Master Thesis

From Generalists to Specialists: Empowering Weak Learners in Deep Ensembles

Jannik Wirtz

Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science of Artificial Intelligence
at the Department of Advanced Computing Sciences
of the Maastricht University

Thesis Committee:

Dr. Christof Seiler
Dr. Kurt Driessens

Maastricht University
Faculty of Science and Engineering
Department of Advanced Computing Sciences

July 6, 2023

Acknowledgements

I would like to express my heartfelt gratitude to the individuals who have played a significant role in the completion of this Master's thesis, which was written as part of the Master's programme in Artificial Intelligence at the Department of Advanced Computing Sciences, Maastricht University.

First and foremost, I would like to thank my supervisor Christof Seiler for his guidance and inspiration throughout this thesis and the previous internship. His infectious enthusiasm and the enjoyable working dynamic we shared made this experience truly memorable. The freedom he granted me during the thesis enabled me to explore my ideas and push my intellectual boundaries.

Additionally, I would like to thank Kurt Driessens for also supervising my thesis.

I would like to take a moment to express my deepest appreciation to my parents for creating an optimal environment for me to write this thesis. Their unwavering support, love, and belief in my abilities have been a driving force behind my work. I am grateful to my mother for instilling in me the values of industriousness, cultivating resilience and work ethic. Furthermore, I would like to thank my father for fostering my intellectual curiosity from an early age and nurturing my passion for technical pursuits.

Finally, I want to thank my good friends Hendrik Böhmer, Markus Horrmann, Nitai Heeb and Nils Wallenfang for their moral support, discussions and motivation.

Jannik Wirtz
Heinsberg, July 2023

Abstract

Ensemble learning is a machine learning technique that leverages the predictions of multiple diverse predictors to enhance overall predictive performance. In recent years the approach has regained significant research attention due to deep ensembles empirically providing state-of-the-art generalisation accuracy and uncertainty estimates.

This thesis provides a critical empirical analysis of state-of-the-art approaches to deep ensembling, highlighting a critical deficiency: their inability to train specialised models. A limitation that leads to inefficient utilisation of learning capacities due to redundant learnings, resulting in compromised accuracy and diminishing returns for larger ensemble sizes.

In response, we propose Stacked Negative Correlation Learning (SNCL), a novel ensembling method that effectively encourages ensemble members to specialise in distinguishing between class subsets. SNCL incorporates negative correlation learning and stacked generalisation through a novel concurrent training scheme for stacked ensembles, effectively overcoming some of their limitations and leveraging the advantages of both methods.

Through extensive experimentation and analysis, we demonstrate that SNCL effectively reduces redundancies within deep ensembles, yielding improved generalisation accuracy with comparable model complexities. Our proposal further enhances the state-of-the-art performance of deep ensembles for uncertainty quantification by reducing in-domain uncertainty, thereby improving out-of-distribution detection capabilities, which are desirable in safety-critical domains.

Our results demonstrate that SNCL can significantly enhance the performance of deep ensembles in high-dimensional multi-class classification settings, presenting a powerful alternative to state-of-the-art techniques by enabling more efficient deep ensembling. We highlight its potential benefits for practical applications and avenues for future research.

Contents

1 Introduction **3**
1.1 Research Questions 4
1.2 Thesis outline 5

2 Fundamentals **6**
2.1 Ensemble Learning 6
2.1.1 Measures of Ensemble Diversity 7
2.2 Ensemble Construction 8
2.2.1 Independent Training 8
2.2.2 Ensemble-aware Training 8
2.2.3 Stacked Generalisation 10
2.3 Predictive Uncertainty 11
2.3.1 In-domain Uncertainty 12
2.3.2 Distribution Shift 14

3 Stacked Negative Correlation Learning **17**
3.1 Motivation 17
3.2 Proposal Details 18

4 Experiments **20**
4.1 Accuracy and Diversity 21
4.2 Cooperation and specialisation 21
4.3 Predictive Uncertainty 21
4.3.1 In-domain Uncertainty 22
4.3.2 OOD Detection 22

5 Results **23**
5.1 Accuracy and Diversity 23
5.1.1 Diversity 23
5.1.2 Accuracy 25
5.2 Cooperation and specialisation 28
5.3 Predictive Uncertainty 32
5.3.1 In-domain Uncertainty 32
5.3.2 OOD Detection 33

6 Discussion **35**

7 Conclusion **37**
7.1 Research Questions . . . 37
7.2 Future Work . . . 38

A Implementation Details **44**

B Supplemental Results **45**
B.1 Accuracy and Diversity . . . 45
B.2 Cooperation and specialisation . . . 47
B.3 Uncertainty . . . 48
B.3.1 In-domain Uncertainty . . . 48
B.3.2 OOD Detection . . . 49

Chapter 1

Introduction

Ensemble learning is a well-established technique in machine learning (ML) to maximise generalisation performance by combining the predictions of multiple weak learners to exploit their diversity and complementary strengths [1]. Although ensembling has been around for over twenty years and is commonly found in winning submissions to ML competitions [2], the technique has recently experienced a considerable surge in attention. Accommodated by a boom in deep learning research, annual publications on ensemble methods are at an all-time high [3].

While ensemble learning has been extensively studied in the context of decision trees, ensembles of deep neural networks (NN) have recently emerged as a topic of growing interest due to their strong empirical performance regarding generalisation accuracy and uncertainty quantification [4, 5]. As such, this thesis explores ways to improve state-of-the-art (SOTA) techniques for constructing deep ensemble classifiers.

The central principle of ensemble learning is that a group of diverse predictors with low correlation make different errors in their predictions, which tend to cancel each other out when combined in the ensemble. For an ensemble to perform adequately, the submodels also referred to as weak learners, must meet two prerequisites: their predictions must be diverse and accurate, where diversity refers to different predictions for the same data, while accuracy entails outperforming random guessing [6]. Hence, the trade-off between the accuracy and diversity of the ensemble members is a crucial factor for optimal results.

Related works have explored many implicit and explicit methods for inducing diversity in deep ensembles, with varying success. Implicit methods train the submodels independently. Data-driven implicit approaches involve training the base learners on different datasets through bootstrap sampling [7] or data augmentation [8] to model more diverse hypotheses and enable better generalisation. Other implicit approaches involve varying the neural architecture to construct heterogeneous ensembles [9], while some rely purely on the random weight initialisations of neural networks [4].

On the other hand, explicit methods typically involve alterations to the loss function to directly optimise diversity during concurrent ensemble-aware training. For instance, Negative Correlation Learning (NCL) [10] introduces a weighted covariance penalty to the loss function to explicitly model diversity. Various works have shown that NCL is preferable to isolated training schemes because it provides parameterised control over the induced diversity and can induce greater diversity, which can further decrease the generalization error of ensembles [11, 12, 13].

Multiple studies have highlighted that deep ensembles suffer diminishing returns with an increase in the number of their members [14, 15]. Likewise, it has been demonstrated that large ensembles maintain their accuracy even after the majority of their members have been removed [16], which suggests an underutilisation of the additional model complexity due to a lack of cooperation. We are convinced that the use of overly simplistic aggregation functions, particularly averaging, the most commonly employed aggregation method in ensemble learning, is the reason for these diminishing returns [1]. We argue that averaging limits the submodels from cooperating beyond making diverse predictive errors.

We propose that the potential of weak learners might be better harnessed by enabling them to specialise and adopt a division-of-labour approach. While we concur that isolated training regimes significantly limit cooperation, we contend that both categories of state-of-the-art methods share a common inefficiency. They restrict ensemble members to function as generalists, proficient across all aspects of the learning problem, instead of specialising in subsets of the task to follow a cooperative division-of-labour approach [17, 18, 19].

Stacked generalisation is an ensembling technique that introduces a parameterised meta-learner to model more complex aggregation functions through element-wise weightings of the class probabilities emitted by the submodels [20]. However, while stacking is generally regarded as superior to averaging, we contend that the way it is applied today is flawed because the approach provides no mutual interaction between the meta-learner and its submodels since it sequentially optimises the meta-learner after having independently trained the submodels. To remedy this flaw, we introduce a training scheme that concurrently fits the two components as part of our proposal.

In this thesis, we introduce Stacked Negative Correlation Learning (SNCL) by combining stacked generalisation [20] with negative correlation learning [10, 11]. We address the individual weaknesses of these two methods by introducing a concurrent training scheme to achieve enhanced utilisation of the allocated model capacities through submodel specialisation. Simultaneously, our method addresses critical limitations of previous attempts at training deep ensembles with specialists [17, 19].

1.1 Research Questions

To explore how submodels can be encouraged to specialise and what benefits specialists can bring to deep ensembles, we devise three **research questions** framed around our proposal. To begin with, we will investigate how SNCL manages to train deep ensembles with specialised members, which leads us to our first research question:

> ***RQ1:*** *How is the proposed SNCL approach able to train specialised weak learners?*

We then proceed to investigate the benefits of specialists within deep ensembles introduced by SNCL in regard to the classical notion of the generalisation error. Hinton et al. [19] have previously motivated and demonstrated that specialists in ensembles can be beneficial when tackling high-dimensional multi-class classification problems with confusable subsets of classes.

> ***RQ2:*** *To what degree can SNCL reduce the generalisation error of deep ensembles?*

Various works have demonstrated that deep ensembles provide state-of-the-art uncertainty estimates and out-of-distribution detection capabilities, which are especially relevant aspects for deployment in safety-critical domains [4, 5]. We will therefore assess how specialisation through SNCL impacts these qualities with our third research question.

> ***RQ3:*** *How does SNCL impact the quality of uncertainty estimates in deep ensembles?*

1.2 Thesis outline

This thesis proceeds by providing a comprehensive introduction to related works and the fundamental principles of ensemble learning in chapter 2, wherein we lay the groundwork for understanding our proposal. We delve into the various theories, methodologies, and related research findings that have shaped our view of deep ensembling before leading into a detailed presentation of our proposal in chapter 3. The subsequent chapter 4 is devoted to our experimental design to compare state-of-the-art techniques to SNCL in terms of accuracy, specialisation and uncertainty quantification, which we designed with the research questions in mind. In chapter 5, we share the detailed results from an extensive series of experiments we performed as part of this thesis. We subsequently interpret and discuss our results and link them to related studies in chapter 6. Finally, in chapter 7, we summarise our key findings and their contribution to the field. We conclude the thesis by addressing the posed research questions and highlighting avenues for future research.

Chapter 2

Fundamentals

2.1 Ensemble Learning

Ensemble learning is a technique that achieves superior generalisation performance by combining the predictions of multiple models, exploiting their diversity and complementary strengths [1]. Some of the most commonly used learning algorithms for ensembling include decision trees and neural networks. Effective ensembling relies on its constituent members to make diverse predictions and errors. Intuitively, averaging highly similar predictions would not result in any significant improvements to generalisation performance.

The importance of diversity within ensembles is well grounded in theory. Despite recent studies suggesting that the bias-variance trade-off of neural network models might not be as applicable to modern overparameterised architectures [21, 22], the dilemma initially studied by Geman et al. [23] is still widely acknowledged today. The intuition remains that as the complexity of a NN increases, its predictions demonstrate reduced bias but increased variance. This implies that neural networks can model a hypothesis that closely fits the training data, which may however show limited applicability to test data, a phenomenon known as overfitting.

Hence, the generalisation error of neural networks decomposes into a bias and a variance term. Because ensembles are constructed from multiple submodels, their generalisation error decomposes into an additional term, the covariance of their constituents. The covariance term introduces a crucial dynamic to the process of ensemble construction, where the optimal submodel diversity is that which optimally balances the three components to minimize the overall error [18]. Recently, Fort et al. [24] confirmed earlier theories that diverse sets of models generalise better because they can cover a larger portion of the hypothesis space [1, 18], by demonstrating that randomly initialised neural networks explore diverse modes within the function space. Beyond diversity, Hinton et al. [19] have demonstrated that specialised submodels can be beneficial when dealing with high-dimensional multi-class classification problems with confusable subsets of classes.

Although generalisation is a key priority for the deployment of ML models in the real world, the use of ensemble learning incurs additional computational and storage costs during development and inference, which presents an important trade-off for practical applications. Various techniques have been shown to be effective in limiting this overhead. Ensemble pruning is a technique aimed at reducing the total number of members within an ensemble by removing the least contributing ones after training [25]. Pruning can typically reduce the inference costs without significant losses to an ensemble's predictive performance, as most ensembles tend to be robust to the removal of constituents [16]. While the time and space complexity of most ensembles scales linearly with their number of members, some more efficient approaches have been proposed for deep ensembles. Studies have shown that neural architectures with shared parameters can reduce the necessary additional parameters for adding a member, with limited sacrifices to the ensemble's diversity, particularly when sharing parameters of early convolutional layers, which focus on similar low-level features [13, 26, 27]. On the other hand, methods of knowledge distillation have shown promising results in compressing the learnings of ensembles into single models, eliminating the computational overhead at inference time [19, 28].

Some studies have shown that ensemble learning can serve a purpose beyond simply maximising predictive performance, as ensembles can more efficiently achieve similar performances to large individual models. In the realm of image classification, it has been shown that ensembles can be more efficient than individual models of equal capacity, as they train faster, utilise fewer floating point operations, and exhibit higher accuracy and memory efficiency [14, 15].

2.1.1 Measures of Ensemble Diversity

Although the importance of diversity for effective ensembling is generally agreed upon, there is no single established metric to quantify it. The two diversity metrics most commonly cited in the literature are the Kullback-Leibler (KL) divergence and the disagreement rate of the submodel predictions [13, 27]. Both metrics are pair-wise measures, i.e. they aim to quantify the dissimilarity of predictions from two submodels at a time. To estimate the diversity of an ensemble, the metrics are computed and averaged for each pair of submodels based on the respective predictions for a test set of samples $X = \{x_n | n = 1, ..., N\}$. As such, the estimates are biased by the sample. The disagreement rate is probably the most intuitive way of measuring diversity, which is simply the ratio of test samples on which two models disagree, i.e. assign a different class label $\hat{y}$. The disagreement rate is computed as follows:

$$D_{\text{Dis.}}(f_i, f_j, X) = \frac{1}{N} \sum_{n=1}^{N} I(\arg\max_{\hat{y}}(f_i(x_n)) \neq \arg\max_{\hat{y}}(f_j(x_n))),$$

where $f_i(x_n)$ is the discrete probability distribution predicted for the sample x_n by model i and $\arg\max_{\hat{y}}(f_i(x_n))$ is the index of the class with the highest likelihood in $f_i(x_n)$.

On the other hand, the KL-divergence is a statistical metric to quantify the dissimilarity of two probability distributions. As opposed to the disagreement rate, the KL-divergence is not just based on the mode of the respective discrete distributions but takes their entirety into account. Therefore it considers all of the secondary choices, which may be argued to be just as important since ensemble learning relies on diverse errors being made. For discrete probability distributions, i.e. in a classification setting, the KL-divergence is computed based on the softmax probabilities per datapoint $x_n \in X$,

$$D_{\text{KL}}(f_i, f_j, X) = \sum_{n=1}^{N} f_i(x_n) \cdot \log\left(\frac{f_i(x_n)}{f_j(x_n)}\right)$$

Since the KL-divergence is an asymmetric measure $D_{\text{KL}}(f_i, f_j, X) \neq D_{\text{KL}}(f_j, f_i, X)$, it is typically computed and averaged for all submodel pairs (f_i, f_j) within an ensemble where $i \neq j$.

2.2 Ensemble Construction

Various approaches exist for inducing and controlling submodel diversity during ensemble training to minimise the generalisation error. Based on related research findings, we will now provide a comprehensive overview of popular explicit diversity management techniques and their implicit counterparts, including their benefits and drawbacks for deep ensembling.

2.2.1 Independent Training

Independent ensemble training approaches fit their submodels to the data without knowledge of each other. As such, they solely rely on exogenous factors to diversify the modelled functions.

One of the earliest yet most popular methods today is bagging [7], which trains algorithms on bootstrap samples of the original dataset, thus expecting to diversify the learned hypotheses. Bagging is a variance-reducing technique which averages the predictions of its submodels to form a final prediction and allows for ensembles of heterogeneous learners. Despite its popularity with decision tree ensembles, bagging has been shown to be ineffective, even detrimental to deep ensemble diversity and performance [26, 29]. An average bootstrap sample contains only 63% of the unique samples present in the original dataset [4, 9]. Nixon et al. [29] have shown that the reduced number of unique datapoints in the training set is the primary determinant, which renders bagging a non-viable option for deep ensembles. Fort et al. [24] have shown that independently trained deep ensembles with random initialisation provide greater diversity and generalisation performance than bagging. To the present day, independent training with random initialisation remains one of the state-of-the-art methods for deep ensemble construction.

More fruitful alternatives to bagging include methods which generate artificial datapoints to create diverse training datasets without reducing the number of unique datapoints [8]. Besides data-driven approaches, heterogeneous ensembles rely on a variety of base models to consider different features of the data for diversity. While some approaches use learning algorithms from entirely different algorithmic families, it has been shown that deep ensembles composed of submodels with varied neural architectures can provide greater diversity than solely relying on random initialisation [9].

2.2.2 Ensemble-aware Training

Ensemble-aware approaches were introduced as a critique of independently trained ensembles, motivated by the decomposition of the ensemble generalisation error and the idea that interaction between the individual learners is essential to encourage cooperation and more optimally balance individual model accuracy and diversity [10, 12, 17, 18].

Lee et al. [26] further critiqued that classical literature commonly refers to ensemble members as experts or specialists, despite a lack of efforts to encourage specialisation. Consequently, Lee et al. [17] proposed stochastic multiple choice learning for end-to-end training of deep ensembles. End-to-end training is a trivial approach to ensemble-aware optimisation as it directly optimises the ensemble as a whole without consideration for the individual learners' performance. In their testing, they report significant diversity and show that their method yields specialised models. However, they find that their models suffer from overspecialisation, leading to decreased overall ensemble accuracy [26]. Independent studies have shown that their method severely underperforms independently trained deep ensembles with random initialisation [11].

Various studies have observed that end-to-end ensemble training tends to offer great diversity but tends to underperform, with some explanations offered in recent studies. Owed to the fact that the individual performance of the base models is entirely neglected during training, many models become overly focused on offsetting the errors of their peers and rely on a single dominant predictor within the ensemble for baseline predictions [16, 30]. As previously mentioned, effective ensembling relies on a set of diverse weak learners who must perform better than random guessing. Therefore, end-to-end training is not a viable option.

Another noteworthy and commonly cited approach for constructing ensembles with specialists was proposed by Hinton et al. [19]. Their multi-step process involves applying a clustering algorithm to the covariance matrices of previously trained generalist models to inform the allocation of frequently confused class subsets, on which they consequently train specialist models. While they observe slight accuracy benefits with this approach, they acknowledge that using the actual discriminative performance of the experts would be preferable for guiding class assignments but argue about the difficulty of parallelisation.

Despite the ineffectiveness of end-to-end training and the limited diversity offered by independent ensemble training, many ensemble-aware methods are concerned with the spectrum in-between the two extremes. We will now provide a detailed introduction to Negative Correlation Learning (NCL) [10], which laid the foundation for modern methods, including our proposal.

Negative Correlation Learning

Negative correlation learning (NCL) was first introduced by Liu et al. in 1999 [10] and has continued to attract research interest to the present day due to its simplicity and effectiveness in inducing and controlling diversity in ensembles of neural networks. It concurrently trains ensemble members and explicitly models diversity by adding a correlation penalty to the loss function to decorrelate the predictions made by the submodels. The penalty term is weighted by the parameter λ, which enables smooth interpolation between independent (λ=0) and end-to-end (λ=1) ensemble training. Like most other ensembling techniques, NCL uses a simple average to form the ensemble prediction $F(x)$ from the submodel predictions $f_m(x), m = 1, ..., M$:

$$F(x) = \frac{1}{M} \sum_{m=1}^{M} f_m(x).$$

When training the ensemble on a batch of labeled samples $\mathcal{B} = \{(x_n, y_n) | n = 1, ..., N\}$ the NCL loss is computed as follows:

$$\mathrm{NCL}(\mathcal{B}) = \frac{1}{N} \sum_{n=1}^{N} \left(\frac{1}{2}(F(x_n) - y_n)^2 + \frac{\lambda}{M} \sum_{m=1}^{M} (f_m(x_n) - F(x_n))^2 \right).$$

With large $\lambda \geq 0.5$, models become interdependent and make negatively correlated predictions [10, 16]. In 2005, Brown et al. [18] established the theoretical underpinning of NCL to support its empirical successes in regression and classification problems [31] by presenting a bias-variance-covariance decomposition of the mean squared error for neural network ensembles. In 2020, Buschjäger et al. [11] presented a generalised decomposition, extending beyond the MSE, thereby encapsulating many approaches comparable to NCL [13, 32]. As such, Generalised Negative Correlation Learning (GNCL) allows for arbitrary losses such as cross-entropy, making it adaptable to various learning problems. The aggregation function $F(.)$ remains a simple average, as in the classical NCL formulation. The GNCL loss is defined as:

$$\mathrm{GNCL}(\mathcal{B}) = \frac{1}{N} \sum_{n=1}^{N} \left(\lambda \mathcal{L}(F(x_n), y_n) + \frac{1-\lambda}{M} \sum_{m}^{M} \mathcal{L}(f_m(x_n), y_n) \right),$$

where $\mathcal{L}(.)$ is an arbitrary loss function computed on a prediction $\hat{y}$ and the ground truth y.

Buschjäger et al. [11] demonstrated that GNCL compares favourably to state-of-the-art ensembling methods with regards to the induced diversity and resulting ensemble accuracy on a variety of benchmark datasets for image classification. They further showed that the optimal trade-off between submodel accuracy and diversity also depends on the individual models' capacity. Therefore diversity should not be the sole concern when determining a favourable trade-off [11]. With models, which manage to achieve training errors close to zero, they found that diversity can even be a detriment to generalisation performance, affirming similar findings by Brown et al. [18]. Overall, their results strongly advocate for explicit methods like NCL, as opposed to implicit approaches, which lack control over the degree to which diversity is encouraged. Their experimental results further indicate that not all diversity is created equal. In specific configurations, stochastic multiple choice learning and bagging achieved diversity levels comparable to GNCL but drastically underperformed in terms of generalisation accuracy [11], reflecting related results by Fort et al. [24] which showed that specific methods can offer superior accuracy-diversity trade-offs to others. Generalised negative correlation learning is an essential foundation for our stacked negative correlation learning proposal, which we present in chapter 3.

2.2.3 Stacked Generalisation

In 1992, Wolpert [20] critiqued the popular aggregation method of averaging for its lack of sophistication and introduced stacked generalisation as a preferable alternative, which can model more complex aggregation functions. The idea of stacking is to construct an ensemble of two levels. The first level, level-0, consists of a diverse set of models independently trained on different folds of the original dataset to induce diversity. Subsequently, the level-1 generaliser, also referred to as meta-learner, is trained on the predictions of the level-0 models to optimally compose the final ensemble prediction, aiming to improve accuracy.

In practice, the second step of training the meta-learner is performed by freezing the parameters of the base models or equivalently constructing a level-1 dataset from the predictions of the level-0 models [33, 34]. For classification tasks, as opposed to regression problems, the level-1 dataset is typically constructed from the concatenated class probabilities emitted by the submodels instead of the predicted class indices [35]. For every datapoint in a batch of labeled samples $\mathcal{B} = \{(x_n, y_n) | n = 1, ..., N\}$, we concatenate the discrete probability distributions for k classes in C emitted by the M predictors f_m, to form the level-1 datapoints. Every classifier f_m predicts a discrete probability distribution over n classes, conditioned on the input x_n:

$$f_m(x_n) = \begin{bmatrix} f_m(C_1|x_n) & f_m(C_2|x_n) & ... & f_m(C_k|x_n) \end{bmatrix}.$$

The level-1 dataset is constructed from the concatenated probability distributions of all predictors and retains the true labels of the original dataset, as illustrated in table 2.1. Figure 2.1 shows an exemplary data flow schematic for a stacked ensemble with three members.

Table 2.1: Level-1 Dataset: Concatenated softmax outputs of M submodels

f_1	f_2	...	f_M	label
$f_1(x_1)$	$f_2(x_1)$	...	$f_M(x_1)$	y_1
$f_1(x_2)$	$f_2(x_2)$	...	$f_M(x_2)$	y_2
...	...	...	...	...
$f_1(x_N)$	$f_2(x_N)$	...	$f_M(x_N)$	y_N

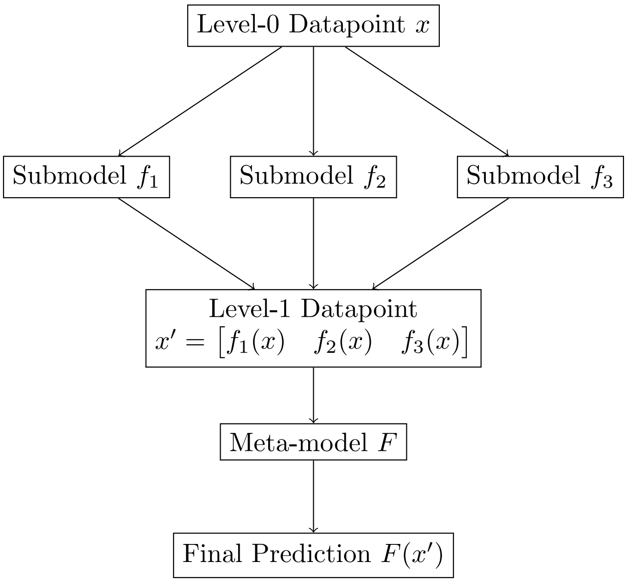


Figure 2.1: Flowchart illustrating the data flow from original data to final prediction

Stacked generalisation encompasses simple and weighted averages as primitive level-1 generalisers [20]. To avoid overfitting, a regularised meta-learner is generally recommended [36]. While stacking was initially proposed with a regularised multi-response least-squares linear regression as meta-model [20], modern applications of stacked ensembling have evolved to utilise shallow neural networks, which allow for non-linear aggregations of the intermediate predictions, yielding further performance enhancements [34, 37]. Stacking is a bias- and variance-reducing ensembling technique and is generally considered superior to simpler alternatives. Yet, it remains one of the least used methods in practice due to its added implementation complexity and overhead [3, 36].

2.3 Predictive Uncertainty

With the widespread application of machine learning across various domains, especially in areas such as medical diagnostics and autonomous driving where human lives are at stake, reliable uncertainty quantification for predictions made by ML algorithms has emerged as a central focus in contemporary research [4, 38]. Robust uncertainty estimates allow practitioners and end-users to assess the degree to which a given prediction should be trusted. Some of the most popular black-box learning algorithms today are deep neural networks. Despite their impressive state-of-the-art performance across various tasks, they suffer from a significant drawback: overconfident erroneous predictions [38].

In classification tasks, neural network models typically produce softmax outputs that are commonly interpreted as confidences. However, it is important to note that these outputs are not reliable probabilities but rather pseudo-probabilities due to the issue of miscalibration in modern neural networks [38, 39]. This miscalibration indicates a lack of correspondence between the predicted confidences and the actual likelihood of a correct prediction. Post-hoc calibration through temperature scaling has been shown to be a simple yet effective method to improve the calibration of softmax outputs [39, 40].

Besides miscalibration, neural networks are vulnerable to erroneous predictions caused by small input perturbations [41] and unrecognisable images, commonly referred to as adversarial attacks [42]. Strauss et al. [43] have demonstrated that deep ensembles are more robust to adversarial attacks than single neural networks, even with sophisticated defence strategies. Furthermore, deep ensembles have been shown to provide uncertainty estimates that are qualitatively comparable to more computationally expensive methods like Bayesian neural networks [4, 5]. Many modern works have tried to explain the superior uncertainty estimates of deep ensembles with the disagreements of their members [4, 24, 44], a long-standing intuition which was most recently contradicted by multiple studies [45, 46].

We will now introduce the concept of in-domain uncertainty and the importance of OOD detection in addressing the challenge of distribution shift in real-world applications of ML.

2.3.1 In-domain Uncertainty

The concept of in-domain uncertainty is concerned with a model's uncertainty on data taken from the same distribution as the training data, also referred to as in-distribution (ID) data. In this setting, models are evaluated on an ID test set to quantify their uncertainty, a procedure analogous to the common hold-out validation approach, that assesses a model's generalisation accuracy on unseen ID data by evaluating it on a random subsample from a large dataset that is excluded from training. An optimal model will make accurate predictions with minimal uncertainty and avoid making confident erroneous predictions, which could have fatal consequences in high-stakes settings.

Recent works in uncertainty estimation use a common set of metrics, including the expected calibration error, the Brier score and the log-likelihood to assess individual models and draw comparisons among different methods [4, 5, 47]. Ashukha et al. [40] recently discussed their various flaws and argued against their use, as they are specifically unsuitable for drawing comparisons between methods. For a fair comparison of in-domain uncertainty, they suggest calibrating all models through temperature scaling before the assessment.

In this work, we will perform assessments with a metric derived from the procedure of conformal prediction, which offers statistical guarantees. Before we introduce conformal prediction, we provide a short description of temperature scaling and entropy.

Temperature Scaling Temperature scaling introduces the parameter T to the softmax outputs of NN classifiers, a scaling factor which is applied to the entire distribution and optimised to minimise the negative log-likelihood on a calibration dataset. Temperature scaling does not affect the accuracy of a model because it does not change the mode of the probability distributions. The logits z_i emitted by a model are divided by T and then transformed into class probabilities q_i as follows:

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

By default, softmax outputs are scaled with $T = 1$, higher values of $T > 1$ soften the output distribution, increasing entropy, and vice versa [39].

Entropy Entropy is an important concept in machine learning, probably most infamous for its role in constructing decision tree models. The Shannon entropy is a measure of uncertainty for discrete probability distributions [48], commonly used in ML literature to quantify predictive uncertainty in classification settings [4, 5, 49]. The Shannon entropy of a prediction p, representing a discrete probability distribution over a set of classes C, is calculated as follows:

$$H(p) := -\sum_{k \in C} p(k) \log p(k)$$

Conformal Prediction

Conformal prediction (CP) is a powerful statistical technique that can transform any uncertainty heuristic, such as softmax outputs, into a rigorous metric of uncertainty with statistical guarantees [50]. The statistical guarantees that CP provides are related to the choice of a significance threshold denoted by α, based on which we can define prediction sets that will contain the true value or class label with a probability of at least $1 - \alpha$.

Thanks to its model-agnostic and distribution-free nature, CP operates independently of the type of model used or the distribution from which the data originates, making it applicable across a wide array of statistical and machine learning models, thereby offering robust uncertainty quantification in a multitude of contexts, including regression settings.

We will specifically focus on split conformal prediction. Although this type of CP sacrifices statistical efficiency because it requires a portion of the available validation data for calibration, it only requires a single fitting of the predictor, making it more suitable for deep learning models, which tend to be expensive to train. We further narrow our focus to the setting of multi-class classification with adaptive prediction sets. This type of conformal procedure produces prediction sets of varying sizes, dependent on the inherent uncertainty of the model for a given datapoint. We will utilise the average set size as a metric to quantify and compare the uncertainty of different models in our experiments in section 4.3.1. For a more elaborate and comprehensive introduction to CP, we refer to the work of Angelopoulos and Bates [50].

We will now formally introduce conformal prediction with adaptive prediction sets. Angelopoulos and Bates [50] suggest to use an i.i.d. calibration dataset $D_{\text{cal.}} = \{(x_i, y_i)|i = 1, ..., n\}$ with $n \geq 1000$ samples. After calibration, prediction sets formed for a new datapoint $(x_{\text{test}}, y_{\text{test}})$, sampled i.i.d from the same distribution as $D_{\text{cal.}}$, provide the conformal coverage guarantee:

$$1 - \alpha \leq \mathbb{P}(y_{\text{test}} \in C(x_{\text{test}})),$$

where $\mathbb{P}(y_{\text{test}} \in C(x_{\text{test}}))$ is referred to as the marginal coverage [50].

1. We define a nonconformal score function $s(x, y) \in \mathbb{R}$, based on an uncertainty heuristic, where smaller scores encode lower uncertainty [50]. Here, we sum the descendingly sorted softmax scores for each class until we reach the true label:

   $$s(x, y) = \sum_{j=1}^{k} \hat{f}(x)_{\pi_j(x)}, \text{ where } y = \pi_k(x)$$

   and $\pi(x)$ is the permutation of classes $\{1, ..., K\}$ that sorts $\hat{f}(x)$ from most likely to least likely, based on the prediction of the pre-trained model.

2. Compute $\hat{q}$ as quantile of the calibration scores $S = \{s(x_i, y_i)|i = 1, ..., n\},$ using $D_{\text{cal.}}$,

   $$\hat{q} = \text{Quantile}\left(S, \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right).$$

3. Use this quantile $\hat{q}$ to form the prediction set for a new example x_{test}:

   $$C(x_{\text{test}}) = \{\pi_1(x_{\text{test}}), ..., \pi_k(x_{\text{test}})\}, \text{ where } k = sup\{k' : \sum_{j=1}^{k'} \hat{f}(x_{\text{test}})_{\pi_j(x_{\text{test}})} < \hat{q}\} + 1.$$

   Figure 2.2 illustrates the construction of an exemplary prediction set.

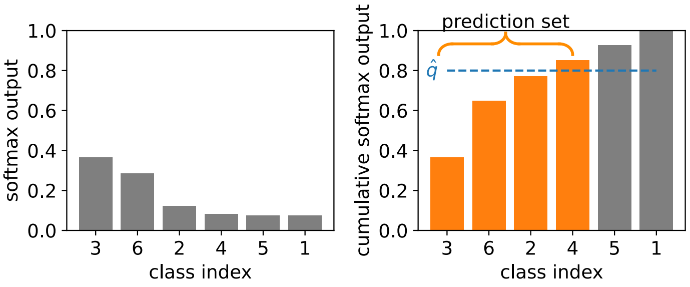


Figure 2.2: Example of prediction set assembly.

2.3.2 Distribution Shift

The usual procedure followed in supervised learning is to collect and label a sample from a real-world distribution and train a ML model to make accurate predictions on this dataset. In practice, the dataset is commonly split using a hold-out validation strategy to monitor generalisation performance and avoid overfitting. However, there are various challenges in real-world settings which can severely impact a model's performance over time. Most existing machine learning models are trained based on the closed world assumption, where the test data is assumed to be drawn i.i.d. from the same distribution as the training data. However, when machine learning models are deployed in the real world, an open world, they are likely to encounter out-of-distribution (OOD) samples.

Dataset shift or domain shift refers to changes in the underlying real-world distribution, which often lead to significant challenges for ML algorithms trained on a previous sample, which no longer accurately reflects the actual distribution. Dataset shift is a phenomenon where the statistical properties of a data distribution change over time and in ways the model was not prepared for during training. Most real-world problems tackled with ML algorithms are vulnerable to various shifts, which can substantially undermine their accuracy, rendering their predictions unreliable, even harmful [51].

While we generally want a model to generalise well to slightly different situations, we also do not want it to remain confident in cases of extreme shifts. When faced with data from a shifted distribution, we would prefer the model's predictions to signal an increase in uncertainty in accordance with the shift's intensity and the associated loss in accuracy [5]. This way, practitioners working on downstream tasks can interpret high uncertainty as a need for human intervention and implement appropriate precautions. The presence of extreme dataset shifts in real-world settings necessitates accurate uncertainty estimates in combination with methods for OOD detection, which enable practitioners to establish a threshold beyond which to call for intervention. This study focuses specifically on covariate and semantic shifts, which we will introduce in the context of out-of-distribution detection.

OOD Detection

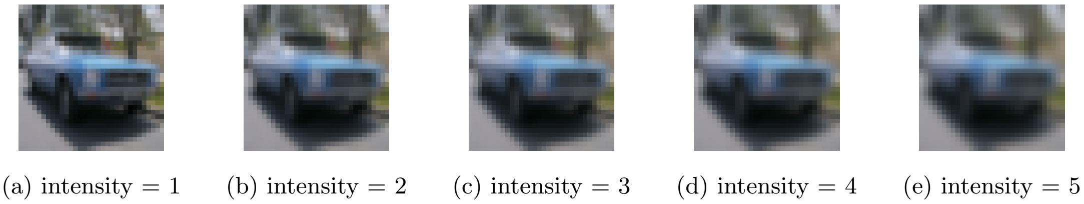
OOD detection is a binary classification problem that distinguishes between in- and out-of-distribution data. While research in this area is still at an early stage, deep ensemble models have shown promising capabilities for this task [45]. In safety-critical domains, we may want our model to detect this type of data to avoid making an otherwise likely erroneous decision and instead rely on human intervention [5]. A common approach to out-of-distribution detection is to perform thresholding based on the inherent uncertainty in a models predictions to classify samples accordingly [5, 38, 45]:

$$g(p,t) = \begin{cases} \text{ID}, & \text{if } U(p) < t, \\ \text{OOD}, & \text{if } U(p) \geq t \end{cases},$$

where $U(p)$ is the uncertainty in a predicted probability distribution p, based on an uncertainty function $U(.)$, like the Shannon entropy $H(.)$ and t is the threshold used to classify samples.

Covariate Shift Covariate shift refers to a kind of dataset shift, which exclusively relates to the distribution of the inputs $P(X)$, where $P(X_{\text{ID}}) \neq P(X_{\text{OOD}})$. The conditional relationship of $P(y|x)$ and the label space $P(Y)$ remain intact [51, 52]. For instance, consider a self-driving car model trained and tested on visual data captured in ideal sunny conditions. However, during deployment in the winter months, the car encounters heavy snowfall, which obstructs its view of the traffic. In this scenario, although the classification targets, such as vehicles and pedestrians, remain the same, the camera footage differs drastically, posing a significant challenge to the model due to changes in the input distribution $P(X)$.

Although the scenario of covariate shift is most commonly used to assess a model's robustness to small perturbations in the input, extreme covariate shifts typically cause severe decreases in model accuracy, requiring human intervention in critical domains to avoid negative consequences.



(a) intensity = 1 (b) intensity = 2 (c) intensity = 3 (d) intensity = 4 (e) intensity = 5

Figure 2.3: Example images of a car from the CIFAR-10C [53] dataset with Gaussian blur of varying intensities

Hendrycks et al. [53] have provided post-processed versions of popular benchmark datasets for image classification to evaluate the robustness of ML models. They applied several common perturbations to widely-used datasets like CIFAR-10 and imagenet. Although originally created for robustness assessments, their datasets have also been commonly used in related works to benchmark uncertainty estimates of deep learning models [5, 47, 49]. Their CIFAR-10C dataset comprises 15 perturbations individually applied to the test set of CIFAR-10 for five different intensity levels, one through five, with five being the most severe. We depict an example image of a car corrupted with varying degrees of Gaussian blur from the CIFAR-10C dataset [53] in figure 2.3.

Semantic Shift Semantic shift is concerned with a shift of the label space $P(Y_{\text{ID}}) \neq P(Y_{\text{OOD}})$, naturally this type of shift also results in a shift of the input distribution $P(X)$ [52]. Continuing with the example of autonomous driving, suppose that a city introduces a new traffic sign outside of its training dataset's scope. In turn, the model will likely fail to recognise this sign or confuse it with a known one, which may lead to incorrect actions and dangerous situations.

The SVHN [54] dataset is a commonly used benchmark for assessing the OOD detection capabilities of models trained on CIFAR-10 or CIFAR-100 for semantic shift [47, 52]. While CIFAR-10 is composed of images featuring animals and vehicles, SVHN [54] includes images of house numbers.

Evaluation Metrics Previous works qualitatively evaluated the OOD detection capabilities of different models and methods by examining the entropy distributions of their predictions on varying datasets to assess how amenable a model is to thresholding for OOD detection [4, 5]. On OOD data, one would expect the predictive distributions to be of high entropy and vice versa for ID data, such that they exhibit limited overlap and are more easily separable.

Recent studies have established quantitative standards to compare the OOD detection capabilities of different methods based on thresholding, specifically the Area Under the Receiver Operator Characteristic curve (AUROC) and the true negative rate (TNR) at 95% true positive rate (TPR), i.e. TNR@TPR95 [47]. While the AUROC allows for a threshold independent assessment of alternatives, the latter assesses the performance at a 95% recall of in-distribution data, simulating a challenging practical setting [55]. The ROC plots the corresponding TPR for every false-positive rate (FPR) between zero and one. We depict an example of the ROC with a random and perfect classifier in figure 2.4. An AUROC of 50% corresponds to a random classifier, whereas 100% corresponds to an idealistic perfect detector.

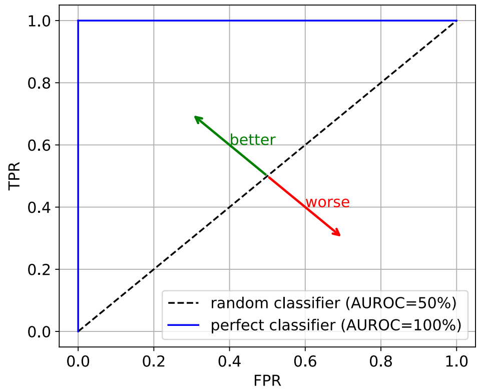


Figure 2.4: ROC visualisation with perfect and random classifier

Chapter 3

Stacked Negative Correlation Learning

In this chapter, we motivate and describe in detail the key contribution of this thesis, the novel method of Stacked Negative Correlation Learning (SNCL) for deep ensemble construction.

3.1 Motivation

Our proposal is motivated by the lack of specialisation among base learners in current approaches to deep ensembling, which commonly leads to redundant learnings, resulting in suboptimal use of the allocated learning capacity. Webb et al. [16] have shown strong evidence of significant redundancies within independently trained ensembles and those constructed with GNCL, even with very large $\lambda = 0.99$. As prefaced, techniques like GNCL decorrelate the predictions made by the individual models to increase diversity but fail to train specialists. We argue that this limitation is due to their use of overly simple aggregation functions, specifically averaging.

Although the GNCL loss allows learners to stray away from purely optimising their individual performances to decrease the ensembles loss cooperatively, they are unable to specialise in a subset of the learning problem because all class predictions are given equal weight during optimisation since the aggregation is carried out at the model level, rather than on a per-class basis. As a result, the GNCL loss function encourages all learners to maximise their performance across all classes, such that they cannot specialise in a subset of the learning problem without negatively impacting the ensemble's predictions through their worsened performance for any neglected classes. In such a setting, we argue that specialisation is not feasible.

Let us now consider how stacked generalisation can alleviate this limitation. The meta-model introduces element-wise scaling to the modelled aggregation function, including negative weights and weights close to zero. Thereby it provides the required flexibility to enable specialisation. Element-wise weightings of the probability vectors emitted by the weak learners will allow the meta-model to effectively assign class subsets to them, such that they are not obligated to form an opinion on the complete probability distribution of a sample. Instead, their objective may primarily revolve around effectively distinguishing between specific subsets of classes within the dataset, as motivated by Hinton et al. [19] and Lee et al. [26]. With the addition of negative weights, the ensemble can move beyond cooperation through diversified errors by allowing the models to offset the errors made by their peers. Furthermore, zero-weighted prediction values are a way of disposing of predictions for classes that a specialist model has neglected, which consequently are less likely to be of value for the final prediction, similar to the notion of a dustbin class for specialists discussed by Hinton et al. [19].

However, simply increasing the degrees of freedom of our aggregation function by introducing a meta-learner does not suffice. Stacking in the form it is widely used today lacks mutual interaction between the submodels and the meta-learner. For the meta-model to influence the training trajectories of the submodels, a feedback loop between the two is required. To address this deficiency, we introduce a concurrent training scheme to facilitate interaction. Therefore, our proposal enjoys the benefits of both stacking and negative correlation learning while improving upon their drawbacks, forming a more powerful ensemble training technique which allows submodels to specialise and focus on their individual strengths during training.

Equally, our proposal will address some significant shortcomings of previous attempts to construct ensembles with specialised models. Since SNCL is based on generalised negative correlation learning, it operates in-between the spectrum of independent and end-to-end learning, such that it will not face the same drawbacks as the previously discussed stochastic multiple choice learning [17], which render it unviable. Furthermore, SNCL also addresses key limitations in the multi-step process presented by Hinton et al. [19], as the meta-learner uses the discriminative performances of the ensemble members to dynamically allocate class subsets for them to specialise in, as part of a holistic and parallelised optimisation process.

3.2 Proposal Details

Having sufficiently motivated our idea, we will now delve into the details of our proposal. Being based upon GNCL [11], our proposal of stacked negative correlation Learning (SNCL) is equally specialised for base learners who are capable of iteratively minimising a loss function, most notably neural networks [18]. Based on the previously discussed ineffectiveness of bagging for deep ensembles, we train the ensemble on the entire training dataset. SNCL allows for heterogeneous neural architectures, data augmentation techniques and supplementation with artificial data [8]. As meta-learner, we use an unregularised single layer of neurons with softmax activation. The weight matrix W is of size $MC \times C$, where C is the number of classes and M is the number of ensemble members. The meta-model maps the level-1 data to the consolidated final prediction vector through matrix multiplication. Our approach allows the same flexibility in choosing a meta-model as stacked generalisation, including more complex meta-learners such as multi-layer perceptron models. To enable the meta-model to guide the submodels during training, we introduce its weightings to the equation of the GNCL loss function from Buschjäger et al. [11] by simply replacing the averaging function $F(.)$ with the parameterised meta-model $F_{\text{meta}}(.)$

$$F_{\text{meta}}(x) = \text{softmax}(\sum_{m=1}^{M} f_m(x) \cdot w_m + b),$$

such that we arrive at the SNCL loss function

$$\text{SNCL}(\mathcal{B}) = \frac{1}{N} \sum_{n=1}^{N} \left(\lambda \mathcal{L}(F_{\text{meta}}(x_n), y_n) + \frac{1-\lambda}{M} \sum_{m=1}^{M} \mathcal{L}(f_m(x_n), y_n) \right),$$

where the probability vector $f_m(x)$ is weighted by the square submatrix w_m of W. In practice, we parallelise the forward passes of all submodels through a composite model $H(x)$, which uses grouped convolutions and modular linear layers in PyTorch, such that we arrive at the equivalent

$$F_{\text{meta}}(x) = \text{softmax}(H(x) \cdot W + b), \text{ where } H(x) = \begin{bmatrix} f_1(x) & ... & f_M(x) \end{bmatrix}.$$

Finally, we introduce a concurrent training scheme for the submodels and the meta-learner, enabling mutual interaction between the two such that the meta-model can steer the learning trajectories of the submodels based on their individual performances. In every step of every epoch, we first update the weights of the submodels based on the SNCL loss function, after which we proceed to update the meta-model weights. Notably, the meta-learner is not fitted to the outputs of the most recently updated models but those of the last forward pass. Therefore we refrain from having to perform two forward passes per optimisation step. As a result of this scheme, the submodels and meta-learner share the same batchsize s.

We detail the pseudocode of the training procedure introduced by SNCL in Alg. 1 and make our Python code publicly available[1], which includes an implementation of stacked negative correlation learning using the PyTorch ML framework.

The parameters of the ensemble members are defined as $\theta_H = \{\theta_m | m = 1, ..., M\}$ and the weights of the meta-learner as θ_F. We devise separate learning rates for the submodels α_H and the meta-model α_F.

Algorithm 1 Stacked Negative Correlation Learning (SNCL)

```
Require: Training dataset $\mathcal{D}$, submodels $H$, meta-learner $F$, loss function $\mathcal{L}$, batchsize $s$
procedure SNCL($\mathcal{D}, H, F, \mathcal{L}, s$)
    for each epoch do
        Randomly shuffle dataset $\mathcal{D}$
        Split the dataset $\mathcal{D}$ into batches $\mathcal{B} = \{b_1, b_2, \ldots, b_N\}$ of size $s$
        for each batch $b_n$ in $\mathcal{B}$ do
            Retrieve inputs and true labels $X, y \leftarrow b_n$
            Reset gradients $\nabla\boldsymbol{\theta_H} = 0$
            Compute level-0 outputs $\hat{\boldsymbol{y}} = H(X)$                                ▷ Forward pass
            Compute SNCL loss $l = \lambda\mathcal{L}(F(\hat{\boldsymbol{y}}), y) + \frac{(1-\lambda)}{M}\sum_{m=1}^{M}\mathcal{L}(\hat{\boldsymbol{y}}_m, y)$
            Compute gradients $\nabla\boldsymbol{\theta_H}$                                      ▷ Backward pass
            Update level-0 model parameters $\boldsymbol{\theta_H} = \boldsymbol{\theta_H} - \alpha_H\nabla\boldsymbol{\theta_H}$     ▷ Optimiser step

            Reset gradients $\nabla\boldsymbol{\theta_F} = 0$
            Compute meta-learner loss $l = \mathcal{L}(F(\hat{\boldsymbol{y}}), y)$                    ▷ Forward pass
            Compute gradients $\nabla\boldsymbol{\theta_F}$                                      ▷ Backward pass
            Update meta-learner parameters $\boldsymbol{\theta_F} = \boldsymbol{\theta_F} - \alpha_F\nabla\boldsymbol{\theta_F}$      ▷ Optimiser step
        end for
    end for
end procedure
```

[1] https://github.com/JannikWirtz/SNCL

Chapter 4

Experiments

This section presents a series of experiments aimed at answering the posed research questions by comparing our proposal with two SOTA deep ensembling approaches, namely independent training [4] and generalised negative correlation learning [11]. The experiments performed in this study are not meant to produce state-of-the-art results on benchmark datasets but rather serve as a relative comparison of SOTA techniques for ensemble construction by exploring a vast space of configurations while remaining cognisant of the associated computational costs.

All ensembles used in the experiments were trained on the two popular benchmark datasets CIFAR-10 and CIFAR-100 [56], using the provided split into five training folds and one fold for testing, with 10.000 datapoints per fold. Each datapoint consists of a coloured and square image with a resolution of 32 pixels, depicting one of 10 or 100 classes, respectively.

As base models, we chose a small convolutional neural network (CNN) with approximately 150.000 parameters each and limited our testing to homogeneous ensembles. We keep the neural architecture consistent throughout the experiments, only adjusting the output dimensions to the given dataset. For reproducibility, we provide its details in table A.1 of the appendix. The CNN comprises six convolutional layers and two dense layers.

We trained every model for 50 epochs on the normalised datasets, to which we applied the same standard data augmentation techniques as Buschjäger et al. [11], namely random cropping and horizontal flipping. We utilised the ADAM optimiser with the categorical cross-entropy as base loss function $\mathcal{L}$ to fit the submodels and meta-models with all methods in this comparison. We used a batchsize of 128 in conjunction with a learning rate of 0.001, which we divided by 10 every 15 epochs for all meta- and submodels. As meta-models, we used single-layer neural networks with the dimensions $MC \times C$ and softmax activation. Our experiments were executed on a desktop machine with an NVIDIA GeForce RTX 3090 GPU and an Intel Core i7-12700K CPU with 5.00 GHz. We used PyTorch to implement our methods and execute our experiments.

The first set of experiments was designed to examine the efficacy of our proposal in lowering the generalisation error of deep ensembles, addressing RQ2. We proceeded with an assessment of cooperation and specialisation within the trained ensembles and a qualitative analysis of the underlying mechanisms of SNCL that facilitate submodel specialisation, to answer RQ1. We concluded our series of experiments by evaluating the quality of the uncertainty estimates of the previously trained ensembles to determine if SNCL could provide reduced in-domain uncertainty and improve the OOD detection capabilities in deep ensembles, answering RQ3.

4.1 Accuracy and Diversity

To assess the efficacy of the proposed SNCL methodology for reducing the generalisation error of deep ensembling, we conducted a comparison its test accuracy and member diversity with GNCL and independently trained deep ensembles. All models were randomly initialised.

To explore how the ensemble size M and different values for the parameter λ impact the relative effectiveness of GNCL and SNCL, we considered a comprehensive parameter space, detailed in table 4.1. GNCL with $\lambda = 0$ represents independent ensemble training. For each configuration, we conducted 10 trials for statistical significance. We assessed the generalisation performance of the ensembles by measuring their test accuracy i.e. the percentage of correctly classified samples of the test set. To compare the induced diversity across methods, we used the average pair-wise disagreement rate and KL-divergence.

In addition, we also evaluated the accuracies obtained by incorporating a meta-learner into ensembles formed with GNCL using the unchanged GNCL loss function. This allowed us to isolate the impact of the meta-learner's weights on the optimization trajectories of the submodels, providing a more nuanced analysis of SNCL's effects.

Table 4.1: Configuration space for experiments in section 5.1

Parameters	Values
Covariance penalty weight λ	[0, 0.25, 0.5, 0.75, 0.9, 1]
Ensemble size M	[3, 5, 10, 15, 20]
Dataset	[CIFAR-10, CIFAR-100]

4.2 Cooperation and specialisation

To evaluate the ability of SNCL to enable greater cooperation and train specialised weak learners, we performed multiple qualitative assessments and comparisons of the previously trained ensembles. Inspired by work from Webb et al. [16], we first examined the degree of interdependence between the submodels within the ensembles by assessing their robustness with regard to the removal of random subsets of ensemble members. As an indication of greater cooperation and lower degrees of redundancy within the ensemble, we would expect to observe more severe drops in the test accuracy of the remaining ensemble as more members are removed. We averaged our results across 20 trials per subset of size $m \in [1, M)$.

To find evidence not just for greater cooperation but specialisation, we proceeded with a closer examination of the impact on the per-class ensemble accuracies after the removal of each individual ensemble member. Finally, to validate our understanding of SNCL, we closely examined how the weights of the trained meta-model correspond to the specialisations of its submodels, to explore the mechanisms through which it induces specialisation.

Finally, to verify our results, we examined the average softmax outputs of the trained ensembles on the CIFAR-10 test set, filtered by the true class labels. This was done to evaluate whether SNCL could enhance the predictive margins between subsets of confusable classes.

4.3 Predictive Uncertainty

Lastly, we offer a critical evaluation of ensemble models constructed with the three alternative procedures through the lens of their predictive uncertainty. Specifically, we evaluated their in-domain uncertainty and capabilities for out-of-distribution detecting to assess their suitability for safety-critical real-world applications. To answer our third research question, we compared the performance of state-of-the-art deep ensemble methods to our proposal.

4.3.1 In-domain Uncertainty

To assess and compare the in-domain uncertainty of the three alternatives, we examined ensembles trained on CIFAR-10 and quantified their uncertainty on the CIFAR-10 test fold. Instead of perpetuating the use of flawed uncertainty metrics and evaluation practices, discussed in detail by Ashukha et al. [40], we performed temperature scaling for all alternatives to not confound our results by measuring the calibration error. We assessed their in-domain uncertainty with a metric derived from conformal prediction.

We conducted 20 trials using the CIFAR-10 test fold with 10.000 datapoints. In each trial we randomly sampled 2.000 datapoints without replacement to be used for temperature scaling and calibration with conformal prediction, we then recorded the average prediction set sizes achieved on the remaining 8.000 datapoints by each respective method. We compared the different methods based on their average prediction set sizes, using CP with adaptive prediction sets, where smaller sets indicate less uncertainty in a model's predictions on the in-domain test data. We examined the average prediction set sizes for the alternatives with $M \in \{5, 20\}$ and $\lambda \in \{0.5, 0.9\}$ and across a range of relevant significance levels $\alpha \in [0.01, 0.3)$, with error bars computed across trials. As a post-hoc test, we also examined the achieved marginal coverages and the class-specific conditional coverages to assess the statistical efficiency of the models. For this experiment, we utilised the Python library MAPIE for conformal prediction.

4.3.2 OOD Detection

In our final experiment, we compared the out-of-distribution (OOD) detection capabilities of the ensembles, based on thresholding performed on the softmax entropies of their predictions. This comparison again utilized the ensembles previously trained on CIFAR-10 using the three alternatives. We conducted trials across models of the same configuration with $M \in \{5, 20\}$ and $\lambda \in \{0.5, 0.9\}$. Their performance was measured via the established metrics AUROC and TNR@TPR95%. Separate experiments were carried out for covariate and semantic shifts. For OOD detection under covariate shift, we utilised the commonly used benchmark dataset of CIFAR-10C provided by Hendrycks et al. [53] and the SVHN [54] dataset for semantic shift detection, which matches the resolution of CIFAR-10. We further confirmed our findings qualitatively by examining and comparing the densities of the softmax entropies on the three datasets and assessing their overlap. The densities were estimated with a Gaussian kernel.

Chapter 5

Results

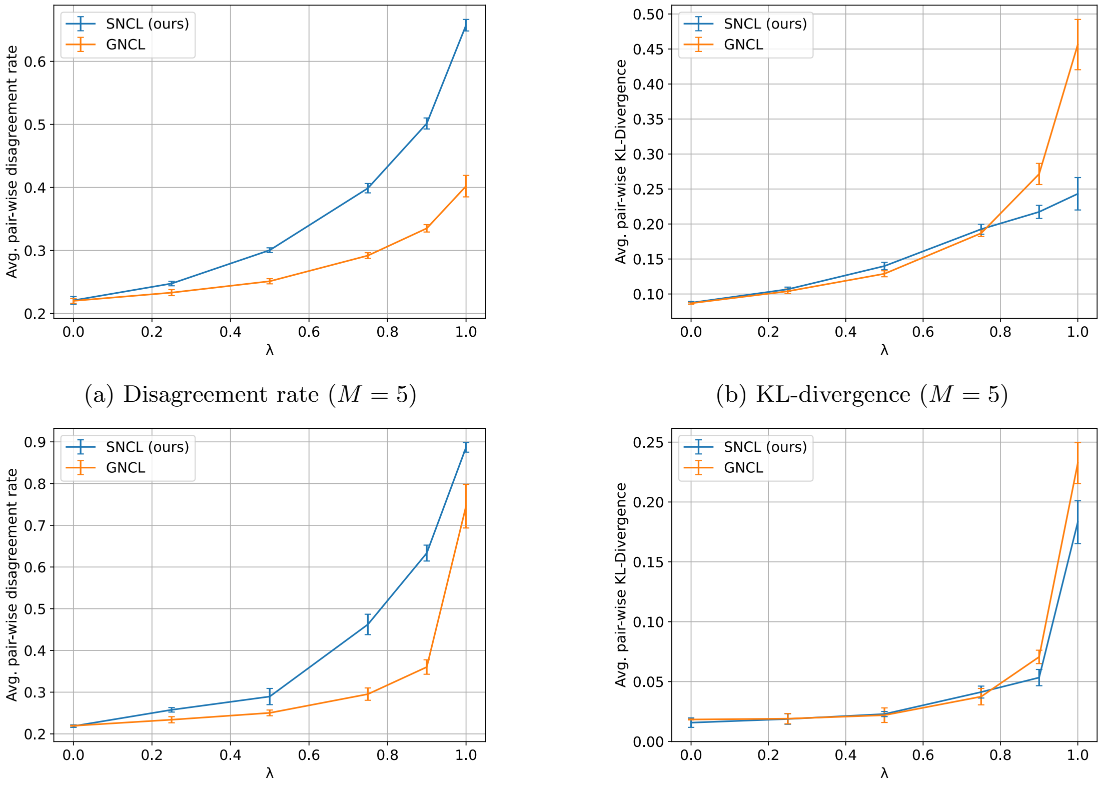
In the ensuing chapter, we present our results from an extensive series of experiments, comparing state-of-the-art deep ensembling approaches to our proposal concerning accuracy, specialisation and predictive uncertainty.

5.1 Accuracy and Diversity

In this section, we present the results of deep ensemble training performed on CIFAR-10. We compared two state-of-the-art methods to our proposal for varying ensemble sizes M and λ with respect to ensemble accuracy and member diversity, with standard error bars derived from 10 trials per configuration. We provide supplemental results for CIFAR-100 in section B.1 of the appendix.

5.1.1 Diversity

In figure 5.1 we display the diversity induced by GNCL and SNCL as measured by the average pairwise disagreement rate and KL-divergence of submodel predictions on the test dataset. We compared the diversity across a range of $\lambda \in \{0, 0.25, 0.5, 0.75, 0.9, 1.0\}$ for ensemble sizes $M \in \{3, 5, 10, 15, 20\}$. For the CIFAR-10 dataset, we observed significantly greater disagreement rates between submodels for ensembles trained with SNCL across all tested M and $\lambda \geq 0.25$. Although SNCL achieved greater rates of disagreement, we observed that the KL-divergence of the two methods tends to be very similar, with GNCL exhibiting higher divergence for small ensembles $M \leq 5$ and $\lambda \geq 0.9$. Similar results were observed in our analogous experiment on the CIFAR-100 dataset, as can be seen in figure B.3 in the appendix. Both methods exhibited generally higher rates of disagreement, but comparable levels of diversity when $M = 5$.



(a) Disagreement rate ($M = 5$)

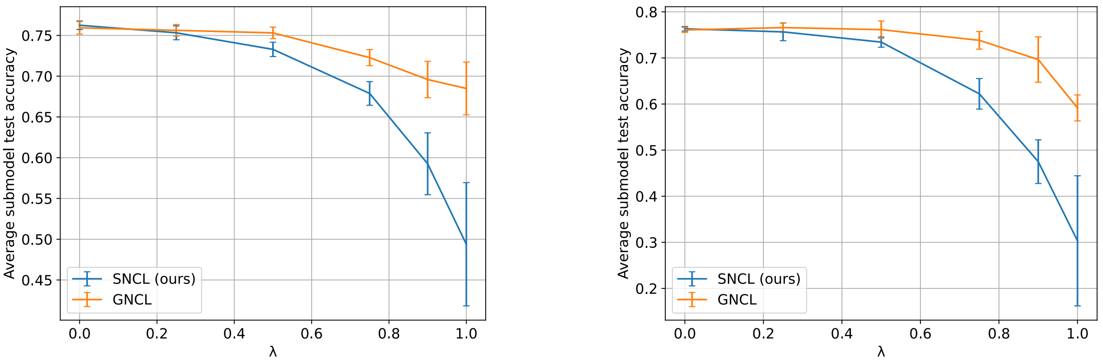
(b) KL-divergence ($M = 5$)

(c) Disagreement rate ($M = 20$)

(d) KL-divergence ($M = 20$)

Figure 5.1: Average pair-wise diversity metrics of ensembles trained with SNCL and GNCL for a spectrum of λ values and $M \in \{5, 20\}$

Figure 5.2 displays the average submodel accuracy for ensemble sizes $M = 5$ and $M = 20$ with $\lambda \in \{0, 0.25, 0.5, 0.75, 0.9, 1.0\}$. With rising diversity, we noted decreasing average submodel accuracies for both methods. We observed greater decreases in submodel accuracies of ensembles trained with SNCL relative to GNCL for the same M and $\lambda > 0.5$.



(a) Average submodel test accuracy's (M=5)

(b) Average submodel test accuracy's (M=20)

Figure 5.2: Average submodel accuracies of ensembles trained with SNCL and GNCL for a spectrum of λ values

Besides the overall lower submodel accuracies of ensembles trained with SNCL, we also observed a greater variance in the per-class accuracies of SNCL ensembles, an example of which is depicted in figure 5.3. The relative levels of the class accuracies are very similar across methods due to the inherent classes' difficulty, i.e. their confusability with other classes. We also observed that all submodels achieve an individual per-class accuracy that surpasses random guessing, for GNCL and SNCL with $\lambda = 0.9$, the largest tested $\lambda < 1$.

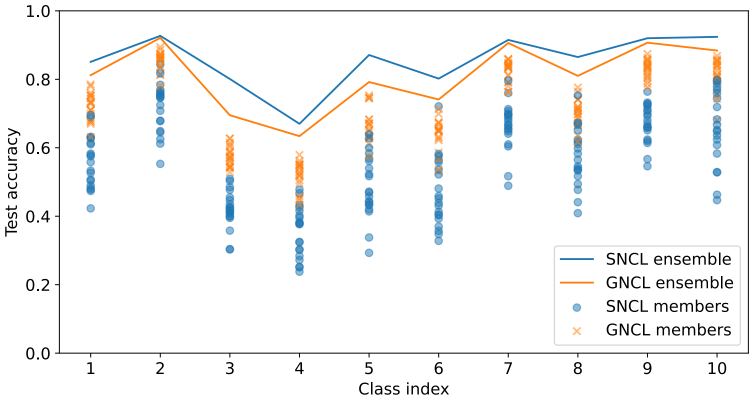
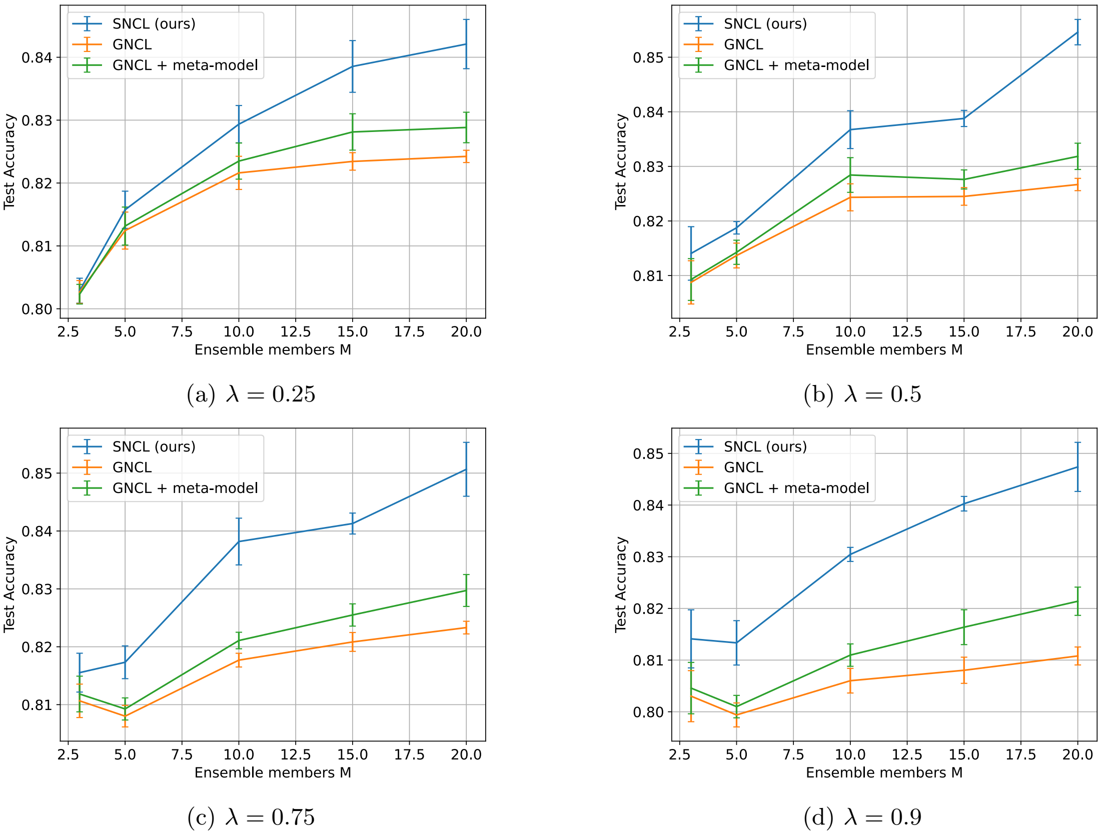


Figure 5.3: Per-class test accuracies of ensembles and their constituents, trained with SNCL and GNCL with M=20 and $\lambda = 0.9$

5.1.2 Accuracy

Figure 5.4 depicts the ensemble accuracies achieved on the CIFAR-10 test fold after 50 epochs of training, drawing a comparison between ensembles trained with GNCL and SNCL for varying ensemble sizes M and λ in-between the spectrum of independent and end-to-end training, with standard error bars derived from 10 trials per configuration.



(a) $\lambda = 0.25$

(b) $\lambda = 0.5$

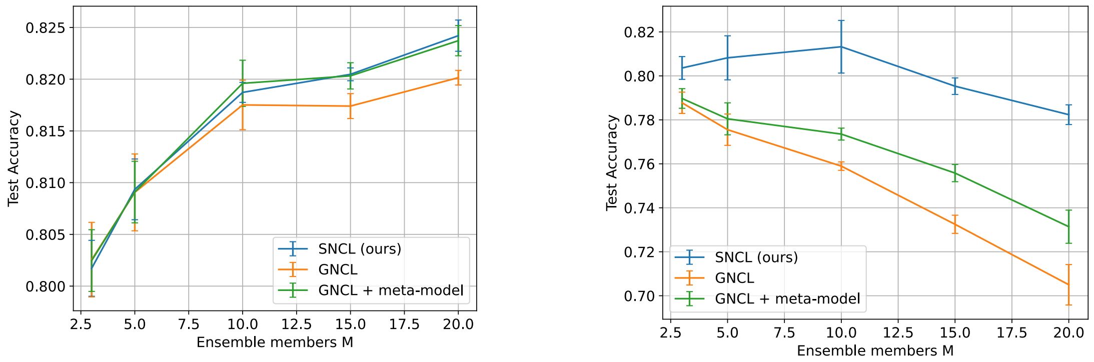
(c) $\lambda = 0.75$

(d) $\lambda = 0.9$

Figure 5.4: CIFAR-10 test accuracy of ensembles trained with SNCL and GNCL for various ensemble sizes M and λ between independent and end-to-end training. Standard error bars were computed across 10 trials each.

We observed that for ensemble sizes $M > 5$ and $\lambda \geq 0.25$, SNCL significantly surpasses the accuracy of GNCL. Interestingly, despite the greatly increased submodel diversity induced by GNCL, we observed only slight improvements to the test accuracy when we trained a meta-model to aggregate their predictions. Under the same parameters and equivalent model complexities, SNCL ensembles performed significantly better than stacked GNCL ensembles. Overall, the performance of SNCL ensembles scaled better with larger ensemble sizes, whereas the two alternatives encountered diminishing returns for $M \geq 10$. In our testing, all methods achieved peak performance for $\lambda = 0.5$.

In figure 5.5 we depict the test accuracies achieved with independent and end-to-end ensemble training on CIFAR-10. As previously explained, GNCL with $\lambda = 0$ is equivalent to independent training, as the ensemble members solely focus on their individual performances. For $\lambda = 0$ we observed that all three alternatives achieve similar accuracies on the test data, and SNCL matched the performance of GNCL with the addition of a meta-model. For end-to-end training with $\lambda = 1$, we observed that SNCL outperforms the alternatives. Independent of the method, we noticed a significant number of submodels that do not surpass individual accuracies beyond random guessing during end-to-end training.

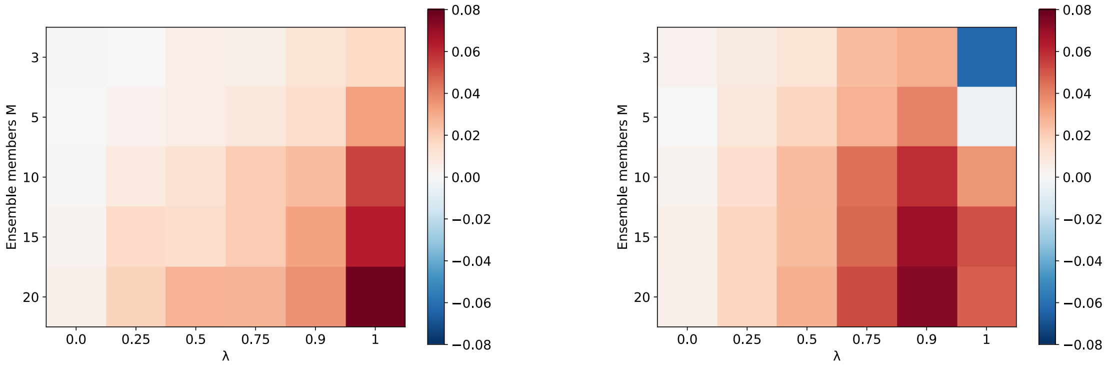


(a) Independent training ($\lambda = 0$)

(b) End-to-end training ($\lambda = 1$)

Figure 5.5: CIFAR-10 test accuracy of SNCL and GNCL with independent and end-to-end training for the tested ensemble sizes

Figure 5.6 summarises the results of our extensive experiments, illustrating the difference in ensemble accuracy between SNCL and GNCL for identical configurations of varying ensemble sizes $M \in \{3, 5, 10, 15, 20\}$ and $\lambda \in \{0, 0.25, 0.5, 0.75, 0.9, 1.0\}$. Overall, we have found that both the ensemble size M and the strength of the negative correlation penalty λ positively impact the performance gap achieved by SNCL over GNCL. While we noticed the greatest increases in diversity and ensemble accuracy for $\lambda \geq 0.5$, we generally found that SNCL tends to perform at least as well as GNCL. Both were superior to independent training in our testing for $\lambda \geq 0.25$. Our findings for CIFAR-10 generally align with our results for the CIFAR-100 dataset, whose detailed results can be found in the appendix (fig. B.1). However, we observed inconsistencies for small ensembles $M < 10$ trained end-to-end ($\lambda = 1$), where GNCL outperformed SNCL significantly (fig. B.2).



(a) CIFAR-10

(b) CIFAR-100

Figure 5.6: Test accuracy difference between SNCL and GNCL for all tested configurations of M and λ.

5.2 Cooperation and specialisation

Figure 5.7 depicts the remaining test accuracies of previously trained ensembles, following the continuous removal of random individual members i.e. the performance of randomly chosen subsets of size $m \leq M$. Since the ensembles constructed with SNCL are tailored specifically to the target ensemble size M, we plot multiple removal trajectories to show that SNCL ensembles are significantly more cooperative in tackling the learning problem of CIFAR-10 across different ensemble sizes $M \in \{5, 10, 20\}$. Although a similar argument could be made for GNCL, we refrain from plotting these additional curves, because they are close to identical.

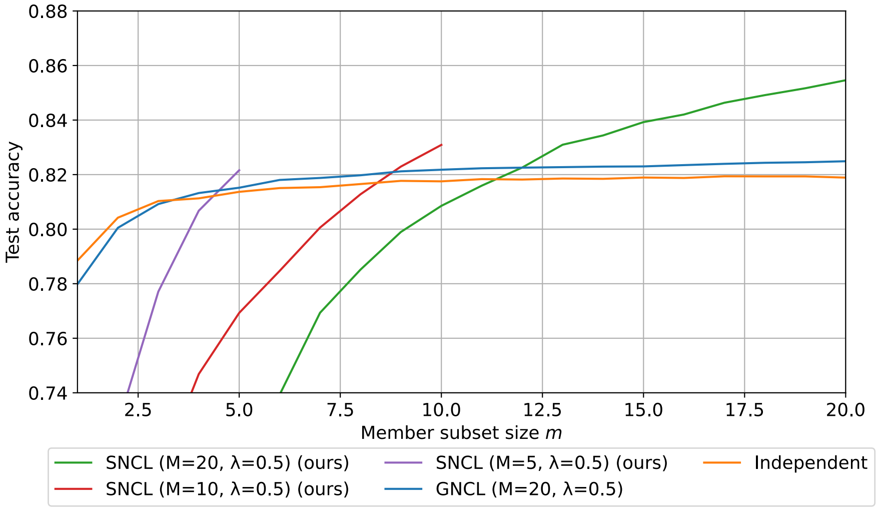
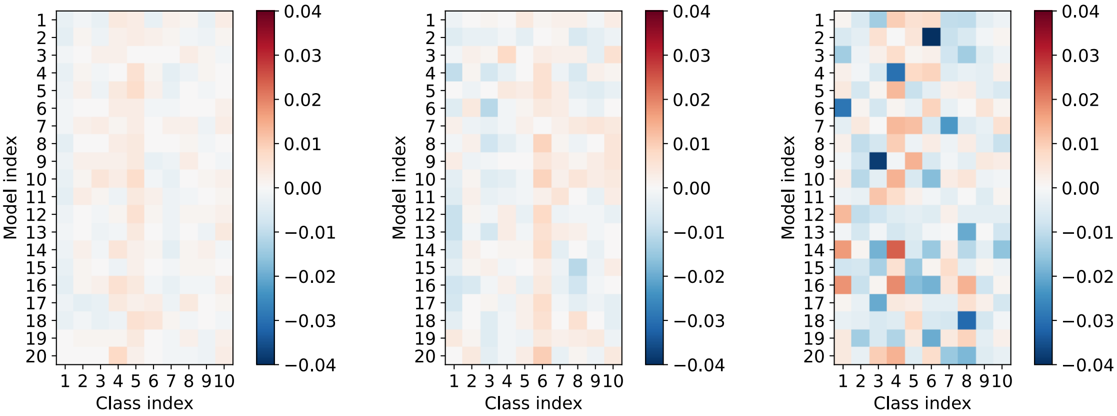


Figure 5.7: Robustness of trained deep ensembles to the removal of submodels at test time

We noted that models trained independently show limited cooperative potential at test time, as suggested by their considerable resilience to the pruning of members, evidencing significant redundancies. Although ensembles trained with GNCL provided greater diversity than those trained independently, the method was equally unable to encourage significant cooperation between its members, as is evident by the almost equal amount of redundancy. The performance of both ensembles only started to decrease significantly after more than half of their members had already been removed, highlighting a severe underutilisation of the added learning capacity with $M \geq 10$.

On the other hand, the members within ensembles trained with SNCL exhibited a much greater interdependence, evidenced by rapid declines in ensemble performance following their removal. We observed significant declines in accuracy across various ensemble sizes $M \in \{5, 10, 20\}$ following the removal of just a single member.

We proceeded by considering a more granular view of the impact of model pruning on the per-class accuracies of the ensembles. Figure 5.8 provides a deeper look at the class-specific accuracy drops after the removal of each individual member from the respective ensemble. We have found that removing individual submodels from the ensembles trained independently and with GNCL impacts all class accuracies of the remaining ensembles similarly little.

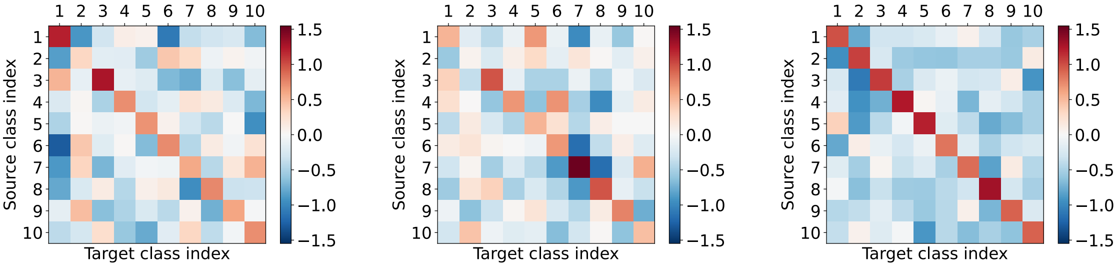


(a) Independent ($\lambda = 0$)

(b) GNCL ($\lambda = 0.5$)

(c) SNCL ($\lambda = 0.5$)

Figure 5.8: Difference in ensemble per-class accuracies following the removal of submodel with model index. Ensemble size $M = 20$



(a) Model 6

(b) Model 7

(c) Model 13

Figure 5.9: Selection of model-specific weight matrices attributed by the meta-model of an SNCL ensemble with $M = 20$ and $\lambda = 0.5$.

On the other hand, our results suggest that a considerable number of members within the SNCL ensemble are specialised to perform well on a subset of classes, evidenced by a significant drop in class-specific accuracies of the remaining ensemble, as a consequence of their removal (fig. 5.8c). We observed similar characteristics for smaller ensembles $M = 5$, depicted in figure B.4 of the appendix. We observed no specialised models for classes 2, 9 and 10, where the ensemble already demonstrates relatively high accuracies. As we have seen in figure 5.3, these classes correspond to those which are most easily distinguishable from the rest.

To gain further insight into how SNCL can train specialists, we proceeded with a qualitative analysis, visualising a selection of weight matrices that were attributed to the individual models 6, 7 and 13 by the meta-learner during training. Remember that with SNCL, the meta-model is optimised in parallel and computes the ensemble prediction through matrix multiplication of the concatenated submodel confidences with the weight matrix. Note that while the original weight matrix W is of dimensions 200×10, in the following, we focus on model-specific sections w_m of size 10×10.

Figure 5.9 visualises the weight matrices of models 6, 7 and 13. Values on the diagonals of these matrices are element-wise weightings of the class probabilities predicted by the submodel, which contribute to the same class in the final ensemble prediction, establishing a baseline. On the other hand, the non-diagonal elements of the weight matrices model linear relationships in-between the class confidences emitted by the individual models. We proceeded with a qualitative analysis of the specific examples to explore mechanisms through which the meta-model enables greater cooperation.

Firstly, as we have seen in figure 5.8c, models 6 and 7 significantly contribute to the accuracy of the ensemble for the respective classes 1 and 7. A look at their weight matrices in figures 5.9a and 5.9b reveals that the classes for which models 6 and 7 have specialised coincide with strong positive weights on the diagonal relative to their other class predictions. As such, the meta-model processes these class predictions to establish a baseline for the final prediction.

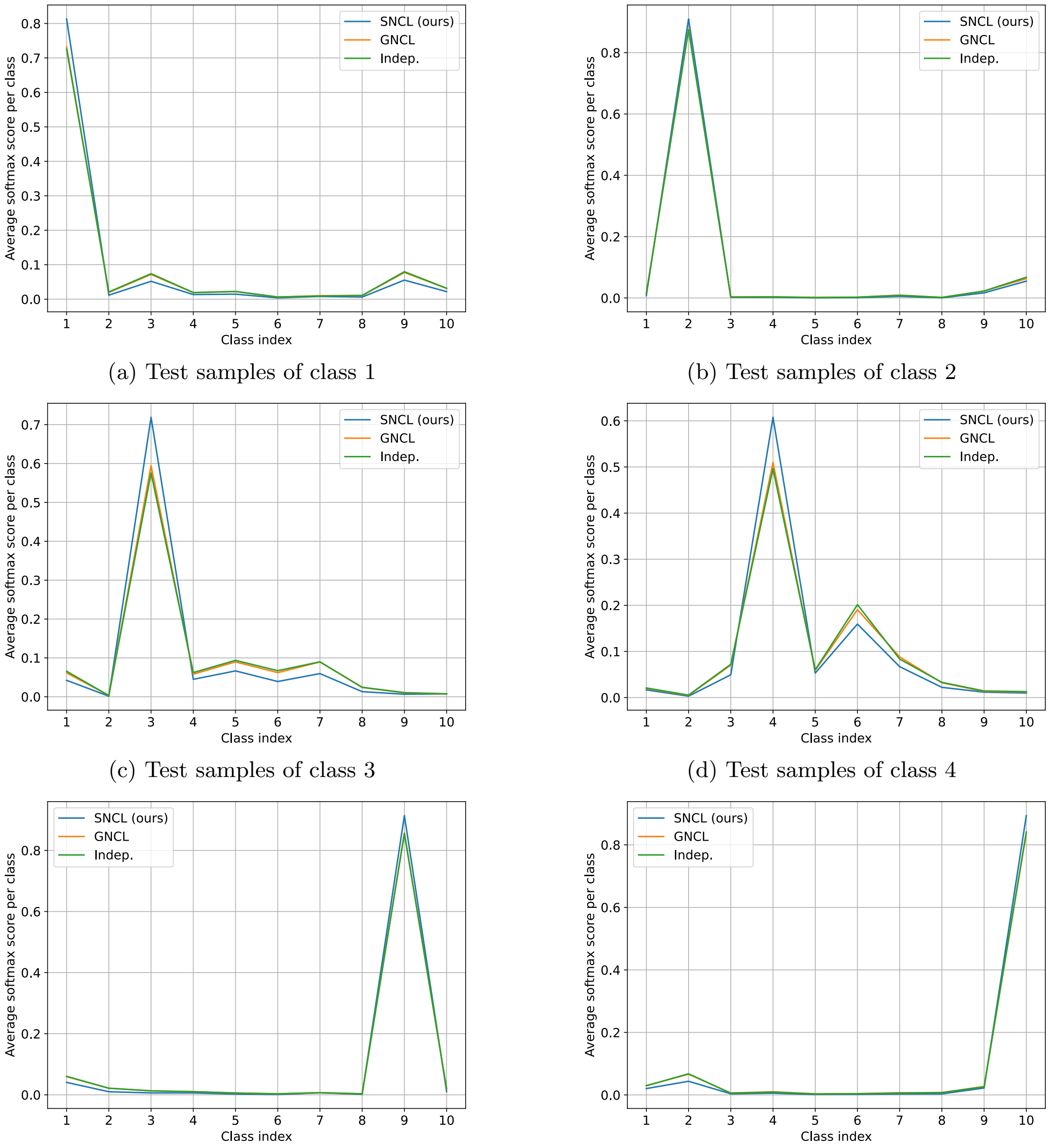
Figure 5.9a displays the weights applied to the probability vectors of the sixth submodel. Besides its strongly positively weighted predictions for class one, we also observed strong negative weights in column one for classes 2 and 6 through 8. Since the chosen meta-learner models linear equations, we can infer that the model's errors for class one are offset by its own emitted probabilities for those classes. Therefore the meta-model is effectively only allowing it to positively contribute to the ensemble's predicted probability for class one if it is unlikely to have confused it with the weighted sum of this subset of classes.

We next considered an example which illustrates that this error correction is not limited to individual models but spans across ensemble members. Looking at the second column in the weight matrix of model six in figure 5.9a, we can see that most weights, including the diagonal element, are weighted very close to zero. The only weight significantly non-zero is for the class probability of class one. Thus, the ensemble uses the confidence for class one emitted by model six to negatively offset its predicted probability of class two.

Another notable phenomenon can be seen in figure 5.9b, which depicts the weight matrix of the seventh model in the SNCL ensemble. The weights for the models' predictions of class two are almost entirely neglected, as all weights in row two are close to zero. Therefore, its predictions for class two are neither significantly considered in the final prediction nor utilised in error corrections for other class predictions. Thereby evidencing that the meta-learner neglects some of the class probabilities emitted by the submodels during optimisation.

Interestingly, we also found multiple models in the ensemble that resemble generalists. One of these is model 13, whose weights we depict in figure 5.9c. We observed similarly strong positive weights on the diagonal of its weight matrix with negative and zero weights on non-diagonal elements.

We further examined the average softmax scores emitted by the models for each of the 10.000 test samples in CIFAR-10, filtered by the true class label, to assess and compare the magnitude of confusions between them in figure 5.10. On average, we observed significantly reduced confidences in non-truth class confidences for SNCL ensembles, where specialists have been trained. On the contrary, we have observed little difference in the predictive margins of classes 2, 9 and 10, for which we find fewer confusions with other classes and have not noticed significantly specialised models in the SNCL ensembles, as evidenced by figure 5.8c.



(a) Test samples of class 1

(b) Test samples of class 2

(c) Test samples of class 3

(d) Test samples of class 4

(e) Test samples of class 9

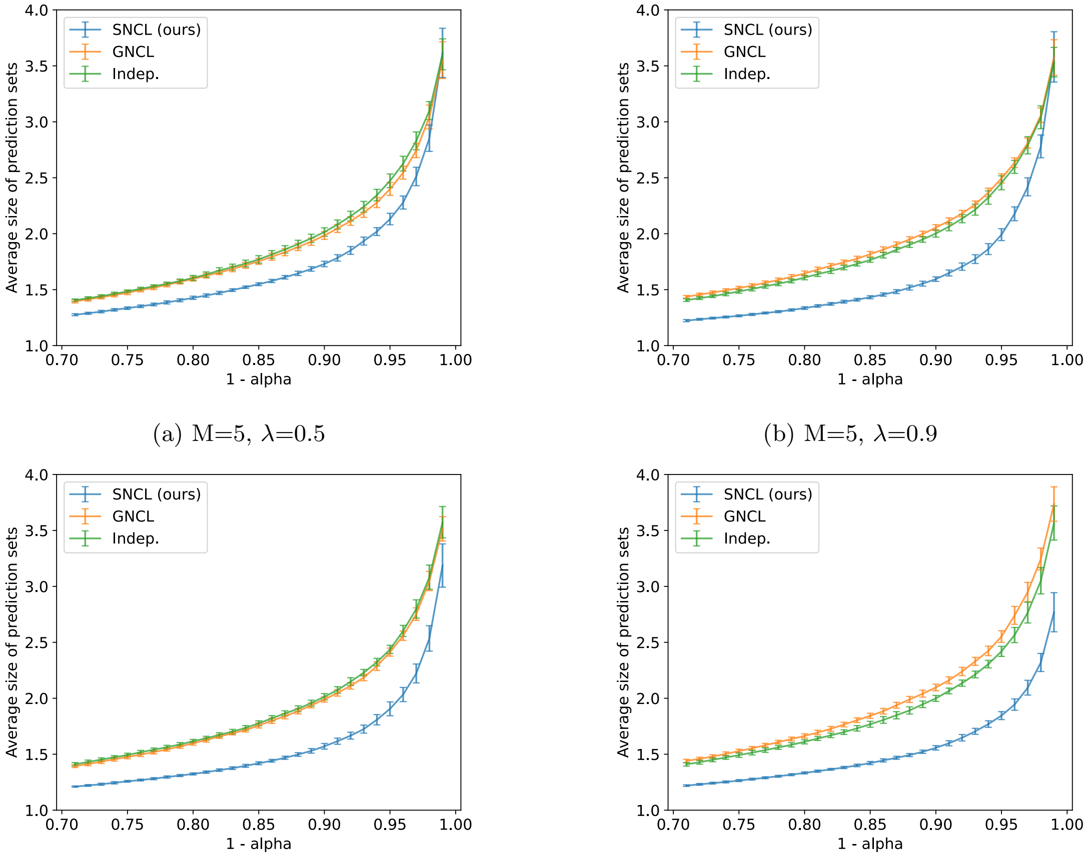
(f) Test samples of class 10

Figure 5.10: Average softmax outputs across CIFAR-10 test samples filtered by true class index. Ensembles trained with $M = 20$ and $\lambda = 0.5$. 1.000 samples per class.

5.3 Predictive Uncertainty

5.3.1 In-domain Uncertainty

Figure 5.11 depicts the average prediction set sizes derived from conformal prediction with adaptive prediction set sizes. We computed error bars across 20 trials, with randomised data splits used for calibration and testing.



(a) M=5, λ=0.5

(b) M=5, λ=0.9

(c) M=20, λ=0.5

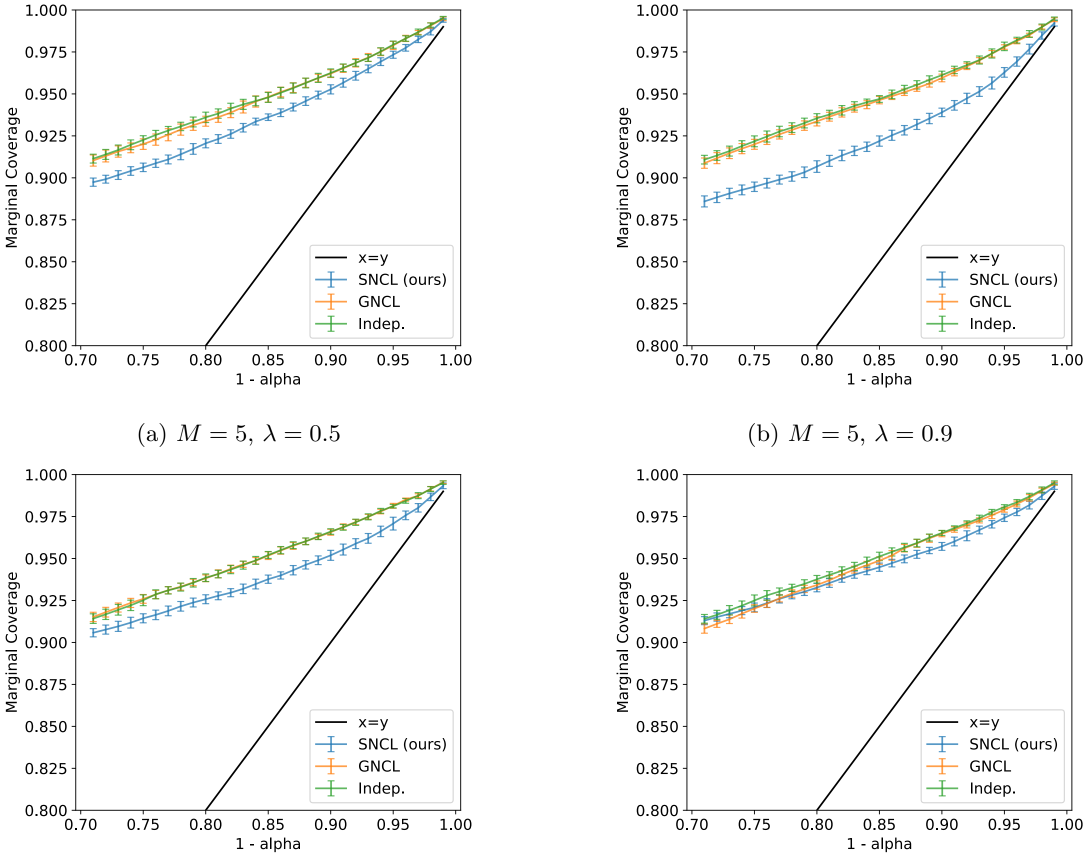
(d) M=20, λ=0.9

Figure 5.11: In-domain uncertainty: Average prediction set sizes using conformal prediction with adaptive set sizes at different significance levels for a selection of ensemble configurations

Overall, we observed that SNCL ensembles were able to provide significantly smaller prediction sets for most relevant significance levels $\alpha \in [0.01, 0.3]$, compared to GNCL and independent training. For small $\alpha \leq 0.02$ and $M = 5$, we observed that SNCL performs on par with the alternative methods, otherwise, our results suggest that SNCL ensembles exhibit significantly less in-domain uncertainty for both small and large ensemble sizes $M \in \{5, 20\}$.

Our results suggest that the gap in in-domain uncertainty between SNCL and the alternatives widens with increasing λ and for larger ensemble sizes M. Furthermore, it was found that inducing explicit diversity with GNCL does not yield any benefits to in-domain uncertainty compared to independently trained ensembles.

Plots of the marginal coverages in figure 5.12 reveal that GNCL and independently trained ensembles tend to provide overly conservative marginal coverages, resulting in larger prediction sets. Overall we find that SNCL provides less conservative marginal coverages, which are more in line with the respective significance level α.



(a) $M = 5$, $\lambda = 0.5$

(b) $M = 5$, $\lambda = 0.9$

(c) $M = 20$, $\lambda = 0.5$

(d) $M = 20$, $\lambda = 0.9$

Figure 5.12: In-domain uncertainty: Average prediction set size using conformal prediction with adaptive set sizes at various significance levels

As a post-hoc test, we also examined the conditional coverages, specific to individual classes, which validated our observations for the marginal coverages. Plots of which we provide for a selection of classes in the appendix, in figures B.5 and B.6.

5.3.2 OOD Detection

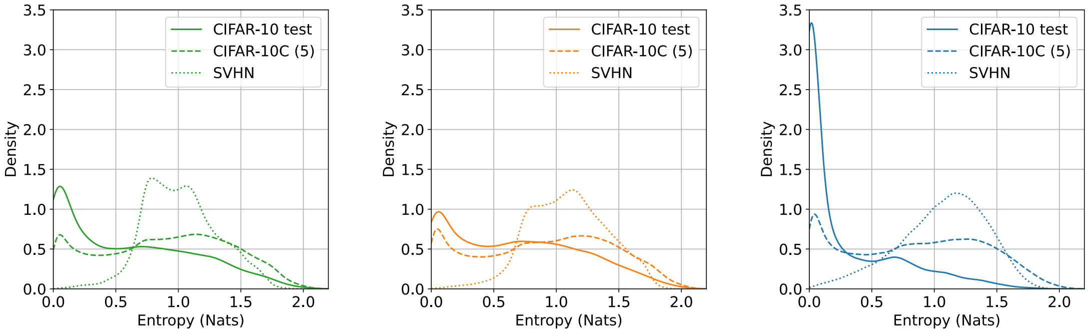
Table 5.1 summarises and compares the results of ensembles trained with the three alternative methods on CIFAR-10. We averaged our results across 5 trials performed with models trained with the respective parameters. We measured the OOD detection capabilities of the models through the AUROC and TNR@TPR95. Our results suggest that SNCL ensembles are able to better discern between ID and OOD data, for the most severe covariate shift in the CIFAR-10C dataset and the semantically different SVHN dataset. Our results are consistent across both evaluation metrics and all tested configurations.

Table 5.1: OOD detection performance for covariate (OOD: CIFAR-10C (5)) and semantic shift (OOD: SVHN) of SNCL, GNCL and independent training on CIFAR-10. All numbers are in percent. Best AUROC and TNR@TPR95 per configuration marked in bold.

ID: CIFAR-10			OOD: CIFAR-10C (5)		OOD: SVHN	
Method	M	λ	AUROC	TNR@TPR95	AUROC	TNR@TPR95
Indep.	5	-	67.91 ± 0.22	19.99 ± 0.95	76.61 ± 0.15	55.87 ± 0.14
GNCL	5	0.5	69.55 ± 0.27	21.90 ± 1.01	78.63 ± 0.53	56.22 ± 0.45
SNCL	5	0.5	$\mathbf{79.40 \pm 0.17}$	$\mathbf{40.02 \pm 0.54}$	$\mathbf{87.78 \pm 0.08}$	$\mathbf{69.28 \pm 0.13}$
Indep.	5	-	67.91 ± 0.22	19.99 ± 0.95	76.61 ± 0.15	55.87 ± 0.14
GNCL	5	0.9	67.52 ± 0.24	19.20 ± 0.77	78.80 ± 0.10	53.42 ± 0.53
SNCL	5	0.9	$\mathbf{87.62 \pm 0.13}$	$\mathbf{58.34 \pm 0.33}$	$\mathbf{92.60 \pm 0.17}$	$\mathbf{75.79 \pm 0.20}$
Indep.	20	-	65.51 ± 0.30	16.97 ± 0.73	74.23 ± 0.22	53.93 ± 0.75
GNCL	20	0.5	66.05 ± 0.34	17.71 ± 0.54	75.97 ± 0.34	54.82 ± 0.34
SNCL	20	0.5	$\mathbf{83.57 \pm 0.19}$	$\mathbf{48.81 \pm 0.31}$	$\mathbf{92.56 \pm 0.10}$	$\mathbf{77.80 \pm 0.52}$
Indep.	20	-	65.51 ± 0.30	16.97 ± 0.73	74.23 ± 0.22	53.93 ± 0.75
GNCL	20	0.9	61.20 ± 0.47	10.92 ± 1.45	73.51 ± 0.16	47.69 ± 0.44
SNCL	20	0.9	$\mathbf{81.04 \pm 0.14}$	$\mathbf{39.96 \pm 0.26}$	$\mathbf{91.12 \pm 0.07}$	$\mathbf{72.44 \pm 0.12}$

Since the detection of data with less severe covariate shifts is more difficult, we also compared the performance decline in OOD detection for all methods and configurations. We have found that the performance gap between SNCL and its alternatives remains stable, as we noted similar relative declines in AUROC for the intensities of 1-4 in CIFAR-10C, illustrated in figure B.8 of the appendix. Across methods, we have observed significant decreases in ensemble accuracy in accordance with the intensity of covariate shifts, see figure B.7 in the appendix.

Examining the exemplary entropy densities of the three methods depicted in figure 5.13, we observe that for in-distribution data, the SNCL ensembles exhibited significantly lower entropy than the alternatives, confirming our previous observations of lower in-domain uncertainty. Across methods, we observed similar entropy levels under covariate and semantic shift, respectively.



(a) Independent ($\lambda = 0$)

(b) GNCL ($\lambda = 0.9$)

(c) SNCL ($\lambda = 0.9$)

Figure 5.13: Density estimates of the softmax entropies of ensembles of size $M = 20$ on various datasets

Chapter 6

Discussion

In this chapter, we interpret and discuss our experimental results and draw comparisons to relevant and recent publications.

Our experimental results have shown that stacked negative correlation learning (SNCL) overcomes several limitations encountered by previous attempts at cultivating specialists in deep ensembles. In contrast to stochastic multiple choice learning [17, 26], SNCL overcomes the issue of overspecialisation. Our results have demonstrated that the weak learners in SNCL ensembles surpass random guessing for all classes. Furthermore, we have addressed a limitation highlighted by Hinton et al. [19] by incorporating a holistic and parallelised process that leverages the discriminative performance of the ensemble members to guide their specialisation. Similar to their multi-step approach, we observed that SNCL ensembles also include generalist models to establish a baseline for the ensemble's predictions. Notably, SNCL implicitly chose the proportion of generalists and specialists during the optimisation procedure, as opposed to the explicit choices made by Hinton et al. [19].

Based on our research findings, it is evident that the primary contribution of SNCL is the introduction of an interactive training scheme to stacked ensembles, which led to the observed benefits. We have found strong evidence suggesting that this feedback loop allows the meta-model to guide the learning trajectories of the submodels towards specialisation, without penalising them for neglected classes in the covariance penalty term of the SNCL loss function. Thus, SNCL follows a division-of-labour approach to the presented classification problems, which resulted in significantly reduced redundancies within ensembles in comparison to the tested state-of-the-art methods [4, 11], as evidenced by the results from a robustness experiment we performed analogously to Webb et al. [16].

We closely analyzed the weight matrices of a single-layer meta-model of an SNCL ensemble and identified three main mechanisms through which it instilled specialisation and demonstrated reduced generalisation error and in-domain uncertainty:

1. **Distribution:** The meta-learner distributes subsets of classes to the submodels, enabling them to specialise in distinguishing between them, without being penalised for specialisation during optimisation in the penalty term of the SNCL loss function. The notion of assigning class subsets to submodels was similarly described in the works of Lee et al. [17] and Hinton et al. [19].

2. **Error correction:** The meta-learner uses the specialist's predictions to facilitate error corrections in the baseline predictions made by their peers, effectively enhancing the predictive margins between confusable classes and thus the overall generalisation accuracy.

3. **Muting:** The concept of muting resembles the idea of a dustbin class for specialists, as discussed by Hinton et al. [19], where specific class predictions of specialists are effectively ignored in the computation of the final ensemble prediction. This mechanism allows submodels to neglect certain classes during training, without hurting the performance of the ensemble's predictions, as would be the case with averaging.

Our experimental results have shown that the mere addition of a meta-model to an ensemble trained with GNCL yields limited benefits to its test accuracy, highlighting the importance of the concurrent training scheme introduced by SNCL. In light of these findings, we hypothesize that specialisation within the label space introduces structure to the induced diversity, which is more advantageous in conjunction with a meta-model than pursuing diversity solely for the sake of it, as in GNCL [11]. It is conceivable that the diversity induced through specialisation is structured by class, which enables the meta-model to craft a better hypothesis of the level-1 data, enhancing its bias- and variance-reducing capabilities [20], ultimately reducing the generalisation error. Although our results suggest that SNCL submodels generally exhibit greater rates of disagreement than those of GNCL ensembles for the same $M \geq 5$ and $\lambda \geq 0.5$, we found that SNCL strongly outperforms GNCL even with comparable diversity, reflecting the results of Buschjäger et al. [11] which indicate that not all diversity is equally beneficial to the accuracy of an ensemble.

In alignment with the observations made by Hinton et al. [19], our findings underscore the significance of specialisation within ensembles when dealing with high-dimensional classification problems with confusable classes. We have seen evidence that the specialists in SNCL ensembles provide enhanced predictive margins between confusable classes compared to state-of-the-art methods which lack specialisation. Although we would argue that any non-trivial classification problem includes confusable classes, we expect a limited relative benefit to the use of SNCL for problems with easily distinguishable classes or low dimensional label spaces which lack the potential for specialisation, e.g. binary classification.

Our results suggest that SNCL provides some unique advantages for practical applications. Firstly, because SNCL significantly reduces redundancies within ensembles, it greatly reduces the need for pruning, such that the ensemble size M can be reliably set based on a predetermined computational budget, making for efficient budgeting and less wasted training resources.

Secondly, we have noted lesser diminishing returns with $M \geq 10$ for SNCL compared to the state-of-the-art, given its improved utilisation of the allocated learning capacity. As some studies have demonstrated, ensembles can be more accurate and computationally efficient than scaling a single model of equal model complexity [14, 15]. Our results suggest that SNCL could expand the design space in which deep ensembles are preferable to individual models in terms of accuracy and efficiency, or enhance the observed benefits in already established scenarios [15].

Finally, in our testing, SNCL ensembles exhibited reduced in-domain uncertainty, which enabled more accurate OOD detection, based on thresholding performed on the softmax entropy. Our results, therefore, suggest that deep ensembles trained with SNCL could provide enhanced uncertainty estimates for deployment in safety-critical domains. Like recent studies [45, 46], we have found no meaningful relationship between an ensembles diversity and the quality of its uncertainty estimates. In our testing, GNCL performed on par with independently trained ensembles regarding in-domain uncertainty and OOD detection, despite greater diversity.

Chapter 7

Conclusion

We now provide a conclusion to this thesis and concisely answer the three research questions, based on our experimental results and the previous discussion. At the end of this chapter, we highlight areas for future research.

Our extensive experiments have provided robust evidence for the efficacy of our proposal, Stacked Negative Correlation Learning (SNCL), to construct more efficient ensembles for high-dimensional classification tasks, by enabling weak learners to specialise within the label space. In our testing, SNCL was able to significantly reduce redundant learnings in deep ensembles, enhancing their generalisation accuracy and reducing their in-domain uncertainty, which resulted in improved OOD detection capabilities.

As a technique founded on generalised negative correlation learning and stacked generalisation, SNCL combines the strengths of both approaches while mitigating their shortcomings. It facilitates mutual interaction between the meta-learner and the submodels through a concurrent training scheme, thereby prompting specialisation among the base learners based on their discriminative performances. Our results evidence that the meta-learner actively guides the training trajectories of the ensemble members, promoting their expertise in discriminating between subsets of classes. In this study, we identified several mechanisms through which the meta-model promotes specialisation and more effectively aggregates the predictions of its constituents to enhance the predictions of the ensemble.

Therefore, we conclude that SNCL offers unique advantages and should be favoured over GNCL and independently trained ensembles for high-dimensional classification tasks, especially in safety-critical settings where accurate uncertainty quantification is paramount.

7.1 Research Questions

RQ1: *How is the proposed SNCL approach able to train specialised weak learners?*

The key innovation through which SNCL manages to train specialists is the introduction of an interactive training scheme between the meta-model and the submodels in stacked ensembles, which encompasses an adapted GNCL loss function through which the meta-models' weights can influence the optimisation trajectories of the weak learners. We have identified and discussed multiple mechanisms through which SNCL promotes specialisation and effectively reduces redundant learnings within deep ensembles, by following a division-of-labour approach it achieves greater utilisation of the allocated learning capacity. Our in-depth analysis of the weight matrices attributed to the submodels by the meta-learner has demonstrated that it effectively assigns confusable class subsets to the weak learners based on their ability to discriminate between them, thereby encouraging them to specialise within the label space of the classification problem.

RQ2: *To what degree can SNCL reduce the generalisation error of deep ensembles?*

We have observed significantly reduced generalisation errors for SNCL ensembles compared to state-of-the-art deep ensembling methods [4, 11] with equivalent model complexities on the benchmark datasets CIFAR-10 and CIFAR-100. Our results suggest that SNCL better utilises the allocated learning capacity by following a division-of-labour approach through specialisation. In our testing, the predictions of SNCL ensembles provided greater predictive margins between confusable classes. Our evidence suggests that the diversity induced through specialisation introduces a structure to the level-1 data which can lead to enhanced bias and variance reductions by the meta-model in the stacked ensemble. Moreover, our findings indicate that SNCL exhibits reduced diminishing returns as the ensemble size increases, particularly for ensemble sizes $M \geq 10$, resulting in a reduced generalisation error.

RQ3: *How does SNCL impact the quality of uncertainty estimates in deep ensembles?*

Our findings indicate that SNCL can significantly reduce the in-domain uncertainty of deep ensembles, as specialised submodels effectively enhance the predictive margins between confusable classes within high-dimensional classification problems. We noted greater relative reductions of in-domain uncertainty between SNCL and the state-of-the-art alternatives in correspondence with greater M and λ. As a result of overall reduced levels of in-domain uncertainty, we observed benefits to its performance in OOD detection for covariate and semantic shift, for small ($M = 5$) and large ($M = 20$) ensemble sizes. Therefore, we conclude that SNCL can positively influence the quality of the uncertainty estimates provided by deep ensembles.

7.2 Future Work

Finally, based on our findings, we have identified several promising avenues for future research. One exciting direction is the incorporation of methods for implicit diversity induction into SNCL, specifically heterogeneous neural architectures [9]. We postulate that specific architectures may be inherently better at distinguishing between certain class subsets than others, by focusing on different sets of features within the data, resulting in unique strengths which could be effectively harnessed with SNCL.

Simultaneously, future research could strive to increase the computational efficiency of SNCL, by introducing shared parameters to the submodels. As previously highlighted, related studies have shown that parameter-sharing, especially within the initial convolutional layers of CNN models, can significantly reduce the computational costs during training and inference, with limited sacrifices to ensemble diversity [13, 26, 27]. This highlights an intriguing prospect for developing more efficient, high-performing deep ensembles with our proposed methodology.

Lastly, since SNCL seems to scale better for large ensemble sizes $M \geq 10$, where traditional approaches typically encounter diminishing returns, it may provide greater opportunities to substitute single large models with more efficient ensembles, which are made up of smaller specialised submodels [14, 15].

Bibliography

[1] M.A. Ganaie, Minghui Hu, A.K. Malik, M. Tanveer, and P.N. Suganthan. “Ensemble deep learning: A review”. In: *Engineering Applications of Artificial Intelligence* 115 (2022), p. 105151. ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2022.105151.

[2] Luis A. Ortega, Rafael Cabañas, and Andrés R. Masegosa. “Diversity and Generalization in Neural Network Ensembles”. In: *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*. Vol. 151. PMLR. Valencia, Spain, 2022. URL: http://proceedings.mlr.press/v151/ortega22a.html.

[3] Suyash Kumar, Prabhjot Kaur, and Anjana Gosain. “A Comprehensive Survey on Ensemble Methods”. In: *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*. 2022, pp. 1–7. DOI: 10.1109/I2CT54291.2022.9825269.

[4] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

[5] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. “Can you trust your model’s uncertainty? Evaluating predictive uncertainty under dataset shift”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf.

[6] Thomas G. Dietterich. “Ensemble Methods in Machine Learning”. In: *Multiple Classifier Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 1–15. ISBN: 978-3-540-45014-6.

[7] Leo Breiman. “Bagging predictors”. In: *Machine Learning* 24.2 (1996), pp. 123–140. ISSN: 1573-0565. DOI: 10.1007/BF00058655.

[8] Prem Melville and Raymond J. Mooney. “Constructing Diverse Classifier Ensembles Using Artificial Training Examples”. In: *Proceedings of the 18th International Joint Conference on Artificial Intelligence*. IJCAI’03. Acapulco, Mexico: Morgan Kaufmann Publishers Inc., 2003, pp. 505–510.

[9] Ulf Johansson and Tuve Löfström. “Producing implicit diversity in ANN ensembles”. In: *The 2012 International Joint Conference on Neural Networks (IJCNN)*. 2012, pp. 1–8. DOI: 10.1109/IJCNN.2012.6252713.

[10] Y. Liu and X. Yao. “Ensemble learning via negative correlation”. In: *Neural Networks* 12.10 (1999), pp. 1399–1404. ISSN: 0893-6080. DOI: `https://doi.org/10.1016/S0893-6080(99)00073-8`.

[11] Sebastian Buschjäger, Lukas Pfahler, and Katharina Morik. “Generalized Negative Correlation Learning for Deep Ensembling”. In: *ArXiv* abs/2011.02952 (2020).

[12] Wenjing Li, Randy Clinton Paffenroth, and David Berthiaume. “Neural Network Ensembles: Theory, Training, and the Importance of Explicit Diversity”. In: *ArXiv* abs/2109.14117 (2021).

[13] Michael Opitz, Horst Possegger, and Horst Bischof. “Efficient Model Averaging for Deep Neural Networks”. In: Mar. 2017, pp. 205–220. ISBN: 978-3-319-54183-9. DOI: `10.1007/978-3-319-54184-6_13`.

[14] Dan Kondratyuk, Mingxing Tan, Matthew Brown, and Boqing Gong. *When Ensembling Smaller Models is More Efficient than Single Large Models.* 2020. arXiv: `2005.00570 [cs.LG]`.

[15] Abdul Wasay and Stratos Idreos. “More or Less: When and How to Build Convolutional Neural Network Ensembles”. In: *International Conference on Learning Representations.* 2021.

[16] Andrew Webb, Charles Reynolds, Wenlin Chen, Henry Reeve, Dan Iliescu, Mikel Luján, and Gavin Brown. “To Ensemble or Not Ensemble: When Does End-to-End Training Fail?” In: *Machine Learning and Knowledge Discovery in Databases.* Cham: Springer International Publishing, 2021, pp. 109–123. ISBN: 978-3-030-67664-3.

[17] Stefan Lee, Senthil Purushwalkam, Michael Cogswell, Viresh Ranjan, David Crandall, and Dhruv Batra. “Stochastic Multiple Choice Learning for Training Diverse Deep Ensembles”. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems.* NIPS’16. Barcelona, Spain: Curran Associates Inc., 2016, pp. 2127–2135. ISBN: 9781510838819.

[18] Gavin Brown, Jeremy L. Wyatt, and Peter Tiño. “Managing Diversity in Regression Ensembles”. In: *Journal of Machine Learning Research* 6.55 (2005), pp. 1621–1650. URL: `http://jmlr.org/papers/v6/brown05a.html`.

[19] Geoffrey Hinton, Jeff Dean, and Oriol Vinyals. “Distilling the Knowledge in a Neural Network”. In: *Advances in Neural Information Processing Systems (NIPS) - Deep Learning Workshop.* Mar. 2014, pp. 1–9.

[20] David H. Wolpert. “Stacked generalization”. In: *Neural Networks* 5.2 (1992), pp. 241–259. ISSN: 0893-6080. DOI: `https://doi.org/10.1016/S0893-6080(05)80023-1`.

[21] Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. “Rethinking Bias-Variance Trade-off for Generalization of Neural Networks”. In: *Proceedings of the 37th International Conference on Machine Learning.* ICML’20. JMLR.org, 2020.

[22] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. “Reconciling modern machine-learning practice and the classical bias–variance trade-off”. In: *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15849–15854. DOI: `10.1073/pnas.1903070116`. eprint: `https://www.pnas.org/doi/pdf/10.1073/pnas.1903070116`.

[23] Stuart Geman, Elie Bienenstock, and René Doursat. “Neural Networks and the Bias/Variance Dilemma”. In: *Neural Computation* 4.1 (1992), pp. 1–58. ISSN: 0899-7667. URL: `http://portal.acm.org/citation.cfm?id=148062`.

[24] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. “Deep Ensembles: A Loss Landscape Perspective”. In: *ArXiv* abs/1912.02757 (2019).

[25] Grigorios Tsoumakas, Ioannis Partalas, and Ioannis Vlahavas. “An Ensemble Pruning Primer”. In: *Applications of Supervised and Unsupervised Ensemble Methods*. Ed. by Oleg Okun and Giorgio Valentini. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 1–13. ISBN: 978-3-642-03999-7. DOI: 10.1007/978-3-642-03999-7_1.

[26] Stefan Lee, Senthil Purushwalkam, Michael Cogswell, David Crandall, and Dhruv Batra. *Why M Heads are Better than One: Training a Diverse Ensemble of Deep Networks*. 2015. arXiv: 1511.06314 [cs.CV].

[27] Marton Havasi, Rodolphe Jenatton, Stanislav Fort, Jeremiah Zhe Liu, Jasper Snoek, Balaji Lakshminarayanan, Andrew Mingbo Dai, and Dustin Tran. “Training independent subnetworks for robust prediction”. In: *International Conference on Learning Representations*. 2021.

[28] Cristian Buciluundefined, Rich Caruana, and Alexandru Niculescu-Mizil. “Model Compression”. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’06. Philadelphia, PA, USA: Association for Computing Machinery, 2006, pp. 535–541. ISBN: 1595933395. DOI: 10.1145/1150402.1150464.

[29] Jeremy Nixon, Balaji Lakshminarayanan, and Dustin Tran. “Why Are Bootstrapped Deep Ensembles Not Better?” In: *”I Can’t Believe It’s Not Better!” NeurIPS 2020 workshop*. 2020.

[30] Alan Jeffares, Tennison Liu, Jonathan Crabbé, and Mihaela van der Schaar. *Joint Training of Deep Ensembles Fails Due to Learner Collusion*. 2023. arXiv: 2301.11323 [cs.LG].

[31] Robert McKay and Hussein Abbass. “Analysing Anti correlation in Ensemble Learning”. In: *Proceedings of 2001 Conferenceon Artificial Neural Networks and Expert Systems*. Jan. 2001, pp. 22–27.

[32] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. “Diversity With Cooperation: Ensemble Methods for Few-Shot Classification”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 3722–3730. DOI: 10.1109/ICCV.2019.00382.

[33] Hyunjin Kwon, Jinhyeok Park, and Youngho Lee. “Stacking Ensemble Technique for Classifying Breast Cancer”. In: *Healthcare Informatics Research* 25 (Oct. 2019), p. 283. DOI: 10.4258/hir.2019.25.4.283.

[34] Dost KHAN. “Sensor-Based Human Activity Recognition Using Deep Stacked Multilayered Perceptron Model”. In: *IEEE Access* 8 (Dec. 2020), pp. 218898–218910. DOI: 10.1109/ACCESS.2020.3041822.

[35] K. M. Ting and I. H. Witten. “Issues in Stacked Generalization”. In: *Journal of Artificial Intelligence Research* 10 (May 1999), pp. 271–289. DOI: 10.1613/jair.594.

[36] Anna Jurek, Yaxin Bi, Shengli Wu, and Chris Nugent. “A survey of commonly used ensemble-based classification techniques”. In: *The Knowledge Engineering Review* 29.5 (2014), pp. 551–581. DOI: 10.1017/S0269888913000155.

[37] Salah A. Faroughi, Ana I. Roriz, and Célio Fernandes. “A Meta-Model to Predict the Drag Coefficient of a Particle Translating in Viscoelastic Fluids: A Machine Learning Approach”. In: *Polymers* 14.3 (2022). ISSN: 2073-4360. DOI: 10.3390/polym14030430. URL: https://www.mdpi.com/2073-4360/14/3/430.

[38] Dan Hendrycks and Kevin Gimpel. “A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks”. In: *International Conference on Learning Representations.* 2017.

[39] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks”. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70.* ICML’17. Sydney, NSW, Australia: JMLR.org, 2017, pp. 1321–1330.

[40] Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. “Pitfalls of In-Domain Uncertainty Estimation and Ensembling in Deep Learning”. In: *International Conference on Learning Representations.* 2020.

[41] Anh Nguyen, Jason Yosinski, and Jeff Clune. “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2015, pp. 427–436. DOI: 10.1109/CVPR.2015.7298640.

[42] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and Harnessing Adversarial Examples”. In: *International Conference on Learning Representations.* 2015. URL: http://arxiv.org/abs/1412.6572.

[43] Thilo Strauss, Markus Hanselmann, Andrej Junginger, and Holger Ulmer. *Ensemble Methods as a Defense to Adversarial Perturbations Against Deep Neural Networks.* 2018. arXiv: 1709.03423 [stat.ML].

[44] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip H. S. Torr, and Yarin Gal. *Deep Deterministic Uncertainty: A Simple Baseline.* 2022. arXiv: 2102.11582 [cs.LG].

[45] Guoxuan Xia and Christos-Savvas Bouganis. *On the Usefulness of Deep Ensemble Diversity for Out-of-Distribution Detection.* 2022. arXiv: 2207.07517 [cs.LG].

[46] Taiga Abe, E. Kelly Buchanan, Geoff Pleiss, Richard Zemel, and John Patrick Cunningham. “Deep Ensembles Work, But Are They Necessary?” In: *Advances in Neural Information Processing Systems.* Ed. by Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho. 2022.

[47] Junjiao Tian, Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. “Exploring Covariate and Concept Shift for Out-of-Distribution Detection”. In: *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications.* 2021.

[48] Eyke Hüllermeier and Willem Waegeman. “Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods”. In: *Machine Learning* 110.3 (Mar. 2021), pp. 457–506. ISSN: 1573-0565. DOI: 10.1007/s10994-021-05946-3.

[49] Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. “Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness”. In: *Advances in Neural Information Processing Systems.* Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 7498–7512. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/543e83748234f7cbab21aa0ade66565f-Paper.pdf.

[50] Anastasios N. Angelopoulos and Stephen Bates. *A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification.* 2022. arXiv: 2107.07511 [cs.LG].

[51] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning.* The MIT Press, Dec. 2008. ISBN: 9780262255103. DOI: 10.7551/mitpress/9780262170055.001.0001.

[52] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, WENXUAN PENG, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. "OpenOOD: Benchmarking Generalized Out-of-Distribution Detection". In: *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track.* 2022.

[53] Dan Hendrycks and Thomas Dietterich. "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations". In: *Proceedings of the International Conference on Learning Representations* (2019).

[54] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. "Reading Digits in Natural Images with Unsupervised Feature Learning". In: *NIPS Workshop on Deep Learning and Unsupervised Feature Learning* (2011).

[55] Y. Hsu, Y. Shen, H. Jin, and Z. Kira. "Generalized ODIN: Detecting Out-of-Distribution Image Without Learning From Out-of-Distribution Data". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* Los Alamitos, CA, USA: IEEE Computer Society, June 2020, pp. 10948–10957. DOI: 10.1109/CVPR42600.2020.01096.

[56] Alex Krizhevsky and Geoffrey Hinton. *Learning multiple layers of features from tiny images.* Tech. rep. 0. Toronto, Ontario: University of Toronto, 2009.

Appendix A

Implementation Details

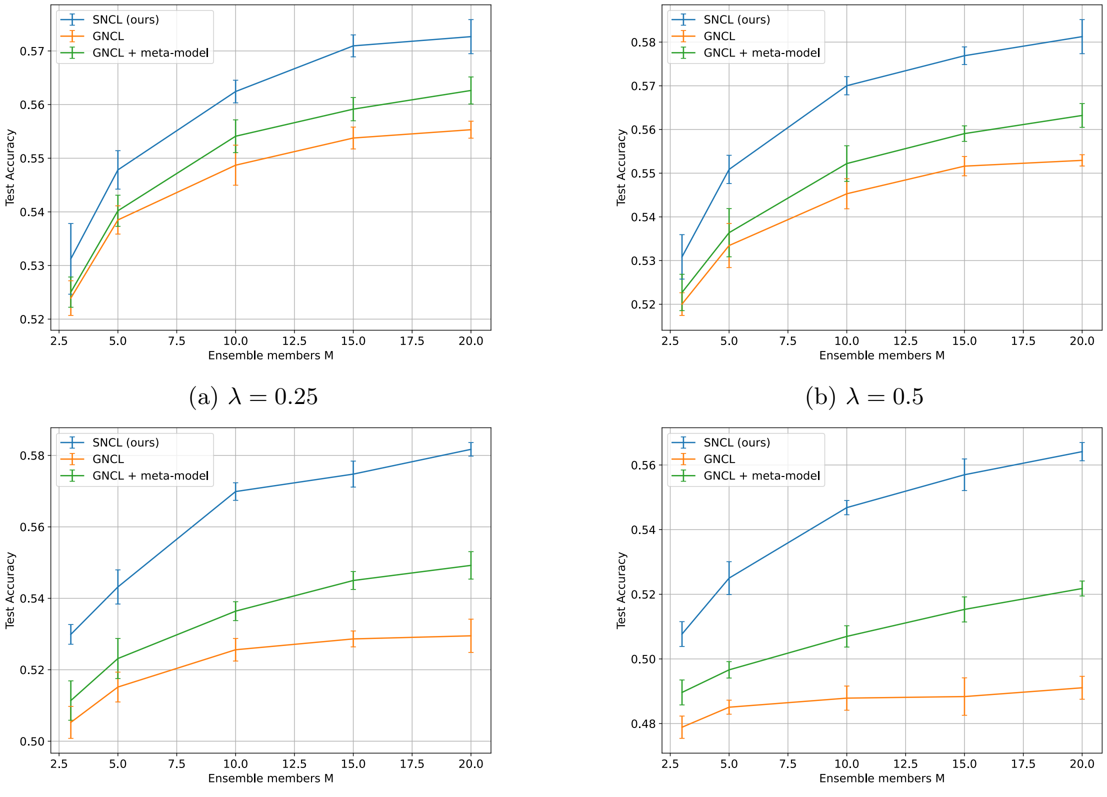
Table A.1: Convolutional Neural Network Architecture for CIFAR-10 and **CIFAR-100** (marked in **bold**). The only difference in the two architectures can be found in the final layer's output channels, which match the number of unique classes in the respective dataset.

Architecture	Layer Type	Input Channels	Output Channels	Kernel Size
Block 1	Conv2d	3	16	3x3
	ReLU	-	-	-
	Dropout(0.2)	-	-	-
	Conv2d	16	16	3x3
	ReLU	-	-	-
	MaxPool2d	-	-	2x2
Block 2	Conv2d	16	32	3x3
	ReLU	-	-	-
	Dropout(0.2)	-	-	-
	Conv2d	32	32	3x3
	ReLU	-	-	-
	MaxPool2d	-	-	2x2
Block 3	Conv2d	32	64	3x3
	ReLU	-	-	-
	Dropout(0.2)	-	-	-
	Conv2d	64	64	3x3
	ReLU	-	-	-
	MaxPool2d	-	-	2x2
Block 4	flatten	-	-	-
	Linear	1024	128	-
	Linear	128	10/**100**	-
	softMax	-	-	-

Appendix B

Supplemental Results

B.1 Accuracy and Diversity



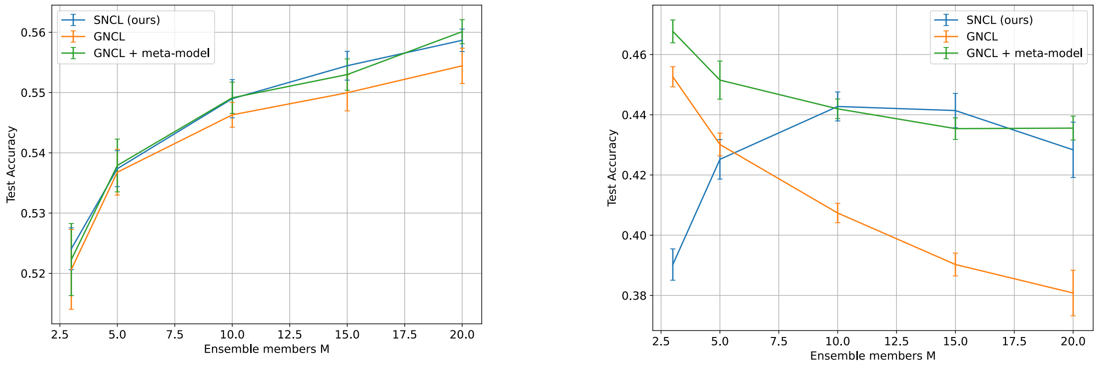
(a) $\lambda = 0.25$

(b) $\lambda = 0.5$

(c) $\lambda = 0.75$

(d) $\lambda = 0.9$

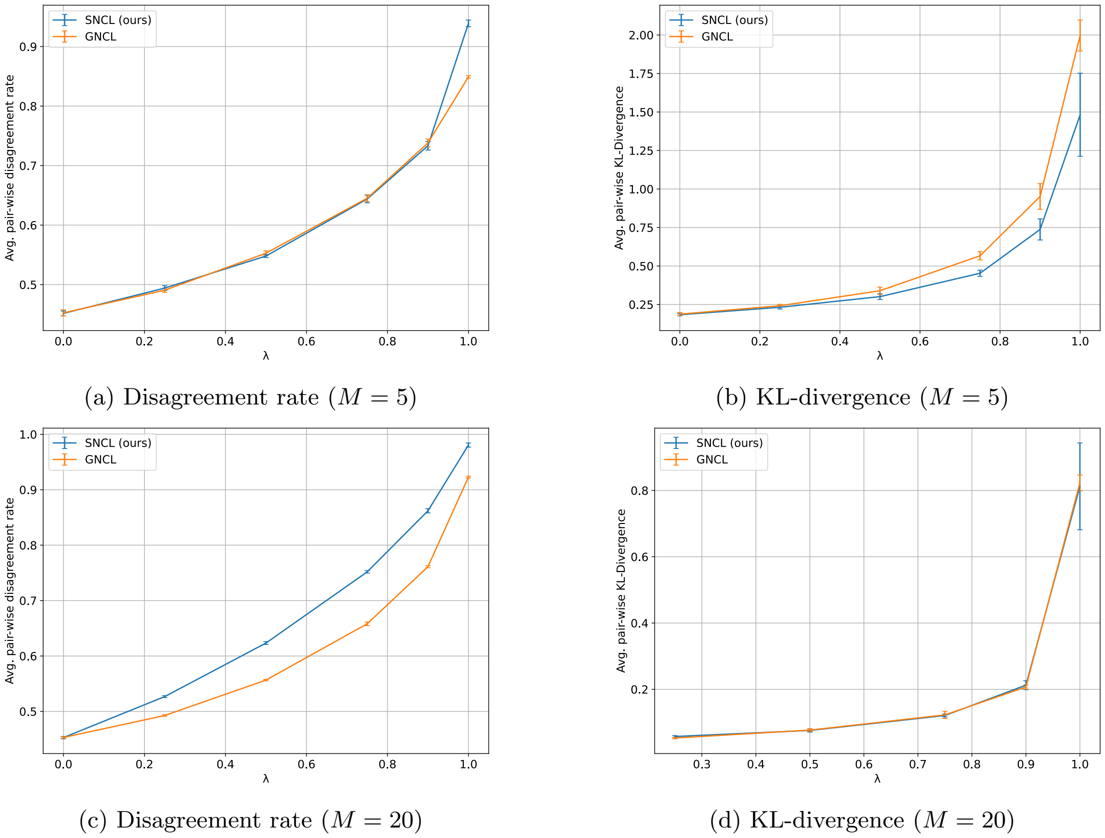
Figure B.1: Test Accuracy of ensembles trained with SNCL and GNCL on CIFAR-100 for a variety of ensemble sizes M and λ between independent and end-to-end training. Standard error bars computed across 10 trials each.



(a) Independent training ($\lambda = 0$)

(b) End-to-end training ($\lambda = 1$)

Figure B.2: CIFAR-100 test accuracy of SNCL and GNCL with independent and end-to-end training for the tested ensemble sizes



(a) Disagreement rate ($M = 5$)

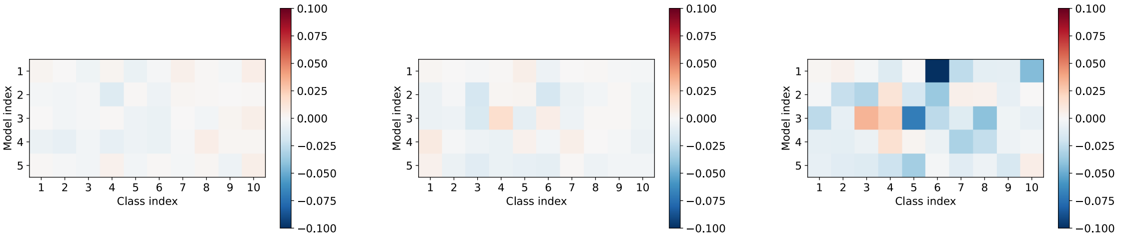
(b) KL-divergence ($M = 5$)

(c) Disagreement rate ($M = 20$)

(d) KL-divergence ($M = 20$)

Figure B.3: Discrepancy of pair-wise diversity metrics of ensembles trained with GNCL and SNCL on CIFAR-100

B.2 Cooperation and specialisation



(a) Independent ($\lambda = 0.0$)

(b) GNCL ($\lambda = 0.5$)

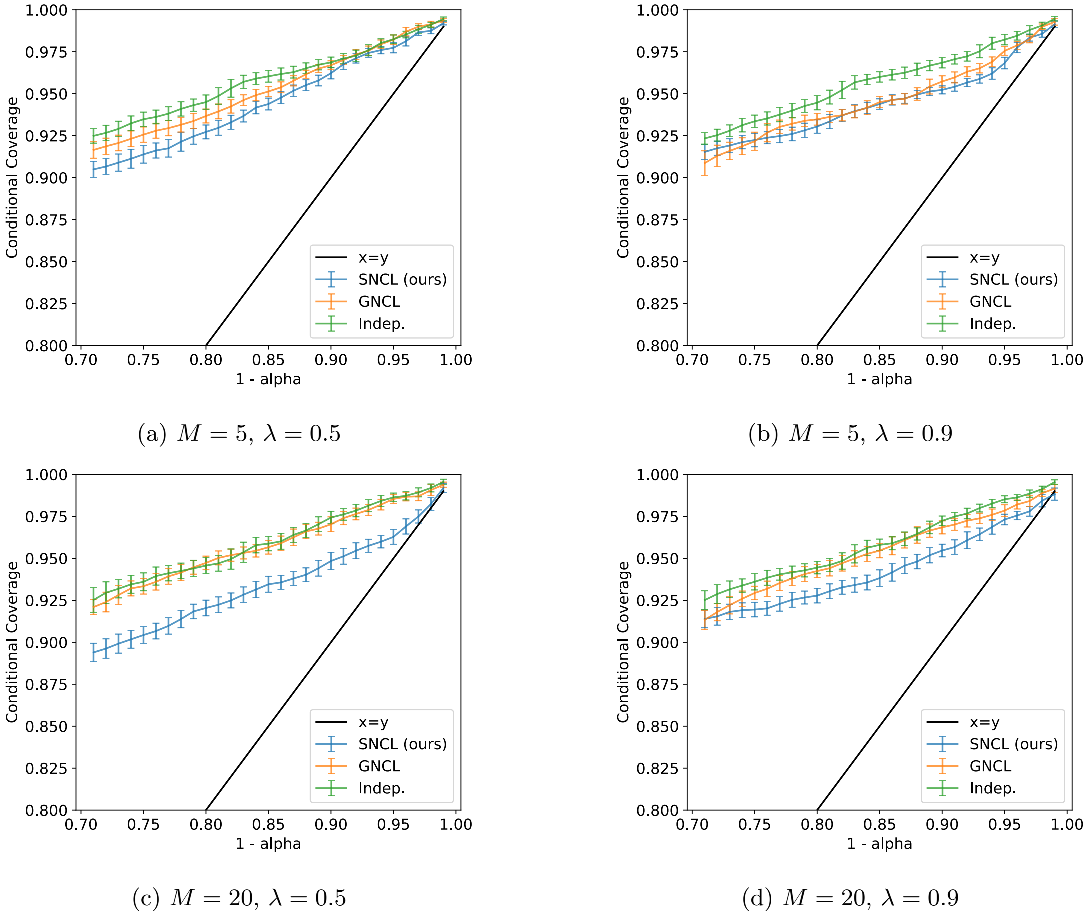
(c) SNCL ($\lambda = 0.5$)

Figure B.4: Change in ensemble per-class accuracies following the removal of submodel with model index. Ensemble size $M = 5$

B.3 Uncertainty

B.3.1 In-domain Uncertainty

Conditional Coverage



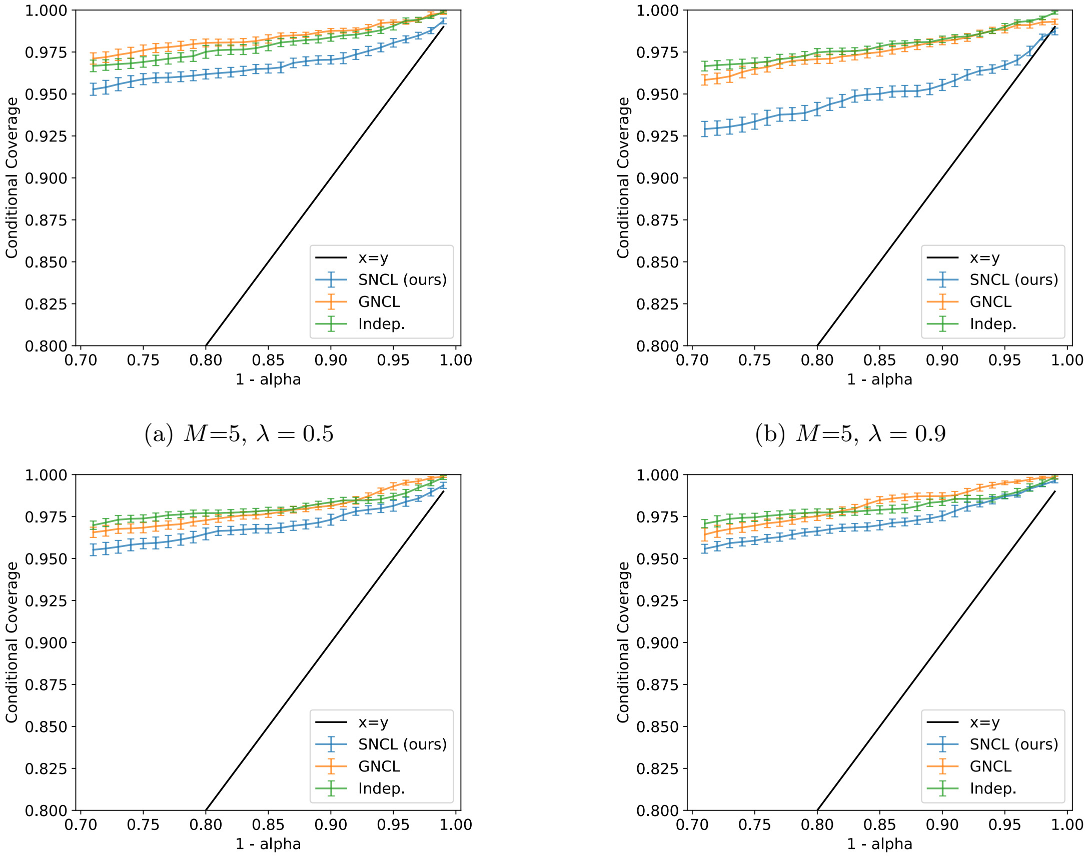
(a) $M = 5$, $\lambda = 0.5$

(b) $M = 5$, $\lambda = 0.9$

(c) $M = 20$, $\lambda = 0.5$

(d) $M = 20$, $\lambda = 0.9$

Figure B.5: Conditional coverages for test samples of class 1 in CIFAR-10



(a) M=5, $\lambda = 0.5$

(b) M=5, $\lambda = 0.9$

(c) M=20, $\lambda = 0.5$

(d) M=20, $\lambda = 0.9$

Figure B.6: Conditional coverages for test samples of class 2 in CIFAR-10

B.3.2 OOD Detection

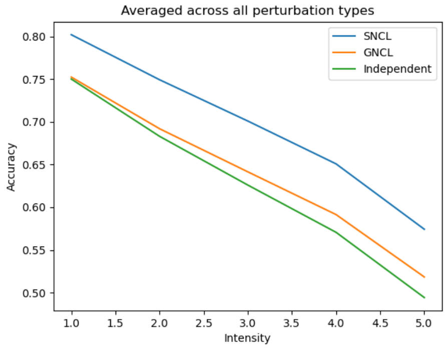
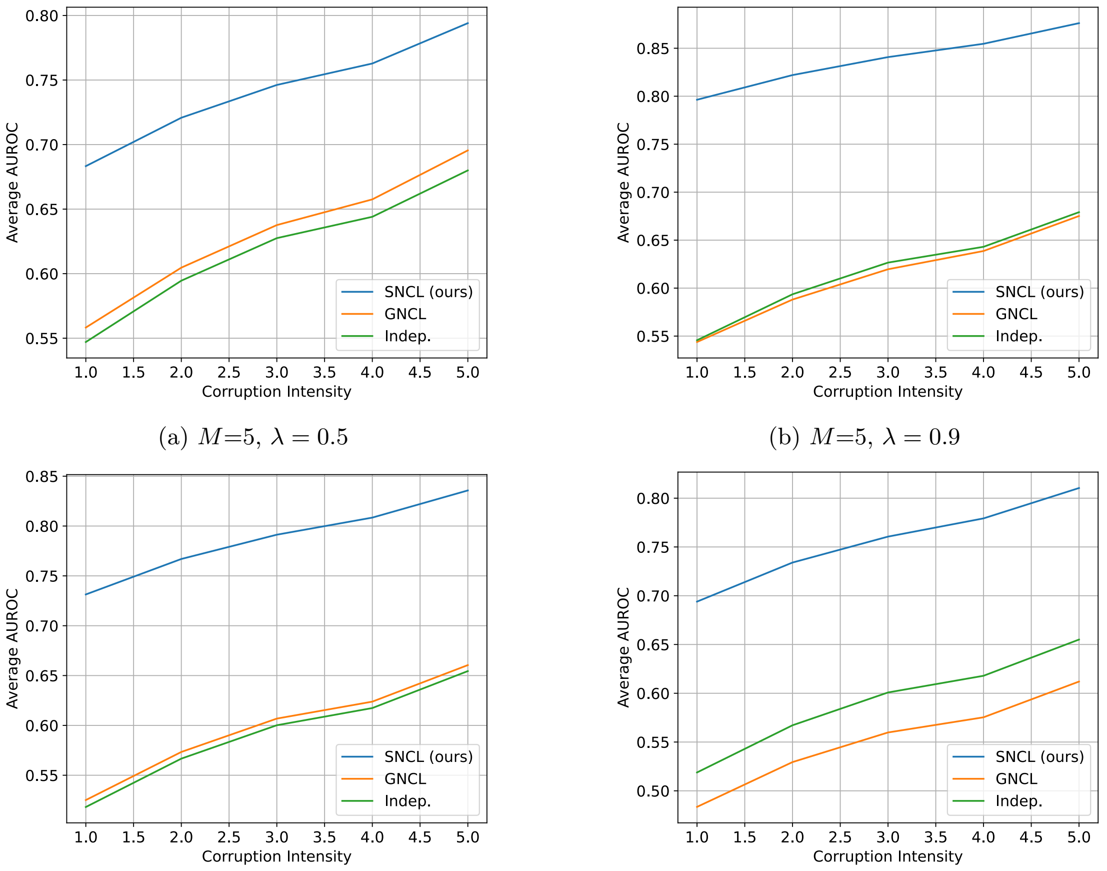


Figure B.7: Accuracy of ensembles trained on CIFAR-10 and evaluated on varying intensities of covariate shift with CIFAR-10C (i) ($M = 20$, $\lambda = 0.9$). Averaged across all 15 perturbations.



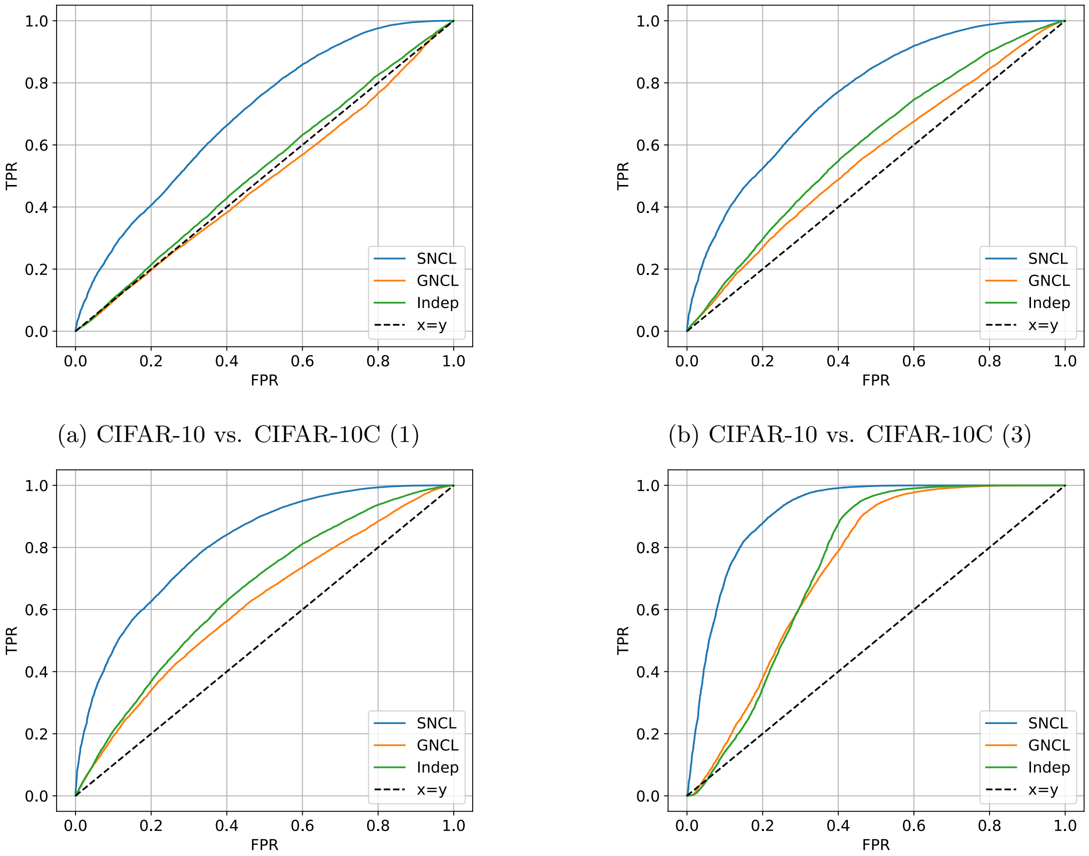
(a) M=5, $\lambda = 0.5$

(b) M=5, $\lambda = 0.9$

(c) M=20, $\lambda = 0.5$

(d) M=20, $\lambda = 0.9$

Figure B.8: AUROC of OOD detection performed with different ensemble configurations for covariate shift of varying intensities (CIFAR-10 vs CIFAR-10C (i))



(a) CIFAR-10 vs. CIFAR-10C (1)

(b) CIFAR-10 vs. CIFAR-10C (3)

(c) CIFAR-10 vs. CIFAR-10C (5)

(d) CIFAR-10 vs. SVHN

Figure B.9: ROC of OOD detection for covariate shift of varying intensities (CIFAR-10 vs. CIFAR-10C) and semantic shift (CIFAR-10 vs. SVHN) performed with ensembles ($M = 20$, $\lambda = 0.9$)