|\Phi(\mathbf{x}_i)\|_2^2. \end{aligned} \quad (5)$$

The proposed approach is summarized in Algorithm 1. We note that our approach can be incorporated in a plug-and-

<sup>1</sup>This is defined only if  $\mathbf{S}$  is positive definite. It can be shown that in our case  $\mathbf{S}$  is positive semi-definite. Thus, in practice, we use a regularized version  $(\mathbf{S} + \epsilon \mathbf{I})$  to ensure the positive definiteness.



---

**Algorithm 1:** One epoch of training with WLD-Reg

---

**Model:** Given a neural network  $f(\cdot)$  with a feature representation  $\phi(\cdot)$ , i.e., last intermediate layer.

**Input:** Training Data:  $\{\mathbf{x}_i, y_i\}_{i=1}^N$

**Parameters:**  $\lambda_1$  and  $\lambda_2$  in equation 4

---

- 1: **for** every mini-batch:  $\{\mathbf{x}_i, y_i\}_{i=1}^m \in \{\mathbf{x}_i, y_i\}_{i=1}^N$  **do**
  - 2: Forward pass the inputs  $\{\mathbf{x}_i\}_{i=1}^m$  into the model to obtain the outputs  $\{f(\mathbf{x}_i)\}_{i=1}^m$  and the feature representations  $\{\Phi(\mathbf{x}_i)\}_{i=1}^m$
  - 3: Compute the standard loss  $\hat{L}(f)$  (equation 2).
  - 4: Compute the extra loss  $\hat{L}_{WLD-Reg}$  (equation 4).
  - 5: Compute the total loss  $\hat{L}_{aug}(f)$  (equation 5)
  - 6: Compute the gradient of the total loss and use it to update the weights of  $f$ .
  - 7: **end for**
  - 8: **return** Return  $f$ .
- 

play manner into any neural network-based approach to augment the original loss and to ensure learning diverse features. We also note that although in this paper, we focus only on applying diversity regularizer to a single layer, i.e., the feature layer, our proposed diversity loss, as in (Cogswell et al. 2016), can be applied to multiple layers within the model.

Our newly proposed loss function defined in equation 5 has two terms. The first term is the classic loss function. It computes the loss with respect to the ground-truth. In the back-propagation, this feedback is back-propagated from the last layer to the first layer of the network. Thus, it can be considered as a between-layer feedback, whereas the second term is computed within a layer. From equation 5, we can see that our proposed approach can be interpreted as a regularization scheme. However, regularization in deep learning is usually applied directly on the parameters, i.e., weights (Goodfellow et al. 2016; Kukačka, Golkov, and Cremers 2017), while in our approach a data-dependent additional term is defined over the output maps of the layers. For a feature layer with  $C$  units and a batch size of  $m$ , the additional computational cost is  $O(C^2(m+1))$  for Direct variant and  $O(C^3 + C^2m)$  for both Det and Logdet variants.

## Related work

**Diversity promoting strategies** have been widely used in ensemble learning (Li, Yu, and Zhou 2012; Yu, Li, and Zhou 2011), sampling (Biyik et al. 2019; Derezinski, Candriello, and Valko 2019; Gartrell et al. 2019), energy-based models (Laakom et al. 2021b; Zhao, Mathieu, and LeCun 2017), ranking (Gan et al. 2020; Yang, Gkatzelis, and Stoyanovich 2019), pruning by reducing redundancy (He et al. 2019; Kondo and Yamauchi 2014; Lee et al. 2020; Singh et al. 2020), and semi-supervised learning (Zbontar et al. 2021). In the deep learning context, various approaches have used diversity as a direct regularizer on top of the weight parameters. Here, we present a brief overview of these regularizers. Based on the way diversity is de-

finied, we can group these approaches into two categories. The first group considers the regularizers that are based on the pairwise dissimilarity of the components, i.e., the overall set of weights is diverse if every pair of weights is dissimilar. Given the weight vectors  $\{\mathbf{w}_m\}_{m=1}^M$ , (Yu, Li, and Zhou 2011) defines the regularizer as  $\sum_{mn} (1 - \theta_{mn})$ , where  $\theta_{mn}$  represents the cosine similarity between  $\mathbf{w}_m$  and  $\mathbf{w}_n$ . In (Bao et al. 2013), an incoherence score defined as  $-\log\left(\frac{1}{M(M-1)} \sum_{mn} \beta |\theta_{mn}|^{\frac{1}{\beta}}\right)$ , where  $\beta$  is a positive hyperparameter, is proposed. In (Xie, Deng, and Xing 2015a; Xie, Zhu, and Xing 2016),  $\text{mean}(\theta_{mn}) - \text{var}(\theta_{mn})$  is used to regularize Boltzmann machines. The authors theoretically analyzed its effect on the generalization error bounds in (Xie, Deng, and Xing 2015b) and extend it to kernel space in (Xie, Liang, and Song 2017). The second group of regularizers considers a more global view of diversity. For example, in (Malkin and Bilmes 2008, 2009; Xie, Singh, and Xing 2017), a weight regularization based on the determinant of the weights' covariance is proposed based on determinantal point process (Kulesza and Taskar 2012; Kwok and Adams 2012).

Unlike the aforementioned methods which promote diversity on the weight level and similar to our method, (Cogswell et al. 2016; Laakom et al. 2022) proposed to enforce dissimilarity on the feature map outputs, i.e., on the activations. To this end, they proposed an additional loss based on the pairwise covariance of the activation outputs. Their additional loss,  $L_{Decov}$ , is defined as the squared sum of the non-diagonal elements of the global covariance matrix  $C$  of the activations:

$$L_{Decov} = \frac{1}{2} (\|C\|_F^2 - \|\text{diag}(C)\|_2^2), \quad (6)$$

where  $\|\cdot\|_F$  is the Frobenius norm. Their approach,  $Decov$ , yielded superior empirical performance. However, correlation is highly sensitive to noise (Kim, Kim, and Ergün 2015), as opposite to the RBF-based distance used in our approach (Savas and Dosis 2019; Haykin 2010). Moreover, the  $Decov$  approach only captures the pairwise diversity between the components, whereas we propose variants of our approach which consider a global view of diversity. Moreover, based on the cross-covariance, their approach is not scale-invariant. In fact, it can be trivially minimized by making all activations in the latent representation small, which has no effect on the generalization since the model can rescale tiny activations to normal values simply by learning large weights on the next layer.

## Experimental results

### CIFAR10 & CIFAR100

We start by evaluating our proposed diversity approach on two image datasets: CIFAR10 and CIFAR100 (Krizhevsky, Hinton et al. 2009). They contain 60,000 (50,000 train/10,000 test)  $32 \times 32$  images grouped into 10 and 100 distinct categories, respectively. We split the original training set (50,000) into two sets: we use the first 40,000 images as the main training set and the last 10,000 as a validation set for hyperparameters optimization. We use our approach on two state-of-the-art CNNs:

- **ResNext-29-08-16**: we consider the standard ResNext Model (Xie et al. 2017) with a 29-layer architecture, a cardinality of 8, and a width of 16.
- **ResNet50**: we consider the standard ResNet model (He et al. 2016) with 50 layers.

We compare against the standard networks<sup>2</sup> as well as networks trained with the DeCov diversity strategy (Cogswell et al. 2016). All the models are trained using stochastic gradient descent (SGD) with a momentum of 0.9, weight decay of 0.0001, and a batch size of 128 for 200 epochs. The initial learning rate is set to 0.1 and is then decreased by a factor of 5 after 60, 120, and 160 epochs, respectively. We also adopt a standard data augmentation scheme that is widely used for these two datasets (He et al. 2016; Huang et al. 2017). For all models, the additional diversity term is applied on top the last intermediate layer. The penalty coefficients  $\lambda_1$  and  $\lambda_2$ , in equation 4, for our approach and the penalty coefficient of Decov are chosen from  $\{0.0001, 0.001, 0.01, 0.1\}$ , and  $\gamma$  in the radial basis function is chosen from  $\{1, 10\}$ . For each approach, the model with the best validation performance is used in the test phase. We report the average performance over three random seeds.

Table 1 reports the average top-1 errors of the different approaches with the two basis networks. We note that, compared to the standard approach, employing a diversity strategy consistently boosts the results for all the two models and that our approach consistency outperforms both competing methods (standard and DeCov) in all the experiments. With ResNet50, the three variants of our proposed approach significantly reduce the test errors compared to standard approach over both datasets: 0.51% – 0.63% improvement on CIFAR10 and 1.25% – 1.44% on CIFAR100.

For CIFAR10, the best performance is achieved by the direct variant and the Logdet variant for ResNext and ResNet models, respectively. For example, with ResNext, our direct variant yields 0.65 boost compared to the standard approach and 0.54 boost compared to DeCov. For CIFAR100, the best performance is achieved by our Logdet variant for both models. This variant leads to 1.4% and 0.85% boost for ResNet and ResNext, respectively. Overall, our three variants consistently outperform DeCov and standard approach in all testing configurations.

## ImageNet

To further demonstrate the effectiveness of our approach and its ability to boost the performance of state-of-the-art neural networks, we conduct additional image classification experiments on the ImageNet-2012 classification dataset (Russakovsky et al. 2015) using four different models: ResNet50 (He et al. 2016), Wide-ResNet50 (Zagoruyko and Komodakis 2016), ResNeXt50 (Xie et al. 2017), and ResNet101 (He et al. 2016). The diversity term is applied on the last intermediate layer, i.e., the global average pooling layer for both DeCov and our method.

<sup>2</sup>For the standard approach, the only difference is not using an additional diversity loss. The remaining regularizers, data augmentation, weight decay etc., are all applied as specified per-experiment.

For the hyperparameters, we fix  $\lambda_1 = \lambda_2 = 0.001$  and  $\gamma = 10$  for all the different approaches. The Scope of this paper is feature diversity. However, in this experiment, we also report results with weight diversity approaches. In particular, we compare with the methods in (Yu, Li, and Zhou 2011), (Xie, Deng, and Xing 2015b), (Rodríguez et al. 2016), and (Ayinde, Inanc, and Zurada 2019).

We use the standard augmentation practice for this dataset as in (Zhang et al. 2018; Huang et al. 2017; Cogswell et al. 2016). All the models are trained with a batch size of 256 for 100 epoch using SGD with Nesterov Momentum of 0.9. The learning rate is initially set to 0.1 and decreases at epochs 30, 60, 90 by a factor of 10.

Table 2 reports the test errors of the different approaches on ImageNet dataset. As it can be seen, feature diversity (our approach and DeCov) reduces the test error of the model and yields a better performance compared to the standard approach. We note that, as opposite to feature diversity, weight diversity does not always yield performance improvement and it can sometimes hurt generalization. Compared to decov, our three variants consistently reach better performance.

For ResNet50 and ResNeXt50, the best performance is achieved by our direct variant, yielding more than 0.5% improvement compared to standard approach for both models. For Wide-ResNet50 and ResNet101, our Det variant yields the top performance with over 0.6% boost for Wide-ResNet50. We note that our approach has a small additional time cost. For example for ResNet50, our direct, Det and Logdet variants take only 0.29%, 0.39%, and 0.49% extra training time, respectively.

## Sensitivity analysis

To further investigate the effect of the proposed diversity strategy, we conduct a sensitivity analysis using ImageNet on the hyperparameters of our methods:  $\lambda_1$  and  $\lambda_2$  which controls the contribution of the global diversity term to the global loss. We analyse the effect of the two parameters on the final performance of ResNet50 on ImageNet dataset. The analysis is presented in Figure 1.

As shown in Figure 1, using a diversity strategy, i.e., three variants of our method, consistently outperform the standard approach and are robust to the hyperparameters. For the Direct variant, the best performance is reached with  $\lambda_1 = 0.005$  and  $\lambda_2 = 0.001$ . With this configuration, the model achieve 0.71% improvement compared to the standard approach. For the Det and the Logdet variants, using  $\lambda_1 = 0.001$  and  $\lambda_2 = 0.0005$ , the model reaches the lowest error rate (23.09%) corresponding to 0.75% accuracy boost. Emphasizing diversity and using a high weights ( $\lambda_1$  and  $\lambda_2$ ) still lead to better results compared to standard approach but can make the total loss dominated by the diversity term. In general, we recommend using  $\lambda_1 = \lambda_2 = 0.001$ . However, this depends on the problem at hand.

## Feature diversity reduces overfitting

In (Laakom et al. 2021a; Cogswell et al. 2016), it has been observed that feature diversity can reduce overfitting. To study the effect of feature diversity on the generalization

Table 1: Classification errors of the different approaches on CIFAR10 and CIFAR100 with three different models. Results are averaged over three random seeds.

method	ResNext-29-08-16		ResNet50	
	CIFAR10	CIFAR100	CIFAR10	CIFAR100
Standard	6.93 $\pm$ 0.10	26.73 $\pm$ 0.10	8.28 $\pm$ 0.41	33.39 $\pm$ 0.42
DeCov	6.82 $\pm$ 0.15	26.70 $\pm$ 0.10	8.03 $\pm$ 0.11	32.26 $\pm$ 0.22
Ours(Direct)	<b>6.28 <math>\pm</math> 0.11</b>	26.20 $\pm$ 0.18	7.77 $\pm$ 0.09	32.09 $\pm$ 0.11
Ours(Det)	6.51 $\pm$ 0.16	26.35 $\pm$ 0.23	7.75 $\pm$ 0.12	32.14 $\pm$ 0.28
Ours(Logdet)	6.38 $\pm$ 0.08	<b>25.88 <math>\pm</math> 0.21</b>	<b>7.65 <math>\pm</math> 0.10</b>	<b>31.99 <math>\pm</math> 0.05</b>

Table 2: Performance of different models with different diversity strategies on ImageNet dataset

	ResNet50	Wide-ResNet50	ResNeXt50	ResNet101
Standard	23.84	22.42	22.70	22.33
(Yu, Li, and Zhou 2011)	23.87	22.48	22.57	22.23
(Ayinde, Inanc, and Zurada 2019)	23.95	22.41	22.67	22.36
(Rodríguez et al. 2016)	24.23	22.70	22.80	23.10
(Xie, Deng, and Xing 2015b)	23.79	22.66	22.64	22.71
DeCov	23.62	22.68	22.57	22.31
Ours(Direct)	<b>23.24</b>	21.95	<b>22.25</b>	22.14
Ours(Det)	23.34	<b>21.75</b>	22.44	<b>21.87</b>
Ours(Logdet)	23.32	21.96	22.40	22.04

gap, in Table 3, we report the final training errors and the generalization gap, i.e., training accuracy - test accuracy for the different feature diversity approaches on ImageNet dataset.

Table 3: Generalization Gap, i.e., training error - test error, of different models with different diversity strategies on ImageNet dataset. \* denotes our approach.

	ERM	DeCov	direct*	det*	logdet*
ResNet50	2.87	2.70	<b>1.15</b>	1.23	1.21
Wide-ResNet50	6.33	6.34	4.44	<b>4.34</b>	4.58
ResNeXt50	5.99	5.85	<b>4.41</b>	4.59	4.48
ResNet101	4.64	4.61	3.68	<b>3.38</b>	3.71

As shown in Table 3, we note that using diversity indeed can reduce overfitting and decrease the empirical generalization gap of neural networks. The three variants of our approach significantly reduce overfitting for all the four models by more than 1% compared standard and DeCov for all the models. For example, our Det variant reduces the empirical generalization gap, compared to the standard approach and DeCov, by 2% for Wide-ResNet model and over 1.2% for the ResNet101 model.

### MLP-based models

Beyond CNN models, we also evaluate the performance of our diversity strategy on modern attention-free, multi-layer perceptron (MLP) based models for image classification (Tolstikhin et al. 2021; Liu et al. 2021; Lee-Thorp et al.

2021). Such models are known to exhibit high overfitting and require regularization. We evaluate how diversity affects the accuracy of such models on CIFAR10. In particular, we conduct a simple experiment using two models: MLP-Mixer (Tolstikhin et al. 2021), gMLP (Liu et al. 2021) with four blocks each.

For the diversity strategies, i.e., ours and Decov, similar to our other experiments, the additional loss has been added on top of the last intermediate layer. The input images are resized to  $72 \times 72$ . We use a patch size of  $8 \times 8$  and an embedding dimension of 256. All models are trained for 100 epochs using Adam with learning rate of 0.002, weight decay with rate 0.0001, batch size 256. Standard data augmentation, i.e., random horizontal flip and random zoom with a factor of 20%, is used. We use 10% of the training data for validation. We also reduce the learning rate by a factor of 2 if the validation loss does not improve for 5 epochs and use early stopping when the validation loss does not improve for 10 epochs. All experiments are repeated over 10 random seeds and the average results are reported.

The results in Table 4 show that employing a diversity strategy can indeed improve the performance of these models, thanks to its ability to help learn rich and robust representation of the input. Our proposed approach consistently outperforms the competing methods for both the MLP-Mixer and gMLP. For example, our direct variant leads to 1.15% and 0.3% boost for MLP-Mixer and gMLP, respectively.

For the MLP-mixer, the top performance is achieved by the Det variant of our approach reducing the error rates by

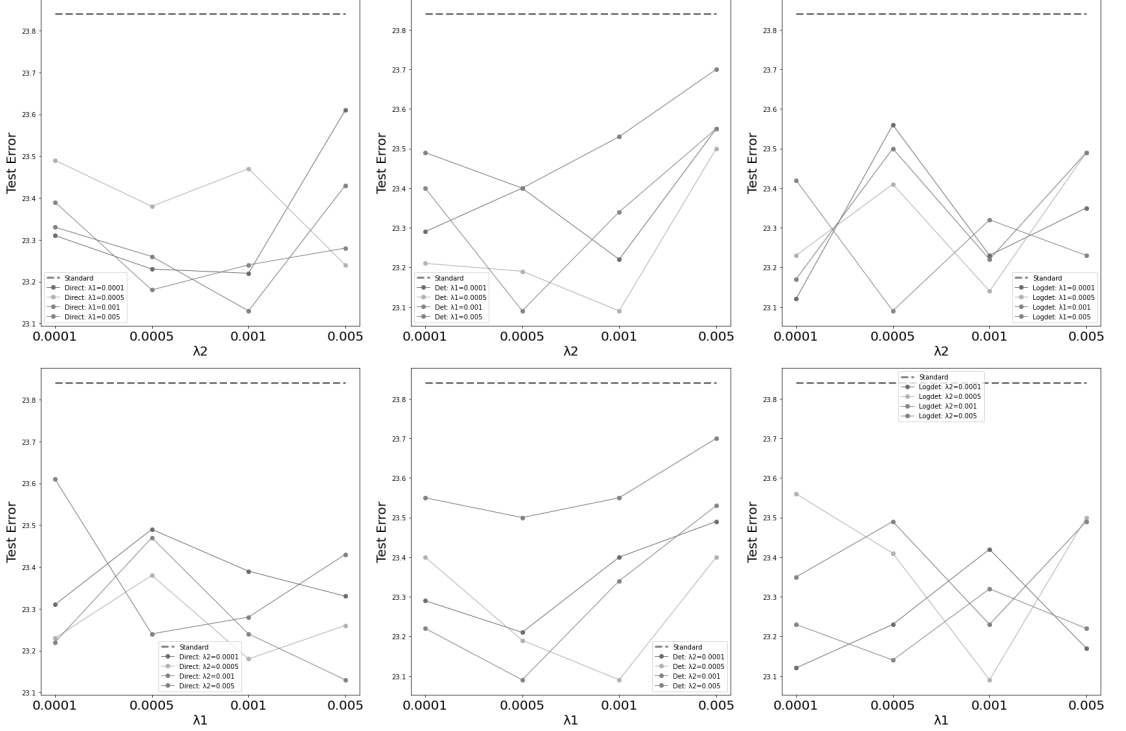


Figure 1: Sensitivity analysis of  $\lambda_1$  and  $\lambda_2$  on the test error using ResNet50 trained on ImageNet. First row contains experiments with fixed  $\lambda_1$  and second row contains experiments with fixed  $\lambda_2$ . From left to right: our Direct variant, our Det variant and our Logdet variant.  $\gamma$  is fixed to 10 in all experiments.

Table 4: Classification errors of modern MLP-based approaches on CIFAR10. Results are averaged over ten random seeds.

	MLP-Mixer	gMLP
Standard	23.93	22.26
DeCov	24.10	22.00
Ours(Direct)	22.78	21.95
Ours(Det)	<b>22.66</b>	21.62
Ours(Logdet)	22.84	<b>21.56</b>

1.27% and 1.44% compared to the standard approach and DeCov, respectively. For the gMLP model, the top performance is achieved by the Logdet variant of our approach boosting the results by 0.7% and 0.44% compared to the standard approach and DeCov, respectively.

### Learning in the presence of label noise

To further demonstrate the usefulness of promoting diversity, we test the robustness of our approach in the presence of label noise. In such situations, standard neural network tend to overfit to the noisy samples and not generalize well to the

test set. Enforcing diversity can lead to better and richer representations attenuating the effect of noise. To show this, we performed additional experiments with label noise (20% and 40%) on CIFAR10 and CIFAR100 using ResNet50. We use the same training protocol used for the original CIFAR10 and CIFAR100: all models are trained using SGD with a momentum of 0.9, weight decay of 0.0001, and a batch size of 128 for 200 epochs. The initial learning rate is set to 0.1 and is then decreased by a factor of 5 after 60, 120, and 160 epochs, respectively. We also adopt a standard data augmentation scheme that is widely used for these two datasets (He et al. 2016; Huang et al. 2017). For all models, the additional diversity term is applied on top the last intermediate layer. For the hyperparameters: The loss weights is chosen from  $\{0.0001, 0.001, 0.01, 0.1\}$  for both our approach ( $\lambda_1$  and  $\lambda_2$ ) and Decov and  $\gamma$  in the radial basis function is chosen from  $\{1, 10\}$ . For each approach, the model with the best validation performance is used in the test phase. The average errors over three random seed are reported.

The results are reported in Table 5. As it can be seen, in the presence of noise, the gap between the standard approach and diversity (Decov and ours) increases. For example, our Logdet variant boosts the results by 1.91% and 2.29% on

Table 5: Classification errors of ResNet50 using different diversity strategies on CIFAR10 and CIFAR100 datasets with different label noise ratios. Results are averaged over three random seeds.

Method	20% label noise		40% label noise	
	CIFAR10	CIFAR100	CIFAR10	CIFAR100
Standard	14.38 $\pm$ 0.29	45.11 $\pm$ 0.52	19.40 $\pm$ 0.80	48.81 $\pm$ 0.57
DeCov	13.75 $\pm$ 0.19	41.93 $\pm$ 0.40	17.60 $\pm$ 0.66	48.23 $\pm$ 0.48
Ours(Direct)	13.31 $\pm$ 0.40	40.10 $\pm$ 0.31	<b>16.96 <math>\pm</math> 0.32</b>	46.73 $\pm$ 0.23
Ours(Det)	13.21 $\pm$ 0.21	40.35 $\pm$ 0.31	17.49 $\pm$ 0.04	46.93 $\pm$ 0.62
Ours(Logdet)	<b>13.01 <math>\pm</math> 0.40</b>	<b>39.97 <math>\pm</math> 0.19</b>	17.24 $\pm$ 0.31	<b>46.52 <math>\pm</math> 0.22</b>

CIFAR10 and CIFAR100 with 40% noise, respectively.

## Conclusions

In this paper, we proposed a new approach to encourage ‘diversification’ of the layer-wise feature map outputs in neural networks. The main motivation is that by promoting within-layer activation diversity, units within the same layer learn to capture mutually distinct patterns. We proposed an additional loss term that can be added on top of any fully-connected layer. This term complements the traditional ‘between-layer’ feedback with an additional ‘within-layer’ feedback encouraging diversity of the activations. Extensive experimental results showing that such a strategy can indeed improve the performance of different state-of-the-art networks across different datasets and different tasks, i.e., image classification, and label noise. We are confident that these results will spark further research in diversity-based approaches to improve the performance of neural networks.

## References

Advani, M. S.; Saxe, A. M.; and Sompolinsky, H. 2020. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132: 428–446.

Arora, S.; Cohen, N.; Hu, W.; and Luo, Y. 2019. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, 7413–7424.

Ayinde, B. O.; Inanc, T.; and Zurada, J. M. 2019. Regularizing deep neural networks by enhancing diversity in feature extraction. *IEEE transactions on neural networks and learning systems*, 30(9): 2650–2661.

Bao, Y.; Jiang, H.; Dai, L.; and Liu, C. 2013. Incoherent training of deep neural networks to de-correlate bottleneck features for speech recognition. In *International Conference on Acoustics, Speech and Signal Processing*, 6980–6984.

Belkin, M.; Hsu, D.; Ma, S.; and Mandal, S. 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32): 15849–15854.

Bietti, A.; Mialon, G.; Chen, D.; and Mairal, J. 2019. A kernel perspective for regularizing deep neural networks. In *International Conference on Machine Learning*, 664–674.

Bytık, E.; Wang, K.; Anari, N.; and Sadigh, D. 2019. Batch active learning using determinantal point processes. *arXiv preprint arXiv:1906.07975*.

Cogswell, M.; Ahmed, F.; Girshick, R. B.; Zitnick, L.; and Batra, D. 2016. Reducing Overfitting in Deep Networks by Decorrelating Representations. In *International Conference on Learning Representations*.

Derezinski, M.; Calandriello, D.; and Valko, M. 2019. Exact sampling of determinantal point processes with sublinear time preprocessing. In *Advances in Neural Information Processing Systems*, 11546–11558.

Dziugaite, G. K.; and Roy, D. M. 2017. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*.

Foret, P.; Kleiner, A.; Mobahi, H.; and Neyshabur, B. 2020. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*.

Gan, L.; Nurbakova, D.; Laporte, L.; and Calabretto, S. 2020. Enhancing Recommendation Diversity using Determinantal Point Processes on Knowledge Graphs. In *Conference on Research and Development in Information Retrieval*, 2001–2004.

Gartrell, M.; Brunel, V.-E.; Dohmatob, E.; and Krichene, S. 2019. Learning nonsymmetric determinantal point processes. In *Advances in Neural Information Processing Systems*, 6718–6728.

Golan, I.; and El-Yaniv, R. 2018. Deep anomaly detection using geometric transformations. In *Advances in Neural Information Processing Systems*, 9758–9769.

Goodfellow, I.; Bengio, Y.; Courville, A.; and Bengio, Y. 2016. *Deep learning*. MIT Press.

Grari, V.; Hajouji, O. E.; Lamprier, S.; and Detyniecki, M. 2021. Learning Unbiased Representations via Rényi Minimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 749–764. Springer.

Han, X.; and Guo, Y. 2021. Continual Learning with Dual Regularizations. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 619–634. Springer.

Haykin, S. 2010. *Neural networks and learning machines*, 3/E. Pearson Education India.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

- He, Y.; Liu, P.; Wang, Z.; Hu, Z.; and Yang, Y. 2019. Filter pruning via geometric median for deep convolutional neural networks acceleration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4340–4349.
- Hinton, G.; Deng, L.; Yu, D.; Dahl, G. E.; Mohamed, A.-r.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T. N.; et al. 2012a. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal processing magazine*, 29(6): 82–97.
- Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. R. 2012b. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.
- Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.
- Kim, Y.; Kim, T.-H.; and Ergün, T. 2015. The instability of the Pearson correlation coefficient in the presence of coincidental outliers. *Finance Research Letters*, 13: 243–257.
- Kondo, Y.; and Yamauchi, K. 2014. A dynamic pruning strategy for incremental learning on a budget. In *International Conference on Neural Information Processing*, 295–303. Springer.
- Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. *Technical report, University of Toronto*.
- Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 1097–1105.
- Kukačka, J.; Golkov, V.; and Cremers, D. 2017. Regularization for deep learning: A taxonomy. *arXiv preprint arXiv:1710.10686*.
- Kulesza, A.; and Taskar, B. 2010. Structured determinantal point processes. In *Advances in Neural Information Processing Systems*, 1171–1179.
- Kulesza, A.; and Taskar, B. 2012. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*.
- Kwok, J. T.; and Adams, R. P. 2012. Priors for diversity in generative latent variable models. In *Advances in Neural Information Processing Systems*, 2996–3004.
- Laakom, F.; Raitoharju, J.; Iosifidis, A.; and Gabbouj, M. 2021a. Learning distinct features helps, provably. *arXiv preprint arXiv:2106.06012*.
- Laakom, F.; Raitoharju, J.; Iosifidis, A.; and Gabbouj, M. 2021b. On Feature Diversity in Energy-based models. In *Energy Based Models Workshop-ICLR 2021*.
- Laakom, F.; Raitoharju, J.; Iosifidis, A.; and Gabbouj, M. 2022. Reducing Redundancy in the Bottleneck Representation of the Autoencoders. *arXiv preprint arXiv:2202.04629*.
- Lee, H. B.; Nam, T.; Yang, E.; and Hwang, S. J. 2019. Meta Dropout: Learning to Perturb Latent Features for Generalization. In *International Conference on Learning Representations*.
- Lee, S.; Heo, B.; Ha, J.-W.; and Song, B. C. 2020. Filter Pruning and Re-Initialization via Latent Space Clustering. *IEEE Access*, 8: 189587–189597.
- Lee-Thorp, J.; Ainslie, J.; Eckstein, I.; and Ontanon, S. 2021. FNet: Mixing Tokens with Fourier Transforms. *arXiv preprint arXiv:2105.03824*.
- Li, N.; Yu, Y.; and Zhou, Z.-H. 2012. Diversity regularized ensemble pruning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 330–345.
- Li, Z.; Gong, B.; and Yang, T. 2016. Improved dropout for shallow and deep learning. In *Advances in Neural Information Processing Systems*, 2523–2531.
- Liu, H.; Dai, Z.; So, D. R.; and Le, Q. V. 2021. Pay Attention to MLPs. *arXiv preprint arXiv:2105.08050*.
- Malkin, J.; and Bilmes, J. 2008. Ratio semi-definite classifiers. In *International Conference on Acoustics, Speech and Signal Processing*, 4113–4116.
- Malkin, J.; and Bilmes, J. 2009. Multi-layer ratio semi-definite classifiers. In *International Conference on Acoustics, Speech and Signal Processing*, 4465–4468.
- Nagarajan, V.; and Kolter, J. Z. 2019. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*, 11615–11626.
- Nakkiran, P.; Kaplun, G.; Bansal, Y.; Yang, T.; Barak, B.; and Sutskever, I. 2020. Deep Double Descent: Where Bigger Models and More Data Hurt. In *International Conference on Learning Representations*.
- Neyshabur, B.; Li, Z.; Bhojanapalli, S.; LeCun, Y.; and Srebro, N. 2018. The role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*.
- Ouali, Y.; Hudelot, C.; and Tami, M. 2021. Spatial contrastive learning for few-shot classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 671–686. Springer.
- Poggio, T.; Kawaguchi, K.; Liao, Q.; Miranda, B.; Rosasco, L.; Boix, X.; Hidary, J.; and Mhaskar, H. 2017. Theory of deep learning III: explaining the non-overfitting puzzle. *arXiv preprint arXiv:1801.00173*.
- Rodríguez, P.; Gonzalez, J.; Cucurull, G.; Gonfaus, J. M.; and Roca, X. 2016. Regularizing cnns with locally constrained decorrelations. *arXiv preprint arXiv:1611.01967*.
- Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252.
- Savas, C.; and Dosis, F. 2019. The impact of different kernel functions on the performance of scintillation detection based on support vector machines. *Sensors*, 19(23): 5219.
- Singh, P.; Verma, V. K.; Rai, P.; and Namboodiri, V. 2020. Leveraging filter correlations for deep model compression. In *The IEEE Winter Conference on Applications of Computer Vision*, 835–844.

- Tolstikhin, I.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Keysers, D.; Uszkoreit, J.; Lucic, M.; et al. 2021. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*.
- Wang, H.; Yang, W.; Zhao, Z.; Luo, T.; Wang, J.; and Tang, Y. 2019. Rademacher dropout: An adaptive dropout for deep neural network via optimizing generalization gap. *Neuro-computing*, 177–187.
- Xie, B.; Liang, Y.; and Song, L. 2017. Diverse neural network learns true target functions. In *Artificial Intelligence and Statistics*, 1216–1224.
- Xie, P.; Deng, Y.; and Xing, E. 2015a. Diversifying restricted boltzmann machine for document modeling. In *International Conference on Knowledge Discovery and Data Mining*, 1315–1324.
- Xie, P.; Deng, Y.; and Xing, E. 2015b. On the generalization error bounds of neural networks under diversity-inducing mutual angular regularization. *arXiv preprint arXiv:1511.07110*.
- Xie, P.; Singh, A.; and Xing, E. P. 2017. Uncorrelation and evenness: a new diversity-promoting regularizer. In *International Conference on Machine Learning*, 3811–3820.
- Xie, P.; Zhu, J.; and Xing, E. 2016. Diversity-promoting bayesian learning of latent variable models. In *International Conference on Machine Learning*, 59–68.
- Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500.
- Yang, K.; Gkatzelis, V.; and Stoyanovich, J. 2019. Balanced Ranking with Diversity Constraints. In *International Joint Conference on Artificial Intelligence*, 6035–6042.
- Yu, Y.; Li, Y.-F.; and Zhou, Z.-H. 2011. Diversity regularized machine. In *International Joint Conference on Artificial Intelligence*.
- Zagoruyko, S.; and Komodakis, N. 2016. Wide Residual Networks. In Richard C. Wilson, E. R. H.; and Smith, W. A. P., eds., *Proceedings of the British Machine Vision Conference (BMVC)*, 87.1–87.12. BMVA Press. ISBN 1-901725-59-6.
- Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; and Deny, S. 2021. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. *arXiv preprint arXiv:2103.03230*.
- Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*.
- Zhao, J.; Mathieu, M.; and LeCun, Y. 2017. Energy-based generative adversarial network. *International Conference on Learning Representations*.





# PUBLICATION

## III

**Efficient CNN with uncorrelated Bag of Features pooling**

F. Laakom, J. Raitoharju, A. Iosifidis and M. Gabbouj

*2022 IEEE Symposium Series on Computational Intelligence (SSCI)2022, 1082–1087*

DOI: 10.1109/SSCI51031.2022.10022157

© 2022 . IEEE. Reprinted, with permission, from Firas Laakom, et al. “*Efficient CNN with uncorrelated Bag of Features pooling*”, IEEE Symposium Series on Computational Intelligence, Dec 2022



# Efficient CNN with uncorrelated Bag of Features pooling

Firas Laakom

*Faculty of Information Technology and Communication Sciences  
Tampere University  
Tampere, Finland  
firas.laakom@tuni.fi*

Jenni Raitoharju

*Faculty of Information Technology  
University of Jyväskylä  
Jyväskylä, Finland  
jenni.k.raitoharju@jyu.fi*

Alexandros Iosifidis

*Department of Electrical and Computer Engineering  
Aarhus University  
Aarhus, Denmark  
ai@ece.au.dk*

Moncef Gabbouj

*Faculty of Information Technology and Communication Sciences  
Tampere University  
Tampere, Finland  
moncef.gabbouj@tuni.fi*

**Abstract**—Despite the superior performance of CNN, deploying them on low computational power devices is still limited as they are typically computationally expensive. One key cause of the high complexity is the connection between the convolution layers and the fully connected layers, which typically requires a high number of parameters. To alleviate this issue, Bag of Features (BoF) pooling has been recently proposed. BoF learns a dictionary, that is used to compile a histogram representation of the input. In this paper, we propose an approach that builds on top of BoF pooling to boost its efficiency by ensuring that the items of the learned dictionary are non-redundant. We propose an additional loss term, based on the pair-wise correlation of the items of the dictionary, which complements the standard loss to explicitly regularize the model to learn a more diverse and rich dictionary. The proposed strategy yields an efficient variant of BoF and further boosts its performance, without any additional parameters.

**Index Terms**—deep learning, CNN, diversity, bag of features pooling

## I. INTRODUCTION

In recent years, Convolutional Neural Networks (CNNs) have significantly advanced many tasks in the computer vision field due to their ability to learn ‘good’ feature representation in an end-to-end manner [1], [2]. However, despite their superior performance across multiple tasks, e.g., image classification [3]–[5], object detection [6]–[8], anomaly detection [9]–[12], deploying CNN-based solutions on low computational power devices, such as mobile phones, is still limited as most of the high-accuracy models are typically computationally expensive [1], [13], [14]. Thus, they are inefficient in terms of time and energy consumption [15]. To alleviate this issue, several approaches have been proposed to reduce the number of parameters required by a CNN model [16]–[20].

The standard CNN model is usually composed of two parts: The first part is formed of convolutional layers typically

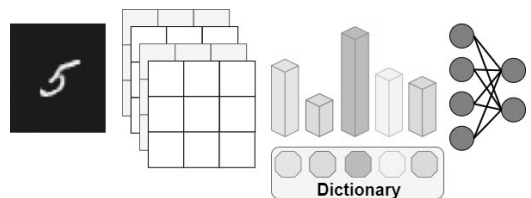


Fig. 1. An illustration of a simple BoF-based CNN model. From left to right: Input image, convolutional layer, BoF layer containing a dictionary and outputting a histogram, and fully connected layers.

coupled with max-pooling operations. Then, in the second part, fully connected layers are connected directly to a flattened version of the last convolutional layer output. This connection dramatically increases the total number of parameters, as convolutional layer outputs usually have high dimensionality. Recent approaches mitigate this problem by developing better mechanisms for connecting both parts, e.g., global average pooling and Bag of Features (BoF) pooling.

BoF pooling is a neural extension [18], [19] of the famous Bag-of-Visual Words [21]–[26]. An illustration of a simple BoF-based CNN model is presented in Figure 1. Based on the convolutional output, BoF pooling learns a codebook (dictionary) and outputs a shallow histogram representation of the input. The items of the dictionary are optimized in an end-to-end manner during the standard back-propagation. This yields powerful and efficient models, that achieve a high performance with a low computational footprint. Recently, CNN models, based on BoF pooling, have been used to solve multiple tasks [27]–[32], such as action recognition [30], information retrieval [33], and illumination estimation [34].

In this paper, we propose an approach that builds on top of BoF pooling to boost its efficiency by ensuring that the items of the learned dictionary are non-redundant. Forcing an

This work was supported by NSF-Business Finland Center for Visual and Decision Informatics (CVDI) project AMALIA.

uncorrelated structure on the codebook yields a more powerful model which can achieve a high performance with minimal dictionary size. Diversity in deep learning context has been shown to lead to better results [35]–[48]. To this end, we propose to augment the loss of the model to penalize pair-wise correlations between the items of the codebook. The proposed technique requires no additional parameters and can be incorporated in any BoF-based CNN model to boost the performance of the CNN model.

The contributions of this paper can be summarized as follows:

- We propose a scheme to avoid redundant items in the dictionary learned by the BoF.
- We propose to augment the CNN-loss to explicitly penalize the pair-wise correlations between codebook items and learn rich compressed dictionary.
- The proposed regularizer acts as an unsupervised regularizer on top of the BoF pooling layer and can be integrated into any BoF-based CNN model in a plug-and-play manner.
- The proposed approach is evaluated with three datasets. The results show a consistent performance boost compared to the standard approach.

The rest of this paper is organized as follows. First, provide a brief overview of BoF in Section II. In Section III, we present the proposed approach. In Section IV, we empirically evaluate the performance of our method on three different datasets. We conclude the paper in Section V.

## II. BAG-OF-FEATURES POOLING

In this section, we briefly describe the BoF pooling mechanism. BoF [18], [19] has been incorporated in a variety of applications and often led to superior results [31], [33], [34]. The BoF pooling is parameterized with a dictionary. Given an input, i.e., the output maps of the last convolutional layer, a histogram representation is compiled based on the dictionary. In the training phase, the items of the dictionary are optimized with the traditional back-propagation. The size of the dictionary is a hyper-parameter that can be adjusted with a validation set to avoid over-fitting.

BoF pooling is formed using two inner layers: a Radial Basis Function (RBF) layer that measures the similarity of the input features to the RBF centers and an accumulation layer that builds a histogram of the quantized feature vectors. Formally, let  $\mathbf{X}$  be the input image and  $\rho(\mathbf{X}) \in \mathbb{R}^{D \times P}$  the output of the convolutional layer, the RBF layer outputs a sequence of quantized representations:

$$\Psi = [\psi_1, \psi_2, \dots, \psi_P] \in \mathbb{R}^{K \times P},$$

where  $\psi_i$  is the representation corresponding to the  $i^{\text{th}}$  feature, i.e.,

$$\psi_i = [\psi_{i,1}, \dots, \psi_{i,K}].$$

The output of the  $i^{\text{th}}$  RBF unit is as follows:

$$\psi_{n,i} = \frac{\exp(-\|\rho(\mathbf{X})_n - \mathbf{c}_i\|/m_i)}{\sum_j \exp(-\|\rho(\mathbf{X})_n - \mathbf{c}_j\|/m_j)}, \quad (1)$$

where  $\mathbf{c}_i$  is the center of the  $i$ -th RBF neuron, and  $m_i$  is a scaling factor. The outputs of the  $P$  RBF neurons are accumulated in the next layer in order to obtain the final representation  $\Phi$  of each image:

$$\Phi = \frac{1}{P} \sum_j \psi_j. \quad (2)$$

To summarize, BoF receives as input a feature representation, usually in high dimension, and quantizes it into a fixed-size shallow histogram representation. The quantization is based on the inner dictionary,  $\{\mathbf{c}_1, \dots, \mathbf{c}_K\}$ , which can be learned jointly with the rest of the parameters in an end-to-end manner.

## III. OUR APPROACH

BoF pooling layer is a key technique that can be used in CNNs to construct powerful models with a low computational cost. The BoF relies on a dictionary, learned during the training, to compute its shallow output. In this paper, we propose an approach that builds on top of BoF pooling to boost its efficiency by explicitly forcing the items of the learned dictionary to be distinct and non-redundant. We propose a simple additional regularizer that penalizes the similarities between the codebook items. This can further boost the performance of the model, without any additional parameters.

The dictionary learned the BoF layer plays a critical role in the global performance of the model. Intuitively, Learning a diverse and rich dictionary yields in a robust codebook and increases the efficiency of the global model. Given a CNN model containing a BoF pooling layer with an inner dictionary  $\{\mathbf{c}_1, \dots, \mathbf{c}_K\}$  of size  $K$ , the similarity  $SIM$  between two elements  $\mathbf{c}_i$  and  $\mathbf{c}_j$  of this dictionary can be measured with the squared correlation:

$$SIM(\mathbf{c}_i, \mathbf{c}_j) = \left( \text{corr}(\mathbf{c}_i, \mathbf{c}_j) \right)^2, \quad (3)$$

where  $\text{corr}(\cdot, \cdot)$  is the correlation operator. We use the square to insure that the similarity is always positive.

Intuitively,  $SIM$  measures how similar two items are. Our goal is to regularize the similarities between the elements of the dictionary. So, the global similarity regularizer can be computed as the sum of the pair-wise similarities, i.e.,

$$\sum_{i \neq j} SIM(\mathbf{c}_i, \mathbf{c}_j). \quad (4)$$

Given the original loss  $L$ , e.g., least squares or cross entropy, we propose to regularize it as follows:

$$L_{new} \triangleq L + \beta \sum_{i \neq j} SIM(\mathbf{c}_i, \mathbf{c}_j), \quad (5)$$

where  $L_{new}$  is the augmented loss and  $\beta$  is a hyper-parameter employed to control the contribution of the supplementary regularizer in the global loss of the model. The computation of the total loss is illustrated in Figure 2.

Setting  $\beta = 0$  corresponds to the standard BoF case, while a higher  $\beta$  yields a loss dominated by the regularizer. In

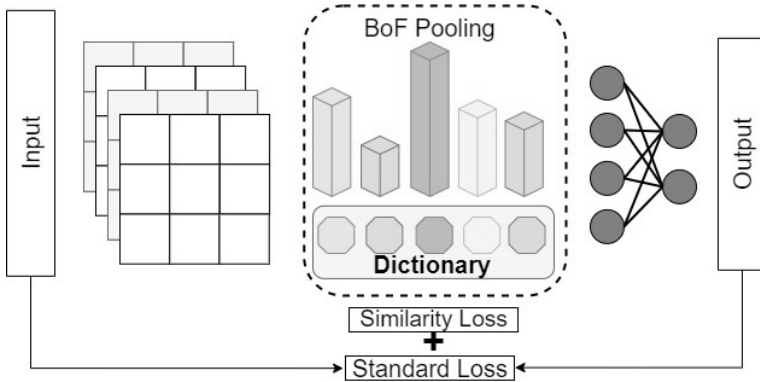


Fig. 2. An illustration on how the BoF-based CNN model loss is computed using our approach. The standard loss can be least squares or cross entropy and the similarity loss corresponds to the second term in (5).

the training phase, at each step in the back-propagation, the gradient of the loss w.r.t. the parameters is computed. The additional term depends only on the elements of the dictionary, i.e.,  $\{\mathbf{c}_1, \dots, \mathbf{c}_K\}$ , of the BoF layer. Thus, the gradient of all the model parameters, except  $\{\mathbf{c}_1, \dots, \mathbf{c}_K\}$ , remains the same as in the standard BoF case. For each of dictionary parameters  $\mathbf{c}_i$ , we have an additional feedback term equal to

$$\beta \frac{\partial \sum_{i \neq j} \text{SIM}(\mathbf{c}_i, \mathbf{c}_j)}{\partial \mathbf{c}_i},$$

which encourages different elements within this dictionary to be distinct.

The proposed approach affects only the training loss and does not require any additional parameters. It can be integrated in a plug-and-play manner in any BoF-based model to improve performance. Intuitively, the proposed schema acts as a penalty on top of the learned dictionary to provide supplementary feedback in the training phase to lessen the correlations of the codebook's items. By explicitly forcing the BoF layer to learn a diverse and rich dictionary, we can increase the model efficiency and reach a high performance with a minimal number of parameters.

#### IV. EXPERIMENTAL RESULTS

In this Section, we present the empirical results of the proposed approach along with the competing methods.

##### A. Experiment Setup

1) *Datasets*: We evaluate the performance of our approach using three different dataset:

- MNIST [49] is a dataset of  $28 \times 28$  images from 10 classes. It contains 50,000 samples for training and 10,000 for testing.
- fashionMNIST [50] is a clothes dataset containing  $28 \times 28$  images from 10 classes. It has in total 50,000 samples for training and 10,000 for testing.

- CIFAR10 [51] is an RGB image dataset containing  $32 \times 32$  images from a total of 10 distinct classes. It has a total of 50,000 and 10,000 samples for training and testing, respectively.

2) *Training & Testing*: In all our experiments, we hold 20% of the training data for validation and hyper-parameter selection. We also experiment with different values for the number of filters in the last convolutional layer. The full topology of the CNN models used in MNIST/fashionMNIST and CIFAR10 experiments are reported in Table I and Table II, respectively.

For MNIST and fashionMNIST experiments, all the models are trained for 50 epochs using Adam [52] regularizer with a 0.001 learning rate and a batch-size of 128. For CIFAR10 experiments, all the models are trained for 200 epochs with standard data augmentation [4] using Adam regularizer with a 0.0001 learning rate and a batch-size of 128.

We report the competitive results of the different pooling strategies, namely global max pooling (GMP) [15], global average pooling (GAP) [15], BoF [18], [19], and our approach. The size of the codebook is a hyperparameter for both BoF and our approach. It is optimized with the validation set from  $\{8, 16, 32, 64, 128\}$  for MNIST and fashionMNIST and from  $\{32, 64, 128, 256\}$  for CIFAR10. The hyper-parameter  $\beta$  in Eq. (5), used for controlling the contribution of the proposed regularizer in the global loss, is selected from  $\{0.1, 0.01, 0.001, 0.0001\}$  using the validation set in all experiments.

##### B. Empirical Results

In Table III and Table IV, we report the average error rates and standard deviations for the different filter sizes, i.e., C in Table I, on MNIST and fashionMNIST datasets, respectively. Compared to standard pooling approaches, i.e., GMP and GAP, we note that both variants of BoF consistently yield a better performance. For the 16 filter case, for example, GMP and GAP reach 3.63% and 4.67% errors on MNIST, respectively,

Input layer
$3 \times 3 \times 32$ - Relu
$3 \times 3 \times 32$ - Relu
$2 \times 2$ max pooling layer
$3 \times 3 \times C$ - Relu
Pooling strategy
Dropout (0.2)
512-Fully connected - Relu
Dropout (0.2)
10-fully connected
softmax layer

TABLE I

TOPOLOGY USED FOR MNIST AND FASHIONMNIST EXPERIMENTS. WE EXPERIMENT WITH DIFFERENT VALUES OF C, I.E., THE NUMBER OF FILTERS IN THE LAST CONVOLUTIONAL LAYER. POOLING STRATEGY REFERS TO THE USED METHOD, E.G., GLOBAL MAX POOLING OR BoF.

Input layer
$3 \times 3 \times 128$ - Relu
$3 \times 3 \times 128$ - Relu
$2 \times 2$ max pooling layer
$3 \times 3 \times 64$ - Relu
$3 \times 3 \times 64$ - Relu
$2 \times 2$ max pooling layer
$3 \times 3 \times C$ - Relu
Pooling strategy
Dropout (0.2)
512-Fully connected - Relu
Dropout (0.2)
10-fully connected
softmax layer

TABLE II

TOPOLOGY USED FOR CIFAR10 EXPERIMENTS. WE EXPERIMENT WITH DIFFERENT VALUES OF C, I.E., THE NUMBER OF FILTERS IN THE LAST CONVOLUTIONAL LAYER. POOLING STRATEGY REFERS TO THE USED METHOD, E.G., GLOBAL MAX POOLING OR BoF.

whereas standard BoF and our variant of BoF reach 1.03 and 1.00% for the same case, respectively.

For the fashionMNIST dataset with a 16-filter model, using BoF reduces the error rates by more than 5%. Compared to the standard BoF, we note that penalizing correlations between the items of the dictionary yields in a consistently better performance in most cases. For example, for fashionMNIST using a 128-filters model, our approach reaches only 8.77% error rate compared to 9.02% reached by BoF.

For MNIST dataset, as shown in Table III, the best performances reached by GMP and GAP pooling are 1.09% and 1.06%, respectively. In both cases, it is achieved by the 128-filters model. We note that our approach requires only 16 filters to achieve a better performance (1.00%). This finding is also in agreement with the results on fashionMNIST dataset in Table IV. In this case, our approach with only 16 filters achieves better results compared to the best GAP case and only 32 filters to achieve better results compared to the best GMP case. The best performance both on MNIST and fashionMNIST is achieved by our approach with the 64-filters and 128-filters model, respectively.

In Table V, we report the results of the different approaches on CIFAR10 dataset with three different filter sizes in the final convolutional layer, namely 32, 64 and 128 filters. As can be seen, the best result is 15.93% error rate which is

achieved by our approach using 128 filters. This constitutes an improvement by 2.87, 1.68%, and 0.17% compared to the best results achieved by GMP, GAP, and the standard BoF, respectively.

## V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a scheme that builds on top of BoF pooling to improve its performance. We proposed a regularizer, based on the pair-wise correlation of the items of the dictionary, which ensures the diversity and the richness of the learned dictionary within the BoF layer. It led to an efficient variant of BoF and further improved its capability, without any additional parameters. The proposed approach can be incorporated in any BoF-based model in a plug-and-play manner. Empirical results over three different dataset showed that the proposed regularizer boosts the performance of the model and led to lower error rates.

Future directions include more extensive experimental evaluation of the proposed approach over larger datasets and proposing more advanced techniques for quantifying the similarities between the codebook elements.

## REFERENCES

- [1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [2] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.
- [3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.
- [4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [5] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.
- [7] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu, "Object detection with deep learning: A review," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3212–3232, 2019.
- [8] Mingxing Tan, Ruoming Pang, and Quoc V Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10781–10790.
- [9] Guansong Pang, Longbing Cao, and Charu Aggarwal, "Deep learning for anomaly detection: Challenges, methods, and opportunities," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 2021, pp. 1127–1130.
- [10] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft, "Deep one-class classification," in *International conference on machine learning*. PMLR, 2018, pp. 4393–4402.
- [11] Hyunjong Park, Jongyoun Noh, and Bumsuh Ham, "Learning memory-guided normality for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14372–14381.
- [12] Taejoon Kim, Sang C Suh, Hyunjoon Kim, Jonghyun Kim, and Jinoh Kim, "An encoding technique for cnn-based network anomaly detection," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 2960–2965.
- [13] Jiasi Chen and Xukan Ran, "Deep learning with edge computing: A review," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1655–1674, 2019.

method	16 filters	32 filters	64 filters	128 filters
GMP	3.63 $\pm$ 0.31	1.97 $\pm$ 0.20	1.39 $\pm$ 0.07	<b>1.09 <math>\pm</math> 0.08</b>
GAP	4.67 $\pm$ 1.17	2.01 $\pm$ 0.09	1.31 $\pm$ 0.05	<b>1.06 <math>\pm</math> 0.02</b>
BoF	1.03 $\pm$ 0.08	<b>0.97 <math>\pm</math> 0.11</b>	1.00 $\pm$ 0.08	1.03 $\pm$ 0.06
BoF (ours)	1.00 $\pm$ 0.06	0.98 $\pm$ 0.06	<b>0.87 <math>\pm</math> 0.10</b>	0.98 $\pm$ 0.08

TABLE III

AVERAGE ERROR RATES AND STANDARD DEVIATION OF DIFFERENT APPROACHES FOR DIFFERENT NUMBER OF FILTERS IN THE LAST CONVOLUTIONAL LAYER ON THE MNIST DATASET. RESULTS ARE AVERAGED OVER 5 RANDOM SEEDS. TOP RESULTS FOR EACH APPROACH ARE IN BOLD AND BEST GLOBAL RESULT IS UNDERLINED.

method	16 filters	32 filters	64 filters	128 filters
GMP	14.94 $\pm$ 0.70	12.13 $\pm$ 0.30	10.46 $\pm$ 0.18	<b>9.48 <math>\pm</math> 0.12</b>
GAP	15.09 $\pm$ 0.19	12.30 $\pm$ 0.20	10.91 $\pm$ 0.21	<b>9.97 <math>\pm</math> 0.06</b>
BoF	9.55 $\pm$ 0.29	9.44 $\pm$ 0.25	9.04 $\pm$ 0.22	<b>9.02 <math>\pm</math> 0.15</b>
BoF (ours)	9.52 $\pm$ 0.29	9.14 $\pm$ 0.12	8.98 $\pm$ 0.18	<b>8.77 <math>\pm</math> 0.22</b>

TABLE IV

AVERAGE ERROR RATES AND STANDARD DEVIATION OF DIFFERENT APPROACHES FOR DIFFERENT NUMBER OF FILTERS IN THE LAST CONVOLUTIONAL LAYER ON THE FASHIONMNIST DATASET. RESULTS ARE AVERAGED OVER 5 RANDOM SEEDS. TOP RESULTS FOR EACH APPROACH ARE IN BOLD AND BEST GLOBAL RESULT IS UNDERLINED.

method	32 filters	64 filters	128 filters
GMP	22.31 $\pm$ 0.48	20.33 $\pm$ 0.67	<b>18.80 <math>\pm</math> 1.03</b>
GAP	20.94 $\pm$ 0.53	20.08 $\pm$ 0.99	<b>17.61 <math>\pm</math> 0.33</b>
BoF	17.15 $\pm$ 0.58	17.05 $\pm$ 0.11	<b>16.10 <math>\pm</math> 0.20</b>
BoF (ours)	17.21 $\pm$ 0.71	16.57 $\pm$ 0.25	<b>15.93 <math>\pm</math> 0.11</b>

TABLE V

AVERAGE ERROR RATES AND STANDARD DEVIATION OF DIFFERENT APPROACHES FOR DIFFERENT NUMBER OF FILTERS IN THE LAST CONVOLUTIONAL LAYER ON THE CIFAR10 DATASET. RESULTS ARE AVERAGED OVER THREE RANDOM SEEDS.

- [14] Firas Laakom, Jenni Raitoharju, Alexandros Iosifidis, Jarno Nikkanen, and Moncef Gabbouj, "Color constancy convolutional autoencoder," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2019, pp. 1085–1090.
- [15] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep learning*, MIT press, 2016.
- [16] Yuchao Li, Shaohui Lin, Baochang Zhang, Jianzhuang Liu, David Doermann, Yongjian Wu, Feiyue Huang, and Rongrong Ji, "Exploiting kernel sparsity and entropy for interpretable cnn compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [17] Shaohui Lin, Rongrong Ji, Xiaowei Guo, Xuelong Li, et al., "Towards convolutional neural networks compression via global error reconstruction," in *IJCAI*, 2016, pp. 1753–1759.
- [18] Nikolaos Passalis and Anastasios Tefas, "Neural bag-of-features learning," *Pattern Recognition*, pp. 277–294, 2017.
- [19] Nikolaos Passalis and Anastasios Tefas, "Learning bag-of-features pooling for deep convolutional neural networks," in *IEEE International Conference on Computer Vision*, 2017.
- [20] Gaurav Menghani, "Efficient deep learning: A survey on making deep learning models smaller, faster, and better," *arXiv preprint arXiv:2106.08962*, 2021.
- [21] Fei-Fei Li and Pietro Perona, "A bayesian hierarchical model for learning natural scene categories," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005, pp. 524–531.
- [22] G. Qiu, "Indexing chromatic and achromatic patterns for content-based colour image retrieval," *Pattern Recognition*, pp. 1675 – 1686, 2002.
- [23] Jun Yang, Yu-Gang Jiang, Alexander G Hauptmann, and Chong-Wah Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proceedings of the international workshop on Workshop on multimedia information retrieval*, 2007, pp. 197–206.
- [24] Gabriela Csukka and Florent Perronnin, "Fisher vectors: Beyond bag-of-visual-words image representations," in *International Conference on*

- Computer Vision, Imaging and Computer Graphics*. Springer, 2010, pp. 28–42.
- [25] Jyoti S Shukla, Kriti Rastogi, Hetal Patel, Gaurav Jain, and Shashikant Sharma, "Bag of visual words methodology in remote sensing—a review," in *Proceedings of the International e-Conference on Intelligent Systems and Signal processing*. Springer, 2022, pp. 475–486.
- [26] Yin Zhang, Rong Jin, and Zhi-Hua Zhou, "Understanding bag-of-words model: a statistical framework," *International journal of machine learning and cybernetics*, vol. 1, no. 1, pp. 43–52, 2010.
- [27] Nikolaos Passalis and Anastasios Tefas, "Training lightweight deep convolutional neural networks using bag-of-features pooling," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 6, pp. 1705–1715, 2018.
- [28] Kateryna Chumachenko, Alexandros Iosifidis, and Moncef Gabbouj, "Self-attention neural bag-of-features," *arXiv preprint arXiv:2201.11092*, 2022.
- [29] Nikolaos Passalis, Anastasios Tefas, Juho Kanninen, Moncef Gabbouj, and Alexandros Iosifidis, "Temporal bag-of-features learning for predicting mid price movements using high frequency limit order book data," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, no. 6, pp. 774–785, 2020.
- [30] Alexandros Iosifidis, Anastasios Tefas, and Ioannis Pitas, "Discriminant bag of words based representation for human action recognition," *Pattern Recognition Letters*, pp. 185 – 192, 2014.
- [31] Nikolaos Passalis, Jenni Raitoharju, Anastasios Tefas, and Moncef Gabbouj, "Efficient adaptive inference for deep convolutional neural networks using hierarchical early exits," *Pattern Recognition*, vol. 105, pp. 107346, 2020.
- [32] Dat Thanh Tran, Nikolaos Passalis, Anastasios Tefas, Moncef Gabbouj, and Alexandros Iosifidis, "Attention-based neural bag-of-features learning for sequence data," *arXiv preprint arXiv:2005.12250*, 2020.
- [33] Yu-Gang Jiang, Chong-Wah Ngo, and Jun Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *ACM International Conference on Image and Video Retrieval*, 2007, pp. 494–501.
- [34] Firas Laakom, Nikolaos Passalis, Jenni Raitoharju, Jarno Nikkanen, Anastasios Tefas, Alexandros Iosifidis, and Moncef Gabbouj, "Bag of color features for color constancy," *IEEE Transactions on Image Processing*, vol. 29, pp. 7722–7734, 2020.
- [35] Huanhuan Chen, *Diversity and regularization in neural network ensembles*, Ph.D. thesis, University of Birmingham, 2008.
- [36] Mehrdad J Gangeh, Ahmed K Farahat, Ali Ghodsi, and Mohamed S Kamel, "Supervised dictionary learning and sparse representation—a review," *arXiv preprint arXiv:1502.05928*, 2015.
- [37] K Rajesh and Atul Negi, "Heuristic based learning of parameters for dictionaries in sparse representations," in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2018, pp. 1013–1019.

- [38] Babajide O Ayinde, Tamer Inanc, and Jacek M Zurada, "Regularizing deep neural networks by enhancing diversity in feature extraction," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2650–2661, 2019.
- [39] Yashar Naderahmadian, Soosan Beheshti, and Mohammad Ali Tinati, "Correlation based online dictionary learning algorithm," *IEEE Transactions on signal processing*, vol. 64, no. 3, pp. 592–602, 2015.
- [40] Hien Van Nguyen, Vishal M Patel, Nasser M Nasrabadi, and Rama Chellappa, "Kernel dictionary learning," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 2021–2024.
- [41] Xiao-Yuan Jing, Rui-Min Hu, Fei Wu, Xi-Lin Chen, Qian Liu, and Yong-Fang Yao, "Uncorrelated multi-view discrimination dictionary learning for recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2014, vol. 28.
- [42] Jiayun Wang, Yubei Chen, Rudrasis Chakraborty, and Stella X Yu, "Orthogonal convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11505–11515.
- [43] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12310–12320.
- [44] Pengtao Xie, Aarti Singh, and Eric P. Xing, "Uncorrelation and evenness: a new diversity-promoting regularizer," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 3811–3820.
- [45] Michael Cogswell, Faruk Ahmed, Ross Girshick, Larry Zitnick, and Dhruv Batra, "Reducing overfitting in deep networks by decorrelating representations," *arXiv preprint arXiv:1511.06068*, 2015.
- [46] Firas Laakom, Jenni Raitoharju, Alexandros Iosifidis, and Moncef Gabbouj, "On feature diversity in energy-based models," in *Energy Based Models Workshop-ICLR 2021*, 2021.
- [47] Firas Laakom, Jenni Raitoharju, Alexandros Iosifidis, and Moncef Gabbouj, "Reducing redundancy in the bottleneck representation of the autoencoders," *arXiv preprint arXiv:2202.04629*, 2022.
- [48] Firas Laakom, Jenni Raitoharju, Alexandros Iosifidis, and Moncef Gabbouj, "Within-layer diversity reduces generalization gap," *arXiv preprint arXiv:2106.06012*, 2021.
- [49] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [50] Han Xiao, Kashif Rasul, and Roland Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.
- [51] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," 2009.
- [52] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.



# PUBLICATION

## IV

**Reducing redundancy in the bottleneck representation of the  
autoencoders**

F. Laakom, J. Raitoharju, A. Iosifidis and M. Gabbouj

*Pattern Recognition Letters* 178.(2024), 202–208

DOI: <https://doi.org/10.1016/j.patrec.2024.01.013>

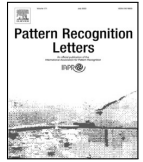
© 2024 . The Authors. Publication is licensed under a Creative  
Commons Attribution 4.0 International License CC-BY





Contents lists available at ScienceDirect

## Pattern Recognition Letters

journal homepage: [www.elsevier.com/locate/patrec](http://www.elsevier.com/locate/patrec)

## Reducing redundancy in the bottleneck representation of autoencoders

Firas Laakom <sup>a,\*</sup>, Jenni Raitoharju <sup>b,c</sup>, Alexandros Iosifidis <sup>d</sup>, Moncef Gabbouj <sup>a</sup><sup>a</sup> Faculty of Information Technology and Communication Sciences, Tampere University, Finland<sup>b</sup> Faculty of Information Technology, University of Jyväskylä, Finland<sup>c</sup> Quality of Information, Finnish Environment Institute, Finland<sup>d</sup> Department of Electrical and Computer Engineering, Aarhus University, Denmark

## ARTICLE INFO

Editor: Alexandru C. Telea

MSC:

68T01

68T10

94A08

68T99

Keywords:

Autoencoders

Unsupervised learning

Diversity

Feature representation

Dimensionality reduction

Image denoising

Image compression

## ABSTRACT

Autoencoders (AEs) are a type of unsupervised neural networks, which can be used to solve various tasks, e.g., dimensionality reduction, image compression, and image denoising. An AE has two goals: (i) compress the original input to a low-dimensional space at the bottleneck of the network topology using an encoder, (ii) reconstruct the input from the representation at the bottleneck using a decoder. Both encoder and decoder are optimized jointly by minimizing a distortion-based loss which implicitly forces the model to keep only the information in input data required to reconstruct them and to reduce redundancies. In this paper, we propose a scheme to explicitly penalize feature redundancies in the bottleneck representation. To this end, we propose an additional loss term, based on the pairwise covariances of the network units, which complements the data reconstruction loss forcing the encoder to learn a more diverse and richer representation of the input. We tested our approach across different tasks, namely dimensionality reduction, image compression, and image denoising. Experimental results show that the proposed loss leads consistently to superior performance compared to using the standard AE loss.

## 1. Introduction

With the progress of data gathering techniques, high-dimensional data are becoming available for training machine learning approaches. The impracticality of working in high dimensional spaces due to the *curse of dimensionality* and the understanding that the data in many problems reside on manifolds with much lower dimensions than those of the original space has led to the development of various approaches which try to learn a mapping of the data representations in the original space to more meaningful lower-dimensional representations.

Autoencoders (AEs) [1] are a powerful data-driven unsupervised approach used to learn a compact representation of a given input distribution. An autoencoder focuses solely on finding a low-dimensional representation, from which the input data can be reconstructed with minimal distortion. Autoencoders have been applied successfully in many tasks, such as transfer learning [2], anomaly detection [3], dimensionality reduction [4], and compression [5].

To accomplish these tasks, an autoencoder has two different parts: an encoder  $g(\cdot)$ , which maps an input  $x \in \mathcal{X}$  to a compact low-dimensional space  $g(x)$ , called the bottleneck representation, and a

decoder  $f(\cdot)$ , which takes the output of the encoder as its input and uses it to reconstruct the original input  $f \circ g(x)$ . Given a distortion metric  $D: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ , which measures the difference between the original input and the reconstructed input. Autoencoders are trained in an end-to-end manner using gradient descent-based optimization [1] to minimize the loss  $L$  defined as the average distortion over the training data  $\{x_i\}_{i=1}^N$ :

$$\min_{f,g} L(\{x_i\}_{i=1}^N) \triangleq \min_{f,g} \frac{1}{N} \sum_{i=1}^N D(x_i, f \circ g(x_i)). \quad (1)$$

Several extensions and regularization techniques have been proposed to augment this loss [5,6] aiming at improving the mapping of the input to a compressed representation at the bottleneck of the autoencoder so that the original inputs can be better reconstructed from these compact representations using the decoder.

By controlling the size of the bottleneck, one can explicitly control the dimensionality of the representation and the compression rate [5]. A low size of the bottleneck increases the complexity of the task of the decoder risking a higher distortion rate. This trade-off forces the model to keep only those variations in the input data that are required

\* Corresponding author.

E-mail addresses: [firmas.laakom@tuni.fi](mailto:firmas.laakom@tuni.fi) (F. Laakom), [jenni.k.raitoharju@jyu.fi](mailto:jenni.k.raitoharju@jyu.fi) (J. Raitoharju), [ai@ece.au.dk](mailto:ai@ece.au.dk) (A. Iosifidis), [moncef.gabbouj@tuni.fi](mailto:moncef.gabbouj@tuni.fi) (M. Gabbouj).<https://doi.org/10.1016/j.patrec.2024.01.013>

Received 28 November 2022; Received in revised form 5 January 2024; Accepted 10 January 2024

Available online 15 January 2024

0167-8655/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

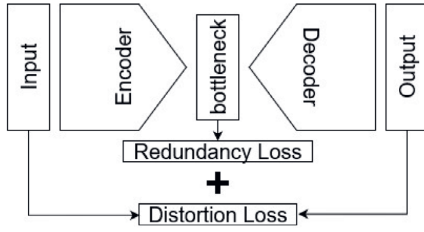


Fig. 1. An illustration of how the autoencoder loss is computed using our approach.

to reconstruct the input and to avoid redundancies and noise within the input. This is achieved implicitly by using error back-propagation for minimizing the reconstruction error, i.e., distortion  $D$ .

In the context of supervised neural networks, it has been shown that reducing redundancy improves generalization [7–9]. Approaches helping to reduce redundancy have been successfully applied, e.g., for pruning [10]. In this paper, we propose to model the feature redundancy in the bottleneck representation and minimize it explicitly. To this end, we propose augmenting the loss  $L$  using a redundancy term computed as the sum of the pairwise covariance between the bottleneck elements. The full scheme is illustrated in Fig. 1. We argue that explicitly penalizing the pairwise covariance between the different units in the bottleneck provides extra feedback for the encoder to avoid redundancy and to learn a richer representation of the input data.

The contributions of this paper can be summarized as follows:

- We propose a scheme to avoid redundant features in the bottleneck representation of autoencoders.
- We propose to augment the autoencoder loss to explicitly penalize the pairwise covariance between the features and learn a diverse compressed embedding of the training data.
- The proposed penalty term acts as an unsupervised regularizer on top of the encoder and can be integrated into any autoencoder-based model in a plug-and-play manner.
- The proposed method is extensively evaluated over three tasks: dimensionality reduction, image compression, and image denoising. Experimental results show consistent performance improvements compared to the standard approach.

The rest of this paper is organized as follows. Section 2 provides the background of autoencoders' training strategies and a brief review of different tasks considered in this work, i.e., dimensionality reduction, image compression, and image denoising. Section 3 describes the proposed approach. Section 4 reports experimental results for the dimensionality reduction task on the Madelon [11], ISOLET [12], and P53 Mutants [13] datasets. Section 5 reports experimental results for the image compression task on the MNIST [14] and CIFAR10 [15] datasets. Section 6 evaluates our approach on the image denoising task using the fashion MNIST [16] and CIFAR10 datasets. Section 7 concludes the paper.

## 2. Related work

Autoencoders are models trained to reconstruct their input, i.e., to approximate the identity function  $f(x) \approx x$ . While the identity function seems a particularly trivial function to learn, enforcing certain constraints on the network topology and particularly using a low number of units in the hidden layers [1] forces the model to learn to efficiently represent the data in a much lower-dimensional space compared to the original space [17,18]. This is a desired property in several tasks, e.g., dimensionality reduction [4], compression [5,19], and image denoising [20,21]. In [3,22–24], different extensions of autoencoders have been proposed to improve their performance in different contexts.

Several approaches based on reducing redundancy have been proposed recently in different contexts [25–29]. In particular, in the context of self-supervised learning, [26,28] proposed a training loss based on the pairwise correlation between the features of two perturbed variants of the same input. [9] proposed a data-dependent regularizer based on the  $L_2$  distance between the units outputs in the last hidden layer of CNNs and showed that such approach can reduce overfitting in the context of supervised learning. In this paper, we explore a similar direction in the context of unsupervised learning with autoencoders. To the best of our knowledge, this is the first work that considers reducing redundancy in this context. We show that it helps improve the performance.

Dimensionality reduction refers to the problem of learning a mapping from a high-dimensional input space  $\mathcal{X} \in \mathbb{R}^D$  into a lower-dimensional space  $\mathcal{Z} \in \mathbb{R}^d$ , where  $d \ll D$ , while preserving features of interest in the input data. Several linear [30–32] and non-linear [33–36] approaches have been proposed to solve this task. Some are supervised approaches, such as Linear Discriminant Analysis (LDA) and its extensions [37], others are unsupervised methods [38], such as Principal Component Analysis (PCA) [39]. Dimensionality reduction is the most straightforward application of autoencoders [4,40], as the mapping can be learned using an autoencoder by setting the size of the bottleneck to  $d$  units and training the model to reconstruct the input.

Image compression is an important task in many applications. Recent advances in deep neural networks [1] have enabled efficient modeling of high-dimensional data and led to outperforming traditional image compression techniques [41,42]. Recently, there has been interest in autoencoders to solve this task [5] due to their flexibility and easiness of training.

Image denoising [43] refers to the task of trying to restore a clean version of the image from its noisy corrupted counterpart. Due to their plug-and-play network architectures, CNN-based autoencoders have been widely adopted to solve this task [44,45]. In particular, an autoencoder is trained using pairs of noisy and clean images. By taking a noisy sample as input and setting its clean version as the target, the model learns to keep only the important information from the image and discard the noise.

## 3. Reducing the pairwise covariance within the bottleneck representation

Autoencoders are a special type of neural networks trained to achieve two objectives: (i) to compress an input into a low-dimensional space, (ii) to reconstruct the original input from the low-dimensional representation. This is achieved by minimizing the reconstruction loss over the training data, which implicitly forces learning a concise 'non-redundant' representation of the data. In this paper, we propose to augment the reconstruction loss with an additional term designed to explicitly minimize redundancy between the features learned at the bottleneck.

Given a training data set  $\{\mathbf{x}_i\}_{i=1}^N$  and an encoder  $g(\cdot) \in \mathbb{R}^D$ , the covariance between the  $i$ th and  $j$ th features,  $g_i$  and  $g_j$ , can be expressed as follows:

$$C(g_i, g_j) = \frac{1}{N} \sum_n \left( g_i(\mathbf{x}_n) - \mu_i \right) \left( g_j(\mathbf{x}_n) - \mu_j \right), \quad (2)$$

where  $\mu_i = \frac{1}{N} \sum_n g_i(\mathbf{x}_n)$  is the average output of the  $i$ th unit. Our aim is to minimize the redundancy of the bottleneck representations which corresponds to minimizing the pairwise covariance between different features. Thus, we augment the loss  $L(\{\mathbf{x}_i\}_{i=1}^N)$  as follows:

$$\begin{aligned} L(\{\mathbf{x}_i\}_{i=1}^N)_{aug} &\triangleq L(\{\mathbf{x}_i\}_{i=1}^N) + \alpha \sum_{i \neq j} C(g_i, g_j) \\ &= \frac{1}{N} \sum_{i=1}^N D(\mathbf{x}_i, f \circ g(\mathbf{x}_i)) \end{aligned} \quad (3)$$

**Table 1**

Statistics of the three datasets used in the dimensionality reduction experiments. # Dim: dimensionality of the data. # Train: number of training samples. # Test: number of test samples. d: projection dimension.

Dataset	# Dim	# Train	# Test	d
Madelon [11]	500	2000	1800	10
ISOLET [12]	617	6238	1559	10
P53 Mutants [13]	5408	21 811	9348	50

$$+ \alpha \sum_{i \neq j} \left( \frac{1}{N} \sum_n (g_i(\mathbf{x}_n) - \mu_i)(g_j(\mathbf{x}_n) - \mu_j) \right), \quad (4)$$

where  $\alpha$  is a hyper-parameter used to control the contribution of the additional term in the total loss.  $L_{aug}$  is composed of two terms, the first term is the standard autoencoder loss that depends on both the encoder and decoder parts to ensure that the autoencoder learns to reconstruct the input, while the second term depends only on the encoder and its aim is to promote the diversity of the learned features.

Intuitively, the proposed approach acts as an unsupervised regularizer on top of the encoder providing extra feedback during training to reduce the redundancy of the encoder's output. The proposed scheme can be embedded into any autoencoder-based model as a plug-in and optimized in a batch-manner, i.e., at each optimization step, we can compute the pairwise covariance using the mini-batch samples. Moreover, it is suitable for different learning strategies and different topologies.

#### 4. Experiments on dimensionality reduction

In this section, we consider the dimensionality reduction task using an autoencoder. We test the proposed approach using three different tabular datasets, namely Madelon [11], ISOLET [12], and P53 Mutants [13]. The Madelon dataset [11] contains samples represented by 500-dimensional vectors grouped in 32 clusters placed on the vertices of a five-dimensional hypercube. The ISOLET dataset [12] is composed of alphabet-speech data from 150 different subjects. Each instance is represented by a 617-feature vector compiled using spectral coefficients, contour features, sonorant features, pre-sonorant features, and post-sonorant features. The P53 Mutants dataset [13] is a large Biophysical dataset with more than 30k samples in total and 5408 attributes per instance. The feature representation is formed by combining the 2D electrostatic and surface-based attributes with the 3D distance-based attributes.

As the autoencoder topology, we use a simple architecture where the encoder maps the input using two intermediate fully-connected layers composed of 64 units with ReLU activation. Then, the bottleneck representation of size  $d$  is obtained using a fully-connected layer with  $d$  units and Leaky ReLU [1] activation. Symmetrically, the decoder is composed of two 64-dimensional fully-connected layers followed by ReLU activation and an output layer with the same size as the input using a sigmoid activation. The number of data dimensions, cardinalities of the training and test sets, and the value of  $d$  for each dataset is specified in Table 1. For training, we use the Adam optimizer with a learning rate of  $10^{-2}$  and the mean square error as the standard training loss  $L$ . The number of epochs and the batch size are set to 50 and 32, respectively, in all experiments. Each experiment is repeated 10 times and the mean and standard deviation of the root mean square error (RMSE) on the test set are reported.

In Table 2, we report the experimental results obtained by training the autoencoder using the standard loss and our proposed augmented loss and different values for the hyper-parameter  $\alpha$ , introduced in (4). It can be seen that, by explicitly penalizing redundancy in the bottleneck representations, the proposed approach consistently achieves lower errors compared to the standard approach on the three datasets. On the Madelon dataset, the best performance is achieved using  $\alpha = 0.005$ . On the ISOLET dataset, using  $\alpha = 0.1$  leads to the highest improvement,

**Table 2**

Reconstruction error on the three datasets used in the dimensionality reduction experiments (average and standard deviation over 10 repetitions).

	Madelon	ISOLET	P53 Mutants
Standard	0.14027 $\pm$ 0.00023	0.13143 $\pm$ 0.00259	0.02777 $\pm$ 0.00159
Ours (0.1)	0.14022 $\pm$ 0.00016	<b>0.12993 <math>\pm</math> 0.00283</b>	0.02717 $\pm$ 0.00087
Ours (0.05)	0.14024 $\pm$ 0.00038	0.13081 $\pm$ 0.00366	<b>0.02689 <math>\pm</math> 0.00054</b>
Ours (0.01)	0.14022 $\pm$ 0.00043	0.13101 $\pm$ 0.00204	0.02709 $\pm$ 0.00052
Ours (0.005)	<b>0.14005 <math>\pm</math> 0.00037</b>	0.13135 $\pm$ 0.00267	0.02694 $\pm$ 0.00051

whereas, on the P53 Mutants dataset, the best performance is achieved using  $\alpha = 0.05$ . It should be noted that while the performance gap is not large compared to the standard approach, the improvement is consistent on all the datasets and the different regularization rates.

In Fig. 2, we provide visualization results comparing the two approaches. We visualize the data in the projected space of the AE trained with the standard loss and the proposed augmented loss using t-SNE [35]. As can be seen, the AE trained with the augmented loss provides a more compact representation of the classes. We also note that by reducing redundancy, the learned embedding is more spread over the projection space and contains fewer empty regions.

Dimensionality reduction is typically applied as a pre-processing step to compile a compact feature representation that can be used to solve another task, such as classification. Intuitively, learning diverse and non-redundant features is crucial to achieve good performance on the task of interest. Here, to further assess the quality of the data representations learned using our approach, we conduct an extra experiment by applying the K-Nearest Neighbor (K-NN) classifier on top of the bottleneck features. In Table 3, we report the classification accuracy for  $K = 3$  and  $K = 5$ . As can be seen, the bottleneck features obtained using our approach yield consistently higher accuracy on the three datasets. For example, for the Madelon dataset, using  $\alpha = 0.005$  leads to 4.49% and 3.46% accuracy improvement compared to the standard approach when using  $K = 3$  and  $K = 5$ , respectively. It is interesting to note also that using the augmented loss consistently leads to more stable performance compared to the standard approach, as shown by the lower variances.

Moreover, for comparative purposes, we report results obtained by using WLD-reg [7], i.e., replacing the proposed regularizer with the direct variant of WLD-reg on top of the bottleneck representation.<sup>1</sup> Consistently with our results, WLD-reg also boosts performance compared to the standard approach showing that reducing redundancy in the bottleneck representation indeed helps to learn better features. Compared to our approach, we also note that the use of pairwise covariance instead of the  $L_2$  distance leads to higher performance on all three datasets.

To further study the effect of the number of dimensions at the bottleneck on performance, we conducted an additional experiment using the Isolet dataset. We plot the average accuracy over training the AE using 10 random seeds for different values of  $d$  in Fig. 2 (right). As can be seen, reducing redundancy improves performance for the different bottleneck sizes. It is interesting to note also that the performance gap is larger for small values of  $d$ . This can be explained by the fact that, when a smaller number of dimensions is used, it is more crucial to learn diverse features for solving the task.

#### 5. Experiments on image compression

In this section, we consider the image compression task using an autoencoder. We start by testing the proposed approach on the MNIST

<sup>1</sup> We note that WLD-reg [7] is a diversity promoting regularizer designed to be added on top of the last intermediate layer of a neural network in a standard supervised learning setting and not designed for unsupervised learning on top of the bottleneck of an autoencoder.

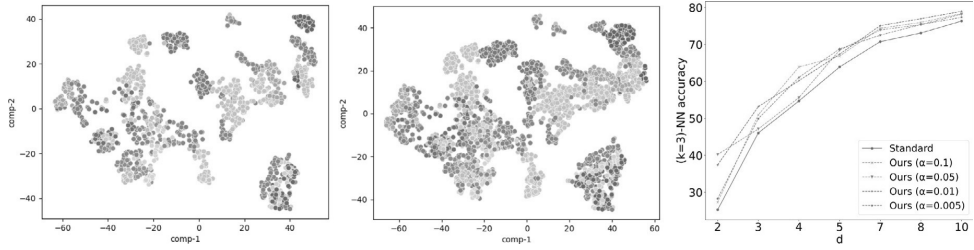


Fig. 2. t-SNE-based visualization of the ISOLET representations obtained by an AE trained by the standard approach (left) and the proposed approach (middle). Each color corresponds to data from a specific class. Average ( $K = 3$ )-NN accuracy as a function of the dimension of the bottleneck size  $d$  (right).

Table 3

Classification accuracy of Nearest Neighbor classifier applied on the bottleneck representations (average and standard deviation over 10 repetitions).

	Madelon ( $K = 3$ )-NN	( $K = 5$ )-NN
Standard	69.33% $\pm$ 2.71	71.32% $\pm$ 2.82
WLD-reg [7]	70.83% $\pm$ 2.08	72.45% $\pm$ 2.01
Ours (0.1)	72.51% $\pm$ 1.73	74.08% $\pm$ 1.63
Ours (0.05)	73.52% $\pm$ 1.91	74.53% $\pm$ 1.49
Ours (0.01)	72.65% $\pm$ 2.21	74.50% $\pm$ 1.87
Ours (0.005)	<b>73.82% <math>\pm</math> 1.83</b>	<b>74.78% <math>\pm</math> 1.78</b>
	ISOLET ( $K = 3$ )-NN	( $K = 5$ )-NN
Standard	76.32% $\pm$ 1.85	77.70% $\pm$ 1.60
WLD-reg [7]	78.22% $\pm$ 0.64	79.73% $\pm$ 0.70
Ours (0.1)	78.35% $\pm$ 0.46	79.82% $\pm$ 0.47
Ours (0.05)	78.18% $\pm$ 0.40	79.43% $\pm$ 0.44
Ours (0.01)	<b>78.96% <math>\pm</math> 0.56</b>	<b>79.83% <math>\pm</math> 0.54</b>
Ours (0.005)	77.34% $\pm$ 0.66	79.29% $\pm$ 0.66
	P53 Mutants ( $K = 3$ )-NN	( $K = 5$ )-NN
Standard	56.42% $\pm$ 0.60	54.99% $\pm$ 0.48
WLD-reg [7]	56.29% $\pm$ 0.36	54.86% $\pm$ 0.46
Ours (0.1)	<b>57.88% <math>\pm</math> 0.46</b>	<b>56.18% <math>\pm</math> 0.59</b>
Ours (0.05)	56.17% $\pm$ 0.46	55.39% $\pm$ 1.09
Ours (0.01)	57.22% $\pm$ 0.50	55.65% $\pm$ 0.46
Ours (0.005)	56.83% $\pm$ 0.41	55.92% $\pm$ 0.46

dataset [14]. It contains grayscale images with resolution of  $28 \times 28$  pixels, which are vectorized to form 784-dimensional vectors. The dataset is split in 50,000 training and 10,000 test images.

For the autoencoder model, we use a simple architecture. The encoder is composed of two fully-connected layers composed of 256 and 128 units, respectively. The final output of the encoder is composed of  $d$  units, where  $d$  is the size of the bottleneck. Similarly, the decoder part takes the encoder's output, maps it to an intermediate layer of 128 units, then 256 units, and outputs a 784-vector. In all the layers, we use ReLU activation except for the final layer, where sigmoid activation is used.

For training, we use the Adam optimizer with a learning rate of  $10^{-2}$  and the mean square loss as the standard training loss  $L$ . We train using 80% of the images in the original training set and hold the remaining 20% of the images as a validation set. During training, the model with the lowest mean square error on the validation set is saved and used in the test phase. We repeat each experiment five times and report the mean and standard deviation of the root-mean-square error (RMSE) errors, the peak signal-to-noise ratio (PSNR), and structural index similarity (SSIM) scores on the test set for the different approaches. We experiment with two different bottleneck sizes, i.e.,  $d = 256$  and  $d = 64$ . The results for different bottleneck sizes are reported in Table 4.

We note that the proposed approach consistently improves performance compared to training with the standard loss, i.e., it leads to lower RMSE values and higher PSNR and SSIM scores. For  $d = 256$ , the

Table 4

RMSE, PSNR, and SSIM on the MNIST dataset (average and standard deviation over 5 repetitions).

	RMSE $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$
	784 $\rightarrow$ 256		
Standard	0.0518 $\pm$ 0.0005	0.9631 $\pm$ 0.0016	26.43 $\pm$ 0.09
Ours (0.1)	0.0508 $\pm$ 0.0005	0.9641 $\pm$ 0.0010	26.57 $\pm$ 0.11
Ours (0.05)	0.0508 $\pm$ 0.0005	0.9636 $\pm$ 0.0007	26.58 $\pm$ 0.08
Ours (0.01)	0.0513 $\pm$ 0.0005	<b>0.9647 <math>\pm</math> 0.0013</b>	26.49 $\pm$ 0.09
Ours (0.005)	<b>0.0506 <math>\pm</math> 0.0007</b>	0.9635 $\pm$ 0.0012	<b>26.61 <math>\pm</math> 0.10</b>
	784 $\rightarrow$ 64		
Standard	0.0596 $\pm$ 0.0021	0.9597 $\pm$ 0.0022	25.25 $\pm$ 0.29
Ours (0.1)	<b>0.0584 <math>\pm</math> 0.0010</b>	<b>0.9607 <math>\pm</math> 0.0012</b>	<b>25.42 <math>\pm</math> 0.16</b>
Ours (0.05)	0.0588 $\pm$ 0.0018	0.9604 $\pm$ 0.0017	25.38 $\pm$ 0.25
Ours (0.01)	0.0593 $\pm$ 0.0010	0.9599 $\pm$ 0.0012	25.30 $\pm$ 0.15
Ours (0.005)	0.0588 $\pm$ 0.0009	0.9602 $\pm$ 0.0013	25.35 $\pm$ 0.13

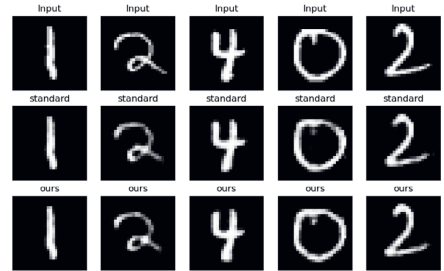


Fig. 3. Visualization of digits reconstructed by an AE trained by using the standard and the proposed training approaches. The first row contains the original inputs. Their reconstructed versions corresponding to the standard approach are shown in the second row, and the proposed approach in the third row.

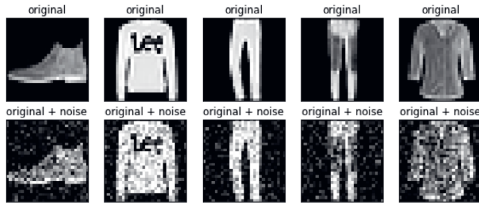
lowest RMSE value is achieved using  $\alpha = 0.005$ , and the highest PSNR and SSIM scores are obtained using  $\alpha = 0.01$  and  $\alpha = 0.005$ , respectively. For  $d = 64$ , using  $\alpha = 0.1$  leads to the best performance across all the metrics. Fig. 3 provides visualization results of images reconstructed from the representations learned using our approach for  $d = 64$ . We note that using the proposed augmented loss to train the AE leads to reconstructed inputs with lower distortion.

We also evaluate our approach on the image compression task with a more challenging dataset, namely the CIFAR10 Dataset [15]. The dataset contains  $32 \times 32$ -pixel color images, which are vectorized to form 3,072-dimensional vectors. For the model topology, we use two hidden layers with 512 and 256 units and ReLU activation. For the bottleneck size  $d$ , we experiment with two configurations,  $d = 128$  and  $d = 256$ . All the models are trained with Adam optimizer with a  $10^{-2}$  learning rate and a batch size of 128 for 50 epochs. The average and standard deviation of the different metrics over 10 random seeds are provided in Table 5. Similar to the results on the MNIST dataset,

Table 5

RMSE, PSNR, and SSIM on the CIFAR10 dataset (average and standard deviation over 5 repetitions).

	RMSE ↓	PSNR ↑	SSIM ↑
3072 → 256			
Standard	0.0888 ± 0.0022	0.6601 ± 0.0068	21.5547 ± 0.2470
Ours (0.01)	0.0882 ± 0.0012	0.6637 ± 0.0064	21.6168 ± 0.1314
Ours (0.005)	0.0882 ± 0.0006	0.6628 ± 0.0060	21.6271 ± 0.0593
Ours (0.001)	0.0882 ± 0.0011	0.6613 ± 0.0068	21.6347 ± 0.1154
Ours (0.0005)	<b>0.0877 ± 0.0011</b>	<b>0.6642 ± 0.0078</b>	<b>21.6829 ± 0.1156</b>
Ours (0.0001)	0.0885 ± 0.0014	0.6610 ± 0.0060	21.5927 ± 0.1518
3072 → 128			
Standard	0.0929 ± 0.0015	0.6151 ± 0.0100	21.2830 ± 0.1455
ours (0.01)	0.0920 ± 0.0010	0.6210 ± 0.0078	21.3765 ± 0.0920
ours (0.005)	0.0927 ± 0.0014	0.6144 ± 0.0089	21.3093 ± 0.1280
ours (0.001)	<b>0.0917 ± 0.0009</b>	<b>0.6246 ± 0.0054</b>	<b>21.4150 ± 0.0888</b>
ours (0.0005)	0.0923 ± 0.0016	0.6190 ± 0.0130	21.3436 ± 0.1527
ours (0.0001)	0.0926 ± 0.0019	0.6182 ± 0.0072	21.3184 ± 0.2070

Fig. 4. Original samples from the fashion MNIST dataset (top), and their noisy versions using  $\beta = 0.2$  (bottom).

we note that the proposed approach consistently leads to performance improvements. For  $d = 256$ , the best performance is achieved by using  $\alpha = 0.0005$ , whereas for  $d = 128$ ,  $\alpha = 0.001$  leads to the best performance.

## 6. Experiments on image denoising

In this section, we consider the image denoising task using an autoencoder. We test the proposed approach using the fashion MNIST dataset [16], which is an image dataset composed of 10 classes. Each sample is a  $28 \times 28$  gray-scale image. The dataset has a total of 60,000 training samples and 10,000 test samples. To construct a noisy dataset, we add a random noise from the normal distribution  $\beta \times \mathcal{N}(0, 1)$ , where  $\beta$  is a hyper-parameter controlling the noise rate. In Fig. 4, we provide examples of original images and their noisy versions.

As the autoencoder model, we use a simple CNN-based architecture. The encoder is composed of two convolutional layers, each of which has 16 and 4 filters, respectively, with kernel size  $3 \times 3$ . Symmetrically, the decoder is composed of two transposed convolutional layers of sizes 4 and 16 and a final convolutional layer with one filter with kernel size  $3 \times 3$ . All the layers have ReLU activation function except for the last layer where we use a sigmoid activation. Each model is trained for 50 epochs using the mean square error loss and Adam optimizer. We repeat each experiment five times and report the mean and standard deviation of RMSE, PSNR, and SSIM scores for different noise rates.

In Table 6, we report the experimental results for three different noise rates, i.e.,  $\beta = 0.1$ ,  $\beta = 0.2$ , and  $\beta = 0.4$ . Except for the hyper-parameter  $\alpha = 0.01$  with noise rates  $\beta = 0.2$  and  $\beta = 0.4$ , we note that our approach by explicitly minimizing the redundancy constantly outperforms the standard approach across all metrics. For the noise rate  $\beta = 0.1$ , the lowest RMSE value and the highest SSIM score are achieved using our approach with  $\alpha = 0.05$ , while the best PSNR is achieved with  $\alpha = 0.005$ . For  $\beta = 0.2$ , the best scores across all metrics correspond to  $\alpha = 0.005$ . For the extreme level of noise case, i.e.,  $\beta = 0.4$ , our approach

Table 6

RMSE, PSNR, and SSIM on the fashion MNIST dataset (average and standard deviation over 5 repetitions).

	RMSE ↓	PSNR ↑	SSIM ↑
$\beta = 0.1$			
Standard	0.0796 ± 0.0016	0.7980 ± 0.0061	22.51 ± 0.19
Ours (0.1)	0.0786 ± 0.0009	0.8018 ± 0.0024	22.64 ± 0.13
Ours (0.05)	<b>0.0772 ± 0.0018</b>	0.8049 ± 0.0062	<b>22.84 ± 0.25</b>
Ours (0.01)	0.0779 ± 0.0019	0.8047 ± 0.0066	22.77 ± 0.27
Ours (0.005)	0.0774 ± 0.0012	<b>0.8058 ± 0.0044</b>	22.82 ± 0.18
$\beta = 0.2$			
Standard	0.0941 ± 0.0026	0.7283 ± 0.0110	20.95 ± 0.25
Ours (0.1)	0.0934 ± 0.0021	0.7301 ± 0.0102	21.03 ± 0.19
Ours (0.05)	0.0933 ± 0.0020	0.7290 ± 0.0079	21.04 ± 0.19
Ours (0.01)	0.0975 ± 0.0034	0.7143 ± 0.1276	20.63 ± 0.31
Ours (0.005)	<b>0.0922 ± 0.0012</b>	<b>0.7357 ± 0.0058</b>	<b>21.14 ± 0.13</b>
$\beta = 0.4$			
Standard	0.1262 ± 0.0021	0.5901 ± 0.0089	18.27 ± 0.16
Ours (0.1)	<b>0.1258 ± 0.0021</b>	<b>0.5954 ± 0.0095</b>	<b>18.30 ± 0.15</b>
Ours (0.05)	0.1260 ± 0.0016	0.5946 ± 0.0067	18.28 ± 0.12
Ours (0.01)	0.1266 ± 0.0014	0.5865 ± 0.0070	18.22 ± 0.09
Ours (0.005)	0.1260 ± 0.0017	0.5911 ± 0.0085	18.28 ± 0.13

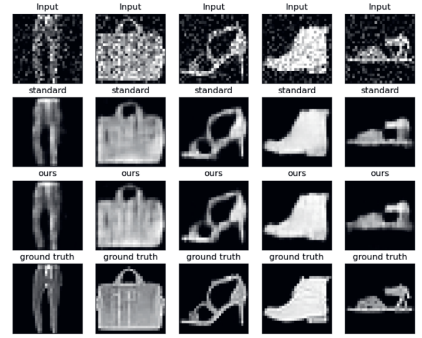


Fig. 5. Visualization of images denoised by AEs trained by using the standard and the proposed training approaches. The first row contains the original inputs. Their denoised versions corresponding to the standard approach are shown in the second row and the proposed approach in the third row. The last row contains the ground truth.

with  $\alpha = 0.1$  achieves the best performance across the three metrics. In Fig. 5, we present visual outputs for our approach. As shown, our approach learns to efficiently discard the noise from the input images.

Next, we evaluate the performance of the proposed approach in image denoising with a more challenging dataset, i.e., CIFAR10. We use the same model topology and experimental protocol used for this dataset in Section 5. We experiment with two levels of noise  $\beta = 0.1$  and  $\beta = 0.2$ . The results over 10 random seeds are presented in Table 7. As can be seen in Table 7, reducing features' redundancy in the bottleneck improves the performance of AE for both noise levels. For  $\beta = 0.1$ , using the augmented loss with  $\alpha = 0.005$  achieved the best performance, while for the high noise rate, i.e.,  $\beta = 0.2$ ,  $\alpha = 0.0001$  led to the best performance across the three metrics. With the same hardware configuration, the standard autoencoder average training time is on average 1,297.7 milliseconds per epoch, whereas using our approach takes on average 1,301.9 milliseconds per epoch. So adding our regularizer leads to performance improvement with less than 0.33% additional time cost.

## 7. Conclusion

In this paper, we proposed a scheme for modeling redundancies at the bottleneck of an autoencoder. We proposed to complement

Table 7

RMSE, PSNR, and SSIM on the CIFAR10 dataset (average and standard deviation over 5 repetitions).

	RMSE ↓	PSNR ↑	SSIM ↑
$\beta = 0.1$			
Standard	0.0954 ± 0.0019	0.6098 ± 0.0121	20.9243 ± 0.1734
Ours (0.01)	0.0948 ± 0.0016	0.6172 ± 0.0093	20.9703 ± 0.1550
Ours (0.005)	<b>0.0940 ± 0.0010</b>	<b>0.6227 ± 0.0056</b>	<b>21.0411 ± 0.1021</b>
Ours (0.001)	0.0952 ± 0.0018	0.6129 ± 0.0116	20.9325 ± 0.1651
Ours (0.0005)	0.0943 ± 0.0012	0.6190 ± 0.0085	21.0238 ± 0.1097
Ours (0.0001)	0.09489 ± 0.0012	0.6157 ± 0.0072	20.9644 ± 0.1102
$\beta = 0.2$			
Standard	0.1001 ± 0.0012	0.5798 ± 0.0081	20.4497 ± 0.0972
Ours (0.01)	0.0996 ± 0.0013	0.5846 ± 0.0089	20.4900 ± 0.1155
Ours (0.005)	0.1000 ± 0.0015	0.5806 ± 0.0104	20.4597 ± 0.1118
Ours (0.001)	0.0999 ± 0.0014	0.5824 ± 0.0090	20.4626 ± 0.1118
Ours (0.0005)	0.0997 ± 0.0015	0.5814 ± 0.0111	20.4881 ± 0.1206
Ours (0.0001)	<b>0.0992 ± 0.0015</b>	<b>0.5884 ± 0.0081</b>	<b>20.5186 ± 0.1370</b>

the training loss with an extra regularization term, which explicitly penalizes the pairwise covariances of the units at the encoder's output and, thus, forces it to learn more diverse and compact representations for the input samples. The proposed approach can be interpreted as an unsupervised regularizer on top of the encoder and can be integrated into any autoencoder-based model in a plug-and-play manner. We empirically demonstrated the effectiveness of our approach across multiple tasks, namely dimensionality reduction, compression, and denoising. We showed that it improves performance compared to the standard approach, with minimal training time cost increase. Even though the proposed regularizer consistently improves the performance of autoencoders, its key limitation is the marginal improvement in certain tasks, as shown in the results, e.g., Table 2. Future directions include proposing more efficient redundancy modeling techniques to further improve the performance of autoencoders and exploring redundancy reduction strategies for variational autoencoders.

#### CRedit authorship contribution statement

**Firas Laakom:** Conceptualization, Methodology, Writing – original draft. **Jenni Raitoharju:** Supervision, Writing – review & editing. **Alexandros Iosifidis:** Supervision, Writing – review & editing. **Moncef Gabbouj:** Supervision, Writing – review & editing.

#### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

#### Data availability

Data will be made available on request.

#### Acknowledgments

This work has been supported by the Academy of Finland Awcha project DN 334566 and NSF-Business Finland Center for Big Learning project AMALIA. The work of Jenni Raitoharju was supported by the Academy of Finland (projects 324475 and 333497).

#### References

- [1] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, *Deep Learning*, MIT Press, 2016.
- [2] F. Zhuang, X. Cheng, P. Luo, S.J. Pan, Q. He, Supervised representation learning: Transfer learning with deep autoencoders, in: *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

- [3] C. Zhou, R.C. Paffenroth, Anomaly detection with robust deep autoencoders, in: *23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.
- [4] S. Petschmann, M. Lux, S. Chatzichristofis, Dimensionality reduction for image features using deep learning and autoencoders, in: *The 15th International Workshop on Content-Based Multimedia Indexing*, 2017.
- [5] L. Theis, W. Shi, A. Cunningham, F. Huszár, Lossy image compression with compressive autoencoders, 2017, arXiv preprint arXiv:1703.00395.
- [6] G.D. Cavalcanti, L.S. Oliveira, T.J. Moura, G.V. Carvalho, Combining diversity measures for ensemble pruning, *Pattern Recognit. Lett.* (2016).
- [7] F. Laakom, J. Raitoharju, A. Iosifidis, M. Gabbouj, WLD-reg: A data-dependent within-layer diversity regularizer, in: *the 37th AAAI Conference on Artificial Intelligence*, 2023.
- [8] M. Cogswell, F. Ahmed, R.B. Girshick, L. Zitnick, D. Batra, Reducing overfitting in deep networks by decorrelating representations, in: *International Conference on Learning Representations*, 2016.
- [9] F. Laakom, J. Raitoharju, A. Iosifidis, M. Gabbouj, On feature diversity in energy-based models, in: *Energy Based Models Workshop-ICLR*, 2021.
- [10] H. Ide, T. Kobayashi, K. Watanabe, T. Kurita, Robust pruning for efficient CNNs, *Pattern Recognit. Lett.* (2020).
- [11] I. Guyon, M. Madelon, 2008, UCI Machine Learning Repository.
- [12] R. Cole, M. Fenty, ISOLET, 1994, UCI Machine Learning Repository.
- [13] R. Lathrop, p53 Mutants, 2010, UCI Machine Learning Repository.
- [14] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE* (1998).
- [15] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images, Technical report, University of Toronto, 2009.
- [16] H. Xiao, K. Rasul, R. Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017, arXiv preprint arXiv:1708.07747.
- [17] J. Guo, X. Yuan, P. Xu, H. Bai, B. Liu, Improved image clustering with deep semantic embedding, *Pattern Recognit. Lett.* (2020).
- [18] Y. Sang, J. Sang, M.S. Alam, Image encryption based on logistic chaotic systems and deep autoencoder, *Pattern Recognit. Lett.* (2022).
- [19] A. Golinski, R. Pourreza, Y. Yang, G. Sautiere, T.S. Cohen, Feedback recurrent autoencoder for video compression, in: *Asian Conference on Computer Vision*, 2020.
- [20] X. Ye, L. Wang, H. Xing, L. Huang, Denoising hybrid noises in image with stacked autoencoder, in: *2015 IEEE International Conference on Information and Automation, IEEE*, 2015.
- [21] L. Gondara, Medical image denoising using convolutional denoising autoencoders, in: *2016 IEEE 16th International Conference on Data Mining Workshops, ICDMW, IEEE*, 2016.
- [22] M. Patacchiola, P. Fox-Roberts, E. Rosten, Y-autoencoders: Disentangling latent representations via sequential encoding, *Pattern Recognit. Lett.* (2020).
- [23] J. Deng, Z. Zhang, E. Marchi, B. Schuller, Sparse autoencoder-based feature transfer learning for speech emotion recognition, in: *Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013.
- [24] P. Baldi, Autoencoders, unsupervised learning, and deep architectures, in: *ICML Workshop on Unsupervised and Transfer Learning, JMLR Workshop and Conference Proceedings*, 2012.
- [25] A. Jeffares, T. Liu, J. Crabbé, F. Imrie, M. van der Schaar, TANGOS: Regularizing tabular neural networks through gradient orthogonalization and specialization, 2023, arXiv preprint arXiv:2303.05506.
- [26] J. Zbontar, L. Jing, I. Misra, Y. LeCun, S. Deny, Barlow twins: Self-supervised learning via redundancy reduction, in: *The 38th International Conference on Machine Learning*, 2021.
- [27] F. Laakom, J. Raitoharju, A. Iosifidis, M. Gabbouj, Efficient CNN with uncorrelated bag of features pooling, in: *2022 IEEE Symposium Series on Computational Intelligence, SSCI, IEEE*, 2022.
- [28] A. Bardes, J. Ponce, Y. LeCun, Vicreg: Variance-invariance-covariance regularization for self-supervised learning, 2021, arXiv preprint arXiv:2105.04906.
- [29] F. Laakom, J. Raitoharju, A. Iosifidis, M. Gabbouj, Learning distinct features helps, provably, in: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer*, 2023.
- [30] W. Zhao, R. Chellappa, P.J. Phillips, Subspace Linear Discriminant Analysis for Face Recognition, *Citeseer*, 1999.
- [31] Y. Koren, L. Carmel, Robust linear dimensionality reduction, *IEEE Trans. Vis. Comput. Graph.* (2004).
- [32] F. Laakom, J. Raitoharju, N. Passalis, A. Iosifidis, M. Gabbouj, Graph embedding with data uncertainty, *IEEE Access* (2022).
- [33] D. DeMers, G.W. Cottrell, Non-linear dimensionality reduction, in: *Advances in Neural Information Processing Systems*, Citeseer, 1993.
- [34] Y.-R. Yeh, S.-Y. Huang, Y.-J. Lee, Nonlinear dimension reduction with kernel sliced inverse regression, *IEEE Trans. Knowl. Data Eng.* (2008).
- [35] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.* (2008).
- [36] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, 2018, arXiv preprint arXiv:1802.03426.



- [37] A. Iosifidis, A. Tefas, I. Pitas, On the optimal class representation in linear discriminant analysis, *IEEE Trans. Neural Netw. Learn. Syst.* (2013).
- [38] A.C. Kumar, Analysis of unsupervised dimensionality reduction techniques, *Comput. Sci. Inf. Syst.* (2009).
- [39] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, *Chemometr. Intell. Laboratory Syst.* (1987).
- [40] S.A. Thomas, A.M. Race, R.T. Steven, I.S. Gilmore, J. Bunch, Dimensionality reduction of mass spectrometry imaging data using autoencoders, in: *IEEE Symposium Series on Computational Intelligence, SSCI*, 2016.
- [41] G. Toderici, D. Vincent, N. Johnston, S. Jin Hwang, D. Minnen, J. Shor, M. Covell, Full resolution image compression with recurrent neural networks, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [42] J. Ballé, V. Laparra, E.P. Simoncelli, End-to-end optimization of nonlinear transform codes for perceptual quality, in: *2016 Picture Coding Symposium, PCS*, IEEE, 2016.
- [43] K. Gupta, S. Gupta, Image denoising techniques-a review paper, *IJITEE* (2013).
- [44] C. Tian, L. Fei, W. Zheng, Y. Xu, W. Zuo, C.-W. Lin, Deep learning on image denoising: An overview, *Neural Netw.* (2020).
- [45] J. Garcia-Gonzalez, J.M. Ortiz-de Laza-Lobato, R.M. Luque-Baena, M.A. Molina-Cabello, E. López-Rubio, Foreground detection by probabilistic modeling of the features discovered by stacked denoising autoencoders in noisy video sequences, *Pattern Recognit. Lett.* (2019).



# PUBLICATION

## V

### **On Feature Diversity in Energy-based models**

F. Laakom, J. Raitoharju, A. Iosifidis and M. Gabbouj

*Energy Based Models Workshop-ICLR2021*

© 2021 . Publication reprinted with the permission of the  
copyright holders



## ON FEATURE DIVERSITY IN ENERGY-BASED MODELS

**Firas Laakom**Faculty of Information Technology  
Tampere University  
Tampere, Finland  
firas.laakom@tuni.fi**Jenni Raitoharju**Programme for Environmental Information  
Finnish Environment Institute  
Jyväskylä, Finland  
jenni.raitoharju@syke.fi**Alexandros Iosifidis**Department of Electrical and Computer Engineering  
Aarhus University  
Aarhus, Denmark  
ai@ece.au.dk**Moncef Gabbouj**Faculty of Information Technology  
Tampere University  
Tampere, Finland  
moncef.gabbouj@tuni.fi

## ABSTRACT

Energy-based learning is a powerful learning paradigm that encapsulates various discriminative and generative approaches. An energy-based model (EBM) is typically formed of one (or many) inner-models which learn a combination of the different features to generate an energy mapping for each input configuration. In this paper, we focus on the diversity of the produced feature set. We extend the probably approximately correct (PAC) theory of EBMs and analyze the effect of the diversity on the performance of EBMs. We derive generalization bounds for various learning contexts, i.e., regression, classification, and implicit regression, with different energy functions and we show that indeed increasing the diversity of the feature set can consistently decrease the gap between the true and empirical expectation of the energy and boosts the performance of the model.

## 1 INTRODUCTION

The energy-based learning paradigm was first proposed by LeCun et al. (2006) as an alternative to probabilistic graphical models (Koller & Friedman, 2009). As their name suggests, energy-based models (EBMs) map each input ‘configuration’ to a single scalar, called the ‘energy’. In the learning phase, the parameters of the model are optimized to associate the desired configurations with small energy values and the undesired ones with higher energy values (Kumar et al., 2019; Song & Ermon, 2019; Yu et al., 2020; Nash & Durkan, 2019; Meng et al., 2020; Arbel et al., 2021). In the inference phase, given an incomplete input configuration, the energy surface is explored to find the remaining variables which yield the lowest energy. EBMs encapsulate solutions to several supervised (LeCun et al., 2006; Fang & Liu, 2016) and unsupervised learning problems (Ranzato et al., 2007b; Haarnoja et al., 2017; Parshakova et al., 2019; Deng et al., 2020; Bakhtin et al., 2021) and provide a common theoretical framework for many learning models, including traditional discriminative (Zhai et al., 2016; Grathwohl et al., 2019; Li et al., 2020; LeCun et al., 2006; Teh et al., 2003) and generative (Zhao et al., 2016; Dai et al., 2017; Ranzato et al., 2007a; Che et al., 2020; Khalifa et al., 2020; Arbel et al., 2021) approaches.

Formally, let us denote the energy function by  $E(W, \mathbf{X}, \mathbf{Y})$ , where  $W$  represents the model parameters to be optimized during training and  $\mathbf{X}, \mathbf{Y}$  are sets of variables. Figure 1 illustrates how classification, regression, and implicit regression can be expressed as EBMs. In Figure 1 (a), a regression scenario is presented. The input  $\mathbf{X}$ , e.g., an image, is transformed using an inner model  $G_W(\mathbf{X})$  and its distance,  $D$ , to the second input  $\mathbf{Y}$  is computed yielding the energy function. A valid energy function in this case can be the  $L_1$  or the  $L_2$  distance. In the binary classification case (Figure 1 (b)), the energy can be defined as  $E(W, \mathbf{X}, \mathbf{Y}) = -\mathbf{Y}G_W(\mathbf{X})$ . In the inference phase, given an input  $\mathbf{X}$ , the label  $\mathbf{Y}^*$  can be obtained by solving the following optimization problem:

$$\mathbf{Y}^* = \arg \min_{\mathbf{Y}} E(W, \mathbf{X}, \mathbf{Y}). \quad (1)$$

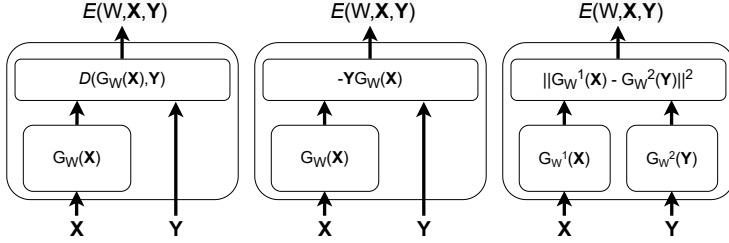


Figure 1: An illustration of energy-based models used to solve (a) a regression problem (b) a binary classification problem (c) an implicit regression problem.

An EBM typically relies on an inner model, i.e.,  $G_w(\mathbf{X})$ , to generate the desired energy landscape (LeCun et al., 2006). Depending on the problem at hand, this function can be constructed as a linear projection, a kernel method, or a neural network and its parameters are optimized in a data-driven manner in the training phase. Formally,  $G_w(\mathbf{X})$  can be written as

$$G_W(\mathbf{X}) = \sum_i^D w_i \phi_i(\mathbf{X}), \quad (2)$$

where  $\{\phi_1(\cdot), \dots, \phi_D(\cdot)\}$  is the feature set, which can be hand-crafted, separately trained from unlabeled data (Zhang & LeCun, 2017), or modeled by a neural network and optimized in the training phase of the EBM model (Du & Mordatch, 2019). In the rest of the paper, we assume that the inner models  $G_W$  defined in the energy-based learning system (Figure 1) are obtained as a weighted sum of different features as expressed in equation 2.

In (Zhang, 2013), it was shown that simply minimizing the empirical energy over the training data does not theoretically guarantee the minimization of the expected value of the true energy. Thus, developing and motivating novel regularization techniques is required (Zhang & LeCun, 2017). We argue that the quality of this feature set, i.e.,  $\{\phi_1(\cdot), \dots, \phi_D(\cdot)\}$ , plays a critical role in the overall performance of the global model. In this work, we extend the theoretical analysis of (Zhang, 2013) and focus on the ‘diversity’ of this set and its effect on the generalization ability of the EBM models. Intuitively, it is clear that a less correlated set of intermediate representations is richer and thus able to capture more complex patterns in the input. Thus, it is important to avoid redundant features for achieving a better performance. However, a theoretical analysis is missing. We start by quantifying the diversity of a set. To this end, we introduce  $\vartheta$ -diversity:

**Definition 1.** ( $\vartheta$ -diversity) A set of feature functions,  $\{\phi_1(\cdot), \dots, \phi_D(\cdot)\}$  is called  $\vartheta$ -diverse, if there exists a constant  $\vartheta \in \mathbb{R}$ , such that for every input  $\mathbf{X}$  we have

$$\sum_{i \neq j}^D (\phi_i(\mathbf{X}) - \phi_j(\mathbf{X}))^2 > \vartheta \quad (3)$$

with a high probability  $\tau$ .

Intuitively, if two feature maps  $\phi_i(\cdot)$  and  $\phi_j(\cdot)$  are different, then with high probability they have different outputs for the same input. However, if for example the features are extracted using a neural network with a ReLU activation function, then there is a high probability that some of the features associated with the input will be zero. Thus, defining a lower bound for the pair-wise diversity directly is impractical. To this end, we quantify diversity as the lower-bound over the sum of the pair-wise distances of the feature maps as expressed in equation 3.  $\vartheta$  measures the diversity of a set.

In machine learning context, diversity has been explored in ensemble learning (Li et al., 2012; Yu et al., 2011), sampling (Derezinski et al., 2019; Bıyık et al., 2019; Gartrell et al., 2019), ranking (Yang et al., 2019; Gan et al., 2020), pruning (Kondo & Yamauchi, 2014; He et al., 2019; Singh et al., 2020; Lee et al., 2020), and neural networks (Xie et al., 2015; 2017). In Xie et al. (2015; 2017), it was shown theoretically and experimentally that employing a diversity strategy over the weights of a neural network using the mutual angles improves the generalization ability of the

model. In this work, we explore a new line of research, where diversity is defined over the feature maps directly, using the  $\vartheta$ -diversity, in the context of energy-based learning. We theoretically study the generalization ability of EBMs in different learning contexts, i.e., regression, classification, implicit regression, and we derive new generalization bounds using the  $\vartheta$ -diversity providing theoretical guarantees that a diverse set of features indeed improves the generalization ability of the model. The contributions of this paper can be summarized as follows:

- We explore a new line of research, where diversity is defined over the features representing the input data and not over the model’s parameters. To this end, we introduce  $\vartheta$ -diversity as a quantification of the diversity of a given feature set.
- We extend the theoretical analysis (Zhang, 2013) and study the effect of the diversity of the feature set on the generalization of the energy-based models (EBMs).
- We derive approximation bounds for the expectation of the true energy in different learning contexts, i.e., regression, classification, and implicit regression, using different energy functions. Our analysis consistently shows that increasing the diversity of the feature set can boost the performance of an energy based model.

## 2 PAC-LEARNING OF EBMS WITH $\vartheta$ -DIVERSITY

In this section, we derive a qualitative justification for  $\vartheta$ -diversity using probably approximately correct (PAC) learning (Valiant, 1984). The PAC-based theory for standard energy based models has been established in (Zhang, 2013). Based on the Rademacher complexity (Bartlett & Mendelson, 2002), several EBMs learning guarantees have been shown. In Lemma 1, we present the principal PAC-learning bound for energy functions with finite outputs.

**Definition 2.** (Bartlett & Mendelson, 2002) *For a given dataset with  $m$  samples  $\mathcal{S} = \{\mathbf{x}_i, y_i\}_{i=1}^m$  generated by a distribution  $\mathcal{D}$  and for a model space  $\mathcal{F} : \mathcal{X} \rightarrow \mathbb{R}$  with a single dimensional output, the empirical Rademacher complexity  $\mathcal{R}_m(\mathcal{F})$  of the set  $\mathcal{F}$  is defined as follows:*

$$\mathcal{R}_m(\mathcal{F}) = \mathbb{E}_{\sigma} \left[ \sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^N \sigma_i f(\mathbf{x}_i) \right], \quad (4)$$

where the Rademacher variables  $\sigma = \{\sigma_1, \dots, \sigma_N\}$  are independent uniform random variables in  $\{-1, 1\}$ .

**Lemma 1.** (Zhang, 2013) *For a well-defined energy function  $E(h, \mathbf{x}, \mathbf{y})$  over hypothesis class  $\mathcal{H}$ , input set  $\mathcal{X}$  and output set  $\mathcal{Y}$  (LeCun et al., 2006), the following holds for all  $h$  in  $\mathcal{H}$  with a probability of at least  $1 - \delta$*

$$\mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}} [E(h, \mathbf{x}, \mathbf{y})] \leq \frac{1}{m} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{S}} E(h, \mathbf{x}, \mathbf{y}) + 2\mathcal{R}_m(\mathcal{E}) + M \sqrt{\frac{\log(2/\delta)}{2m}}, \quad (5)$$

where  $\mathcal{E}$  is the energy function class defined as  $\mathcal{E} = \{E(h, \mathbf{x}, \mathbf{y}) | h \in \mathcal{H}\}$ ,  $\mathcal{R}_m(\mathcal{E})$  is its Rademacher complexity, and  $M$  is the upper bound of  $\mathcal{E}$ .

Lemma 1 provides a generalization bound for energy-based models with well-defined (non-negative) and bounded energy. The expected energy is bounded using the sum of three terms: The first term is the empirical expectation of energy over the training data, the second term depends on the Rademacher complexity of the energy class, and the third term involves the number of the training data  $m$  and the upper-bound of the energy function  $M$ . This shows that merely minimizing the empirical expectation of energy, i.e., the first term, may not yield a good approximation of the true expectation. In (Zhang & LeCun, 2017), it has been shown that regularization using unlabeled data reduces the second and third terms, thus, leading to better generalization. In this work, we express these two terms using the  $\vartheta$ -diversity and show that employing a diversity strategy may also decrease the gap between the true and empirical expectation of the energy. In Section 2.1, we consider the special case of regression and derive two bounds relative to two energy functions based on  $L_1$  and  $L_2$  distances. In Section 2.2, we derive the bound relative to the binary classification task using as energy function  $E(\mathbf{W}, \mathbf{x}, y) = -yG_{\mathbf{W}}(\mathbf{x})$  (LeCun et al., 2006). In Section 2.3, we consider the

case of implicit regression, which encapsulates different learning problems such as metric learning, generative models, and denoising (LeCun et al., 2006). For this case, we use the  $L_2$  distance between the inner models as the energy function.

## 2.1 REGRESSION TASK

Regression can be formulated as an energy-based learning problem (Figure 1 (a)) using the inner model  $G_{\mathbf{W}}(\mathbf{x}) = \sum_{i=1}^D w_i \phi_i(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x})$ . We also suppose that the feature set is well-defined over the input domain  $\mathcal{X}$ , i.e.,  $\forall \mathbf{x} \in \mathcal{X} \|\Phi(\mathbf{x})\|_2 \leq A$ . The two valid energy functions which can be used for regression are:  $E_2(\mathbf{W}, \mathbf{x}, y) = \frac{1}{2} \|G_{\mathbf{W}}(\mathbf{x}) - y\|_2^2$  and  $E_1(\mathbf{W}, \mathbf{x}, y) = \|G_{\mathbf{W}}(\mathbf{x}) - y\|_1$  (LeCun et al., 2006). Theorem 1 and Theorem 2 express the special cases of Lemma 1 using the  $\vartheta$ -diversity of the feature set  $\{\phi_1(\cdot), \dots, \phi_D(\cdot)\}$ .

**Theorem 1.** *For the energy function  $E(h, \mathbf{x}, \mathbf{y}) = \frac{1}{2} \|G_{\mathbf{W}}(\mathbf{x}) - y\|_2^2$ , over the input set  $\mathcal{X} \in \mathbb{R}^N$ , hypothesis class  $\mathcal{H} = \{G_{\mathbf{W}}(\mathbf{x}) = \sum_{i=1}^D w_i \phi_i(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) \mid \Phi \in \mathcal{F}, \forall \mathbf{x} \|\Phi(\mathbf{x})\|_2 \leq A\}$ , and output set  $\mathcal{Y} \subset \mathbb{R}$ , if the feature set  $\{\phi_1(\cdot), \dots, \phi_D(\cdot)\}$  is  $\vartheta$ -diverse with a probability  $\tau$ , then with a probability of at least  $(1 - \delta)\tau$ , the following holds for all  $h$  in  $\mathcal{H}$*

$$\mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim D}[E(h, \mathbf{x}, \mathbf{y})] \leq \frac{1}{m} \sum_{(\mathbf{x}, \mathbf{y}) \in S} E(h, \mathbf{x}, \mathbf{y}) + 8D\|\mathbf{w}\|_{\infty}(\|\mathbf{w}\|_{\infty}\sqrt{DA^2 - \vartheta^2} + B)\mathcal{R}_m(\mathcal{F}) \\ + (\|\mathbf{w}\|_{\infty}\sqrt{DA^2 - \vartheta^2} + B)^2 \sqrt{\frac{\log(2/\delta)}{2m}}, \quad (6)$$

where  $B$  is the upper-bound of  $\mathcal{Y}$ , i.e.,  $y \leq B, \forall y \in \mathcal{Y}$ .

**Theorem 2.** *For the energy function  $E(h, \mathbf{x}, \mathbf{y}) = \|G_{\mathbf{W}}(\mathbf{x}) - y\|_1$ , over the input set  $\mathcal{X} \in \mathbb{R}^N$ , hypothesis class  $\mathcal{H} = \{G_{\mathbf{W}}(\mathbf{x}) = \sum_{i=1}^D w_i \phi_i(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) \mid \Phi \in \mathcal{F}, \forall \mathbf{x} \|\Phi(\mathbf{x})\|_2 \leq A\}$ , and output set  $\mathcal{Y} \subset \mathbb{R}$ , if the feature set  $\{\phi_1(\cdot), \dots, \phi_D(\cdot)\}$  is  $\vartheta$ -diverse with a probability  $\tau$ , then with a probability of at least  $(1 - \delta)\tau$ , the following holds for all  $h$  in  $\mathcal{H}$*

$$\mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim D}[E(h, \mathbf{x}, \mathbf{y})] \leq \frac{1}{m} \sum_{(\mathbf{x}, \mathbf{y}) \in S} E(h, \mathbf{x}, \mathbf{y}) + 4D\|\mathbf{w}\|_{\infty}\mathcal{R}_m(\mathcal{F}) \\ + 2(\|\mathbf{w}\|_{\infty}\sqrt{DA^2 - \vartheta^2} + B)\sqrt{\frac{\log(2/\delta)}{2m}}, \quad (7)$$

where  $B$  is the upper-bound of  $\mathcal{Y}$ , i.e.,  $y \leq B, \forall y \in \mathcal{Y}$ .

The proofs are available in the Appendix. We note that, in Theorem 1 and Theorem 2, we consistently find that the bound of the true expectation of the energy is a decreasing function with respect to  $\vartheta$ . This proves that for the regression task employing a diversity strategy can improve the generalization performance of the energy-based model.

## 2.2 TWO-CLASS CLASSIFIER

Here, we consider the problem of binary classification, as illustrated in Figure 1 (b). Using the same assumption as in regression for the inner model, i.e.,  $G_{\mathbf{W}}(\mathbf{x}) = \sum_{i=1}^D w_i \phi_i(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x})$ , energy function of  $E(\mathbf{W}, \mathbf{x}, y) = -yG_{\mathbf{W}}(\mathbf{x})$  (LeCun et al., 2006), and the  $\vartheta$ -diversity of the feature set, we express Lemma 1 for this specific configuration in Theorem 3.

**Theorem 3.** *For the energy function  $E(h, \mathbf{x}, \mathbf{y}) = -yG_{\mathbf{W}}(\mathbf{x})$ , over the input set  $\mathcal{X} \in \mathbb{R}^N$ , hypothesis class  $\mathcal{H} = \{G_{\mathbf{W}}(\mathbf{x}) = \sum_{i=1}^D w_i \phi_i(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) \mid \Phi \in \mathcal{F}, \forall \mathbf{x} \|\Phi(\mathbf{x})\|_2 \leq A\}$ , and output set  $\mathcal{Y} \subset \mathbb{R}$ , if the feature set  $\{\phi_1(\cdot), \dots, \phi_D(\cdot)\}$  is  $\vartheta$ -diverse with a probability  $\tau$ , then with a probability of at least  $(1 - \delta)\tau$ , the following holds for all  $h$  in  $\mathcal{H}$*

$$\mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim D}[E(h, \mathbf{x}, \mathbf{y})] \leq \frac{1}{m} \sum_{(\mathbf{x}, \mathbf{y}) \in S} E(h, \mathbf{x}, \mathbf{y}) + 4D\|\mathbf{w}\|_{\infty}\mathcal{R}_m(\mathcal{F}) \\ + \|\mathbf{w}\|_{\infty}\sqrt{DA^2 - \vartheta^2} \sqrt{\frac{\log(2/\delta)}{2m}}, \quad (8)$$



The proof is available in the Appendix. Similar to the regression task, we note that the upper-bound of the true expectation is a decreasing function with respect to the diversity term. Thus, a more diverse feature set, i.e., higher  $\vartheta$ , has a lower upper-bound for the true energy.

### 2.3 IMPLICIT REGRESSION

In this section, we consider the problem of implicit regression. This is a general formulation of a different set of problems such as metric learning, where the goal is to learn a distance function between two domains, image denoising, or object detection as illustrated in (LeCun et al., 2006). This form of EBM (Figure 1 (c)) has two inner models,  $G_{\mathbf{W}}^1(\cdot)$  and  $G_{\mathbf{W}}^2(\cdot)$ , which can be equal or different according to the problem at hand. Here, we consider the general case, where the two models correspond to two different combinations of different features, i.e.,  $G_{\mathbf{W}}^1(\mathbf{x}) = \sum_{i=1}^{D^{(1)}} w_i^1 \phi_i^1(\mathbf{x})$  and  $G_{\mathbf{W}}^2(\mathbf{y}) = \sum_{i=1}^{D^{(2)}} w_i^2 \phi_i^2(\mathbf{y})$ . Thus, we have a different  $\vartheta$ -diversity term for each set. The final result is presented in Theorem 4.

**Theorem 4.** *For the energy function  $E(h, \mathbf{x}, \mathbf{y}) = \frac{1}{2} \|G_{\mathbf{W}}^1(\mathbf{x}) - G_{\mathbf{W}}^2(\mathbf{y})\|_2^2$ , over the input set  $\mathcal{X} \in \mathbb{R}^N$ , hypothesis class  $\mathcal{H} = \{G_{\mathbf{W}}^1(\mathbf{x}) = \sum_{i=1}^{D^{(1)}} w_i^{(1)} \phi_i^{(1)}(\mathbf{x}) = \mathbf{w}^{(1)T} \Phi^{(1)}(\mathbf{x}), G_{\mathbf{W}}^2(\mathbf{y}) = \sum_{i=1}^{D^{(2)}} w_i^{(2)} \phi_i^{(2)}(\mathbf{y}) = \mathbf{w}^{(2)T} \Phi^{(2)}(\mathbf{y}) \mid \Phi^{(1)} \in \mathcal{F}_1, \Phi^{(2)} \in \mathcal{F}_2, \forall \mathbf{x} \|\Phi^{(1)}(\mathbf{x})\|_2 \leq A^{(1)}, \forall \mathbf{y} \|\Phi^{(2)}(\mathbf{y})\|_2 \leq A^{(2)}\}$ , and output set  $\mathcal{Y} \subset \mathbb{R}^N$ , if the feature set  $\{\phi_1^{(1)}(\cdot), \dots, \phi_{D^{(1)}}^{(1)}(\cdot)\}$  is  $\vartheta^{(1)}$ -diverse with a probability  $\tau_1$  and the feature set  $\{\phi_1^{(2)}(\cdot), \dots, \phi_{D^{(2)}}^{(2)}(\cdot)\}$  is  $\vartheta^{(2)}$ -diverse with a probability  $\tau_2$ , then with a probability of at least  $(1 - \delta)\tau_1\tau_2$ , the following holds for all  $h$  in  $\mathcal{H}$*

$$\begin{aligned} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim D}[E(h, \mathbf{x}, \mathbf{y})] &\leq \frac{1}{m} \sum_{(\mathbf{x}, \mathbf{y}) \in S} E(h, \mathbf{x}, \mathbf{y}) \\ &\quad + 8(\sqrt{\mathcal{J}_1} + \sqrt{\mathcal{J}_2}) (D^{(1)} \|\mathbf{w}^{(1)}\|_{\infty} \mathcal{R}_m(\mathcal{F}_1) + D^{(2)} \|\mathbf{w}^{(2)}\|_{\infty} \mathcal{R}_m(\mathcal{F}_2)) \\ &\quad + 2(\mathcal{J}_1 + \mathcal{J}_2) \sqrt{\frac{\log(2/\delta)}{2m}}, \quad (9) \end{aligned}$$

where  $\mathcal{J}_1 = \|\mathbf{w}^{(1)}\|_{\infty}^2 (D^{(1)} A^{(1)2} - \vartheta^{(1)2})$  and  $\mathcal{J}_2 = \|\mathbf{w}^{(2)}\|_{\infty}^2 (D^{(2)} A^{(2)2} - \vartheta^{(2)2})$ .

The proof of Theorem 4 is available in the Appendix. The upper-bound of the energy model depends on the diversity variable of both feature sets. Moreover, we note that the bound for the implicit regression decreases proportionally to  $\vartheta^2$ , as opposed to the classification case for example, where the bound is proportional to  $\vartheta$ .

We note that the theory developed in our paper (Theorems 1 to 4) is agnostic to the loss function (LeCun et al., 2006) or the optimization strategy used (Kumar et al., 2019; Song & Ermon, 2019; Yu et al., 2020). We show that increasing the diversity of the features consistently decreases the upper-bound of the true expectation of the energy and, thus, can boost the generalization performance of the energy-based model. We note that our analysis is independent of how the features are obtained, e.g., handcrafted or optimized. In fact, in the recent state-of-the-art EBMs (Khalifa et al., 2020; Bakhtin et al., 2021; Nash & Durkan, 2019; Yu et al., 2020), the features are typically parameterized using a deep learning model and optimized during the training. Thus, our theory suggests the use of a diversity strategy, for example in the form of a regularization as in (Cogswell et al., 2016), to avoid learning redundant features can improve the performance of the model and decrease the gap between the expectation of the true and the empirical energy.

## 3 CONCLUSION

The energy-based learning is a powerful learning paradigm that encapsulates various discriminative and generative systems. An EBM is typically formed of one (or many) inner models which learn a combination of different features to generate an energy mapping for each input configuration. In this paper, we introduced the feature diversity concept, i.e.,  $\vartheta$ -diversity, and we used it to extend the PAC theory of EBMs. We derived different generalization bounds for various learning contexts, i.e., regression, classification, and implicit regression, with different energy functions and we consistently

found that increasing the diversity of the feature set can improve the approximation error of the true expectation of the energy function. We also note that our theory is independent of the loss function or the training strategy used to optimize the parameters of the EBM.

Future directions include developing practical strategies to promote the diversity of the feature set in case the features are optimized following a data-driven process, like the training phase of a neural network.

## REFERENCES

- Michael Arbel, Liang Zhou, and Arthur Gretton. Generalized energy based models. In *International Conference on Learning Representations*, 2021.
- Anton Bakhtin, Yuntian Deng, Sam Gross, Myle Ott, Marc’Aurelio Ranzato, and Arthur Szlam. Residual energy-based models for text. *Journal of Machine Learning Research*, 22(40):1–41, 2021.
- Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, pp. 463–482, 2002.
- Erdem Bıyık, Kenneth Wang, Nima Anari, and Dorsa Sadigh. Batch active learning using determinantal point processes. *arXiv preprint arXiv:1906.07975*, 2019.
- Tong Che, Ruixiang Zhang, Jascha Sohl-Dickstein, Hugo Larochelle, Liam Paull, Yuan Cao, and Yoshua Bengio. Your gan is secretly an energy-based model and you should use discriminator driven latent sampling. *arXiv preprint arXiv:2003.06060*, 2020.
- Michael Cogswell, Faruk Ahmed, Ross B. Girshick, Larry Zitnick, and Dhruv Batra. Reducing overfitting in deep networks by decorrelating representations. In *International Conference on Learning Representations*, 2016.
- Zihang Dai, Amjad Almahairi, Philip Bachman, Eduard Hovy, and Aaron Courville. Calibrating energy-based generative adversarial networks. *arXiv preprint arXiv:1702.01691*, 2017.
- Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc’Aurelio Ranzato. Residual energy-based models for text generation. *arXiv preprint arXiv:2004.11714*, 2020.
- Michal Dereziński, Daniele Calandriello, and Michal Valko. Exact sampling of determinantal point processes with sublinear time preprocessing. In *Advances in Neural Information Processing Systems*, pp. 11546–11558, 2019.
- Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- Yi Fang and Mengwen Liu. A unified energy-based framework for learning to rank. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, pp. 171–180, 2016.
- Lu Gan, Diana Nurbakova, Léa Laporte, and Sylvie Calabretto. Enhancing recommendation diversity using determinantal point processes on knowledge graphs. In *Conference on Research and Development in Information Retrieval*, pp. 2001–2004, 2020.
- Mike Gartrell, Victor-Emmanuel Brunel, Elvis Dohmatob, and Syrine Krichene. Learning nonsymmetric determinantal point processes. In *Advances in Neural Information Processing Systems*, pp. 6718–6728, 2019.
- Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.
- Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pp. 1352–1361. PMLR, 2017.

- Yang He, Ping Liu, Ziwei Wang, Zhilan Hu, and Yi Yang. Filter pruning via geometric median for deep convolutional neural networks acceleration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4340–4349, 2019.
- Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. A distributional approach to controlled text generation. *arXiv preprint arXiv:2012.11635*, 2020.
- Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
- Yusuke Kondo and Koichiro Yamauchi. A dynamic pruning strategy for incremental learning on a budget. In *International Conference on Neural Information Processing*, pp. 295–303. Springer, 2014.
- Rithesh Kumar, Sherjil Ozair, Anirudh Goyal, Aaron Courville, and Yoshua Bengio. Maximum entropy generators for energy-based models. *arXiv preprint arXiv:1901.08508*, 2019.
- Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.
- Seunghyun Lee, Byeongho Heo, Jung-Woo Ha, and Byung Cheol Song. Filter pruning and re-initialization via latent space clustering. *IEEE Access*, 8:189587–189597, 2020.
- Nan Li, Yang Yu, and Zhi-Hua Zhou. Diversity regularized ensemble pruning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 330–345, 2012.
- Shuang Li, Yilun Du, Guido M van de Ven, Antonio Torralba, and Igor Mordatch. Energy-based models for continual learning. *arXiv preprint arXiv:2011.12216*, 2020.
- Chenlin Meng, Lantao Yu, Yang Song, Jiaming Song, and Stefano Ermon. Autoregressive score matching. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.
- Charlie Nash and Conor Durkan. Autoregressive energy machines. In *International Conference on Machine Learning*, pp. 1735–1744. PMLR, 2019.
- Tetiana Parshakova, Jean-Marc Andreoli, and Marc Dymetman. Distributional reinforcement learning for energy-based sequential models. *arXiv preprint arXiv:1912.08517*, 2019.
- Marc Ranzato, Christopher Poultney, Sumit Chopra, Yann LeCun, et al. Efficient learning of sparse representations with an energy-based model. *Advances in neural information processing systems*, 19:1137, 2007a.
- Marc’Aurelio Ranzato, Y-Lan Boureau, Sumit Chopra, and Yann LeCun. A unified energy-based framework for unsupervised learning. In *Artificial Intelligence and Statistics*, pp. 371–379. PMLR, 2007b.
- Pravendra Singh, Vinay Kumar Verma, Piyush Rai, and Vinay Namboodiri. Leveraging filter correlations for deep model compression. In *The IEEE Winter Conference on Applications of Computer Vision*, pp. 835–844, 2020.
- Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL <https://proceedings.neurips.cc/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf>.
- Yee Whye Teh, Max Welling, Simon Osindero, and Geoffrey E Hinton. Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research*, 4(Dec):1235–1260, 2003.
- Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.
- Michael M Wolf. *Mathematical foundations of supervised learning*, 2018.

- Bo Xie, Yingyu Liang, and Le Song. Diverse neural network learns true target functions. In *Artificial Intelligence and Statistics*, pp. 1216–1224. PMLR, 2017.
- Pengtao Xie, Yuntian Deng, and Eric Xing. On the generalization error bounds of neural networks under diversity-inducing mutual angular regularization. *arXiv preprint arXiv:1511.07110*, 2015.
- Ke Yang, Vasilis Gkatzelis, and Julia Stoyanovich. Balanced ranking with diversity constraints. In *International Joint Conference on Artificial Intelligence*, pp. 6035–6042, 2019.
- Lantao Yu, Yang Song, Jiaming Song, and Stefano Ermon. Training deep energy-based models with f-divergence minimization. In *International Conference on Machine Learning*, pp. 10957–10967. PMLR, 2020.
- Yang Yu, Yu-Feng Li, and Zhi-Hua Zhou. Diversity regularized machine. In *International Joint Conference on Artificial Intelligence*, 2011.
- Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection. In *International Conference on Machine Learning*, pp. 1100–1109. PMLR, 2016.
- Xiang Zhang. *Pac-learning for energy-based models*. PhD thesis, Citeseer, 2013.
- Xiang Zhang and Yann LeCun. Universum prescription: Regularization using unlabeled data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.
- Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

## 4 APPENDIX

## 4.1 PROOF OF THEOREM 1

**Lemma 2.** *With a probability of at least  $\tau$ , we have*

$$\sup_{\mathbf{x}, h} |h(\mathbf{x})| \leq \sqrt{\mathcal{J}}, \quad (10)$$

where  $\mathcal{J} = \|\mathbf{w}\|_\infty^2 (DA^2 - \vartheta^2)$  and  $A = \sup_{\mathbf{x}} \|\phi(\mathbf{x})\|_2$ .

*Proof.*

$$\begin{aligned} h^2(\mathbf{x}) &= \left( \sum_{i=1}^D w_i \phi_i(\mathbf{x}) \right)^2 \leq \left( \sum_{i=1}^D \|\mathbf{w}\|_\infty \phi_m(\mathbf{x}) \right)^2 = \|\mathbf{w}\|_\infty^2 \left( \sum_{i=1}^D \phi_i(\mathbf{x}) \right)^2 \\ &= \|\mathbf{w}\|_\infty^2 \left( \sum_{i,j} \phi_i(\mathbf{x}) \phi_j(\mathbf{x}) \right) = \|\mathbf{w}\|_\infty^2 \left( \sum_i \phi_i(\mathbf{x})^2 + \sum_{i \neq j} \phi_i(\mathbf{x}) \phi_j(\mathbf{x}) \right) \end{aligned} \quad (11)$$

We have  $\|\Phi(\mathbf{x})\|_2 \leq A$ . For the first term in equation 11, we have  $\sum_m \phi_m(\mathbf{x})^2 \leq A^2$ . By using the identity  $\phi_m(\mathbf{x}) \phi_n(\mathbf{x}) = \frac{1}{2} (\phi_m(\mathbf{x})^2 + \phi_n(\mathbf{x})^2 - (\phi_m(\mathbf{x}) - \phi_n(\mathbf{x}))^2)$ , the second term can be rewritten as

$$\sum_{m \neq n} \phi_m(\mathbf{x}) \phi_n(\mathbf{x}) = \frac{1}{2} \sum_{m \neq n} \left( \phi_m(\mathbf{x})^2 + \phi_n(\mathbf{x})^2 - (\phi_m(\mathbf{x}) - \phi_n(\mathbf{x}))^2 \right). \quad (12)$$

In addition, we have with a probability  $\tau$ ,  $\frac{1}{2} \sum_{m \neq n} \|\phi_m(\mathbf{x}) - \phi_n(\mathbf{x})\|_2 \geq \vartheta$ . Thus, we have with a probability at least  $\tau$ :

$$\sum_{m \neq n} \phi_m(\mathbf{x}) \phi_n(\mathbf{x}) \leq \frac{1}{2} (2(D-1)A^2 - 2\vartheta^2) = (D-1)A^2 - \vartheta^2. \quad (13)$$

By putting everything back to equation 11, we have with a probability  $\tau$ ,

$$h^2(\mathbf{x}) \leq \|\mathbf{w}\|_\infty^2 (A^2 + (D-1)A^2 - \vartheta^2) = \|\mathbf{w}\|_\infty^2 (DA^2 - \vartheta^2) = \mathcal{J}. \quad (14)$$

Thus, with a probability  $\tau$ ,

$$\sup_{\mathbf{x}, h} |h(\mathbf{x})| \leq \sqrt{\sup_{\mathbf{x}, h} h(\mathbf{x})^2} \leq \sqrt{\mathcal{J}}. \quad (15)$$

□

**Lemma 3.** *With a probability of at least  $\tau$ , we have*

$$\sup_{\mathbf{x}, y, f} |E(h(\mathbf{x}), y)| \leq (\sqrt{\mathcal{J}} + B)^2. \quad (16)$$

*Proof.* We have  $\sup_{\mathbf{x}, y, h} |h(\mathbf{x}) - y| \leq 2 \sup_{\mathbf{x}, y, h} (|h(\mathbf{x})| + |y|) = 2(\sqrt{\mathcal{J}} + B)$ . Thus  $\sup_{\mathbf{x}, y, h} |E(h(\mathbf{x}), y)| \leq (\sqrt{\mathcal{J}} + B)^2$ . □

**Lemma 4.** *With a probability of at least  $\tau$ , we have*

$$\mathcal{R}_m(\mathcal{E}) \leq 4D \|\mathbf{w}\|_\infty (\sqrt{\mathcal{J}} + B) \mathcal{R}_m(\mathcal{F}) \quad (17)$$

*Proof.* Using the decomposition property of the Rademacher complexity (if  $\phi$  is a  $L$ -Lipschitz function, then  $\mathcal{R}_m(\phi(\mathcal{A})) \leq L \mathcal{R}_m(\mathcal{A})$ ) and given that  $E(\cdot, y) = \|\cdot - y\|^2$  is  $K$ -Lipschitz with a constant  $K = \sup_{\mathbf{x}, y, h} \|h(\mathbf{x}) - y\| \leq 2(\sqrt{\mathcal{J}} + B)$ , we have  $\mathcal{R}_m(\mathcal{E}) \leq K \mathcal{R}_m(\mathcal{F}) \leq 2(\sqrt{\mathcal{J}} + B) \mathcal{R}_m(\mathcal{H})$ , where  $\mathcal{H} = \{G_{\mathbf{W}}(\mathbf{x}) = \sum_{i=1}^D w_i \phi_i(\mathbf{x}) \mid \|\mathbf{w}\|_1 \leq D \|\mathbf{w}\|_\infty\}$ . Next, similar to the proof of Theorem 2.10 in (Wolf, 2018), we note that  $\sum_{i=1}^D w_i \phi_i(\mathbf{x}) \in (D \|\mathbf{w}\|_\infty \text{conv}(\mathcal{F} + (-\mathcal{F}))) := \mathcal{G}$ , where  $\text{conv}$  denotes the convex hull and  $\mathcal{F}$  is the set of  $\phi$  functions. Thus,  $\mathcal{R}_m(\mathcal{H}) \leq \mathcal{R}_m(\mathcal{G}) = D \|\mathbf{w}\|_\infty \mathcal{R}_m(\text{conv}(\mathcal{F} + (-\mathcal{F}))) = D \|\mathbf{w}\|_\infty \mathcal{R}_m(\mathcal{F} + (-\mathcal{F})) = 2D \|\mathbf{w}\|_\infty \mathcal{R}_m(\mathcal{F})$ . □

**Theorem 1** For the energy function  $E(h, \mathbf{x}, \mathbf{y}) = \frac{1}{2} \|G_{\mathbf{W}}(\mathbf{x}) - \mathbf{y}\|_2^2$ , over the input set  $\mathcal{X} \in \mathbb{R}^N$ , hypothesis class  $\mathcal{H} = \{G_{\mathbf{W}}(\mathbf{x}) = \sum_{i=1}^D w_i \phi_i(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) \mid \forall \mathbf{x} \|\Phi(\mathbf{x})\|_2 \leq A\}$ , and output set  $\mathcal{Y} \subset \mathbb{R}$ , if the feature set  $\{\phi_1(\cdot), \dots, \phi_D(\cdot)\}$  is  $\vartheta$ -diverse with a probability  $\tau$ , then with a probability of at least  $(1 - \delta)\tau$ , the following holds for all  $h$  in  $\mathcal{H}$

$$\mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim D}[E(h, \mathbf{x}, \mathbf{y})] \leq \frac{1}{m} \sum_{(\mathbf{x}, \mathbf{y}) \in S} E(h, \mathbf{x}, \mathbf{y}) + 8D\|\mathbf{w}\|_{\infty}(\|\mathbf{w}\|_{\infty}\sqrt{DA^2 - \vartheta^2} + B)\mathcal{R}_m(\mathcal{F}) \\ + (\|\mathbf{w}\|_{\infty}\sqrt{DA^2 - \vartheta^2} + B)^2 \sqrt{\frac{\log(2/\delta)}{2m}}, \quad (18)$$

where  $B$  is the upper-bound of  $\mathcal{Y}$ , i.e.,  $y \leq B, \forall y \in \mathcal{Y}$ ,

*Proof.* We replace the variables in Lemma 1 using Lemma 3 and Lemma 4.  $\square$

#### 4.2 PROOF OF THEOREM 2

**Lemma 5.** With a probability of at least  $\tau$ , we have

$$\sup_{\mathbf{x}, y, f} |E(h(\mathbf{x}), y)| \leq 2(\sqrt{\mathcal{J}} + B). \quad (19)$$

*Proof.* We have  $\sup_{\mathbf{x}, y, h} |h(\mathbf{x}) - y| \leq 2 \sup_{\mathbf{x}, y, h} (|h(\mathbf{x})| + |y|) = 2(\sqrt{\mathcal{J}} + B)$ .  $\square$

**Lemma 6.** With a probability of at least  $\tau$ , we have

$$\mathcal{R}_m(\mathcal{E}) \leq 2D\|\mathbf{w}\|_{\infty}\mathcal{R}_m(\mathcal{F}) \quad (20)$$

*Proof.*  $|\cdot|$  is 1-Lipschitz, Thus  $\mathcal{R}_m(\mathcal{E}) \leq \mathcal{R}_m(\mathcal{H})$ .  $\square$

**Theorem 2** For the energy function  $E(h, \mathbf{x}, \mathbf{y}) = \|G_{\mathbf{W}}(\mathbf{x}) - \mathbf{y}\|_1$ , over the input set  $\mathcal{X} \in \mathbb{R}^N$ , hypothesis class  $\mathcal{H} = \{G_{\mathbf{W}}(\mathbf{x}) = \sum_{i=1}^D w_i \phi_i(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) \mid \forall \mathbf{x} \|\Phi(\mathbf{x})\|_2 \leq A\}$ , and output set  $\mathcal{Y} \subset \mathbb{R}$ , if the feature set  $\{\phi_1(\cdot), \dots, \phi_D(\cdot)\}$  is  $\vartheta$ -diverse with a probability  $\tau$ , then with a probability of at least  $(1 - \delta)\tau$ , the following holds for all  $h$  in  $\mathcal{H}$

$$\mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim D}[E(h, \mathbf{x}, \mathbf{y})] \leq \frac{1}{m} \sum_{(\mathbf{x}, \mathbf{y}) \in S} E(h, \mathbf{x}, \mathbf{y}) + 4D\|\mathbf{w}\|_{\infty}\mathcal{R}_m(\mathcal{F}) \\ + 2(\|\mathbf{w}\|_{\infty}\sqrt{DA^2 - \vartheta^2} + B) \sqrt{\frac{\log(2/\delta)}{2m}}, \quad (21)$$

*Proof.* We replace the variables in Lemma 1 using Lemma 5 and Lemma 6.  $\square$

#### 4.3 PROOF OF THEOREM 3

**Lemma 7.** With a probability of at least  $\tau$ , we have

$$\sup_{\mathbf{x}, y, f} |E(h(\mathbf{x}), y)| \leq \sqrt{\mathcal{J}} \quad (22)$$

*Proof.* We have  $\sup -yG_{\mathbf{W}}(\mathbf{x}) \leq \sup |G_{\mathbf{W}}(\mathbf{x})| \leq \sqrt{\mathcal{J}}$   $\square$

**Lemma 8.** With a probability of at least  $\tau$ , we have

$$\mathcal{R}_m(\mathcal{E}) \leq 2D\|\mathbf{w}\|_{\infty}\mathcal{R}_m(\mathcal{F}) \quad (23)$$

*Proof.* We note that for  $y \in \{-1, 1\}$ ,  $\sigma$  and  $-y\sigma$  follow the same distribution. Thus, we have  $\mathcal{R}_m(\mathcal{E}) = \mathcal{R}_m(\mathcal{H})$ . Next, we note that  $\mathcal{R}_m(\mathcal{H}) \leq 2D\|\mathbf{w}\|_{\infty}\mathcal{R}_m(\mathcal{F})$   $\square$

**Theorem 3** For a well-defined energy function  $E(h, \mathbf{x}, \mathbf{y})$  (LeCun et al., 2006), over hypothesis class  $\mathcal{H}$ , input set  $\mathcal{X}$  and output set  $\mathcal{Y}$ , if it has upper-bound  $M$ , then with a probability of at least  $1 - \delta$ , the following holds for all  $h$  in  $\mathcal{H}$

$$\mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim D}[E(h, \mathbf{x}, \mathbf{y})] \leq \frac{1}{m} \sum_{(\mathbf{x}, \mathbf{y}) \in S} E(h, \mathbf{x}, \mathbf{y}) + 4D\|\mathbf{w}\|_{\infty} \mathcal{R}_m(\mathcal{F}) + \|\mathbf{w}\|_{\infty} \sqrt{DA^2 - \vartheta^2} \sqrt{\frac{\log(2/\delta)}{2m}}, \quad (24)$$

*Proof.* We replace the variables in Lemma 1 using Lemma 7 and Lemma 8.  $\square$

#### 4.4 PROOF OF THEOREM 4

**Lemma 9.** With a probability of at least  $\tau_1 \tau_2$ , we have

$$\sup_{\mathbf{x}, \mathbf{y}, f} |E(h(\mathbf{x}), \mathbf{y})| \leq 2(\mathcal{J}_1 + \mathcal{J}_2) \quad (25)$$

*Proof.* We have  $\|G_{\mathbf{W}}^{(1)}(\mathbf{x}) - G_{\mathbf{W}}^{(2)}(\mathbf{y})\|_2^2 \leq 2(\|G_{\mathbf{W}}^{(1)}(\mathbf{x})\|_2^2 + \|G_{\mathbf{W}}^{(2)}(\mathbf{y})\|_2^2)$ . Similar to Theorem 1, we have  $\sup \|G_{\mathbf{W}}^{(1)}(\mathbf{x})\|_2^2 \leq \|\mathbf{w}^{(1)}\|_{\infty}^2 (D^{(1)}A^{(1)^2} - \vartheta^{(1)^2}) = \mathcal{J}_1$  and  $\sup \|G_{\mathbf{W}}^{(2)}(\mathbf{y})\|_2^2 \leq \|\mathbf{w}^{(2)}\|_{\infty}^2 (D^{(2)}A^{(2)^2} - \vartheta^{(2)^2}) = \mathcal{J}_2$   $\square$

**Lemma 10.** With a probability of at least  $\tau_1 \tau_2$ , we have

$$\mathcal{R}_m(\mathcal{E}) \leq 4(\sqrt{\mathcal{J}_1} + \sqrt{\mathcal{J}_2})(D^{(1)}\|\mathbf{w}^{(1)}\|_{\infty} \mathcal{R}_m(\mathcal{F}_1) + D^{(2)}\|\mathbf{w}^{(2)}\|_{\infty} \mathcal{R}_m(\mathcal{F}_2)) \quad (26)$$

*Proof.* Let  $f$  be the square function, i.e.,  $f(x) = \frac{1}{2}x^2$  and  $\mathcal{E}_0 = \{G_{\mathbf{W}}^{(1)}(x) - G_{\mathbf{W}}^{(2)}(y) \mid x \in \mathcal{X}, y \in \mathcal{Y}\}$ . We have  $\mathcal{E} = f(\mathcal{E}_0 + (-\mathcal{E}_0))$ .  $f$  is Lipschitz over the input space, with a constant  $L$  bounded by  $\sup_{x, \mathbf{w}} G_{\mathbf{W}}^{(1)}(x) + \sup_{y, \mathbf{w}} G_{\mathbf{W}}^{(2)}(y) \leq \sqrt{\mathcal{J}_1} + \sqrt{\mathcal{J}_2}$ . Thus, we have  $\mathcal{R}_m(\mathcal{E}) \leq (\sqrt{\mathcal{J}_1} + \sqrt{\mathcal{J}_2})\mathcal{R}_m(\mathcal{E}_0 + (-\mathcal{E}_0)) \leq 2(\sqrt{\mathcal{J}_1} + \sqrt{\mathcal{J}_2})\mathcal{R}_m(\mathcal{E}_0)$ . Next, we note that  $\mathcal{R}_m(\mathcal{E}_0) = \mathcal{R}_m(\mathcal{H}_1 + (-\mathcal{H}_2)) = \mathcal{R}_m(\mathcal{H}_1) + \mathcal{R}_m(\mathcal{H}_2)$ . Using same as technique as in Lemma 4, we have  $\mathcal{R}_m(\mathcal{H}_1) \leq 2D^{(1)}\|\mathbf{w}^{(1)}\|_{\infty} \mathcal{R}_m(\mathcal{F}_1)$  and  $\mathcal{R}_m(\mathcal{H}_2) \leq 2D^{(2)}\|\mathbf{w}^{(2)}\|_{\infty} \mathcal{R}_m(\mathcal{F}_2)$   $\square$

**Theorem 4** For the energy function  $E(h, \mathbf{x}, \mathbf{y}) = \frac{1}{2}\|G_{\mathbf{W}}^{(1)}(\mathbf{x}) - G_{\mathbf{W}}^{(2)}(\mathbf{y})\|_2^2$ , over the input set  $\mathcal{X} \in \mathbb{R}^N$ , hypothesis class  $\mathcal{H} = \{G_{\mathbf{W}}^{(1)}(\mathbf{x}) = \sum_{i=1}^{D^{(1)}} w_i^{(1)} \phi_i^{(1)}(\mathbf{x}) = \mathbf{w}^{(1)^T} \Phi^{(1)}(\mathbf{x}), G_{\mathbf{W}}^{(2)}(\mathbf{y}) = \sum_{i=1}^{D^{(2)}} w_i^{(2)} \phi_i^{(2)}(\mathbf{y}) = \mathbf{w}^{(2)^T} \Phi^{(2)}(\mathbf{y}) \mid \Phi^{(1)} \in \mathcal{F}_1, \Phi^{(2)} \in \mathcal{F}_2, \forall \mathbf{x} \|\Phi^{(1)}(\mathbf{x})\|_2 \leq A^{(1)}, \forall \mathbf{y} \|\Phi^{(2)}(\mathbf{y})\|_2 \leq A^{(2)}\}$ , and output set  $\mathcal{Y} \subset \mathbb{R}^N$ , if the feature set  $\{\phi_1^{(1)}(\cdot), \dots, \phi_{D^{(1)}}^{(1)}(\cdot)\}$  is  $\vartheta^{(1)}$ -diverse with a probability  $\tau_1$  and the feature set  $\{\phi_1^{(2)}(\cdot), \dots, \phi_{D^{(2)}}^{(2)}(\cdot)\}$  is  $\vartheta^{(2)}$ -diverse with a probability  $\tau_2$ , then with a probability of at least  $(1 - \delta)\tau_1 \tau_2$ , the following holds for all  $h$  in  $\mathcal{H}$

$$\mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim D}[E(h, \mathbf{x}, \mathbf{y})] \leq \frac{1}{m} \sum_{(\mathbf{x}, \mathbf{y}) \in S} E(h, \mathbf{x}, \mathbf{y}) + 8(\sqrt{\mathcal{J}_1} + \sqrt{\mathcal{J}_2})(D^{(1)}\|\mathbf{w}^{(1)}\|_{\infty} \mathcal{R}_m(\mathcal{F}_1) + D^{(2)}\|\mathbf{w}^{(2)}\|_{\infty} \mathcal{R}_m(\mathcal{F}_2)) + 2(\mathcal{J}_1 + \mathcal{J}_2) \sqrt{\frac{\log(2/\delta)}{2m}}, \quad (27)$$

where  $\mathcal{J}_1 = \|\mathbf{w}^{(1)}\|_{\infty}^2 (D^{(1)}A^{(1)^2} - \vartheta^{(1)^2})$  and  $\mathcal{J}_2 = \|\mathbf{w}^{(2)}\|_{\infty}^2 (D^{(2)}A^{(2)^2} - \vartheta^{(2)^2})$ .

*Proof.* We replace the variables in Lemma 1 using Lemma 9 and Lemma 10.  $\square$





# PUBLICATION

## VI

**Class-Wise Generalization Error: An Information-Theoretic  
Analysis**

F. Laakom, Y. Bu and M. Gabbouj

*arXiv preprint arXiv:2401.02904. Submitted to ICML (2024)*

© 2024 . Publication reprinted with the permission of the  
copyright holders





