

Víctor Jiménez Rodríguez

# Improved robustness of deep learning models through posterior agreement based model selection

## Master Thesis

Institute for Machine Learning  
Swiss Federal Institute of Technology (ETH) Zurich

## Supervision

Dr. João Borges de Sá Carvalho, Dr. Alessandro Torcinovich  
Prof. Dr. Joachim M. Buhmann

September 2024

900-0030-00L



# Preface

The research presented in this thesis was conducted within the Information Science and Engineering group at the Institute for Machine Learning (ETH Zurich), under the supervision of Prof. Dr. Joachim M. Buhmann. The thesis was co-supervised at Universitat Politècnica de Catalunya by Prof. Dr. Alexandre Parera i Lluna.



# Contents

<b>Abstract</b>	<b>v</b>
<b>Notation</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 The robustness challenge . . . . .	1
1.1.1 Adversarial setting . . . . .	2
1.1.2 Out-of-distribution setting . . . . .	4
1.2 Related work . . . . .	6
1.3 Objectives . . . . .	7
<b>2 Theoretical background</b>	<b>9</b>
2.1 The learning framework . . . . .	9
2.2 Learning with neural networks . . . . .	10
2.3 Posterior agreement . . . . .	11
2.3.1 Posterior distribution . . . . .	12
2.3.2 Generalization error . . . . .	13
2.3.3 Maximum posterior agreement . . . . .	14
<b>3 Experimental setup</b>	<b>17</b>
3.1 Problem formulation . . . . .	17
3.2 Robustness in covariate shift settings . . . . .	18
3.3 Adversarial setting . . . . .	20
3.4 Domain generalization setting . . . . .	21
3.5 Robust learners . . . . .	22
3.6 Robustness assessment with posterior agreement . . . . .	23
3.6.1 Posterior in classification tasks . . . . .	23
3.6.2 The posterior agreement kernel . . . . .	24
3.6.3 Implementation . . . . .	26
<b>4 Robustness assessment</b>	<b>29</b>
4.1 In-distribution setting . . . . .	29
4.2 Adversarial setting . . . . .	33
4.2.1 Adversarial robustness assessment with PA . . . . .	36
4.2.2 Interpretability of PA in the adversarial setting . . . . .	41
4.3 Out-of-distribution setting . . . . .	44
<b>5 Model selection</b>	<b>49</b>
5.1 Model selection under controloed experimental conditions . . . . .	49
5.2 ID model selection . . . . .	52
5.3 Model selection on benchmark datasets . . . . .	55
<b>6 Conclusions</b>	<b>57</b>

<b>A Theoretical Proofs and Derivations</b>	<b>59</b>
A.1 Proof of problem formulation . . . . .	59
A.2 Properties of the PA kernel . . . . .	61
<b>B Supplementary Results</b>	<b>65</b>
B.1 PA as a robustness metric . . . . .	65
B.1.1 Empirical behaviour . . . . .	65
B.2 Adversarial setting . . . . .	67
<b>C Dataset reference</b>	<b>74</b>
C.1 Robustness assessment . . . . .	75
C.2 Model selection . . . . .	75
C.2.1 Controlled conditions . . . . .	75
C.2.2 ID model selection . . . . .	77

# Abstract

Posterior Agreement (PA) has been proposed as a theoretically-grounded alternative for model robustness assessment in covariate shift settings. In this work, we provide further evidence in favor of these findings and we explore the use of PA as a model selection criterion for deep learning models in supervised image classification tasks.

Starting from the theoretical principles leading to the formulation of PA, we derive an operative version for discrete hypothesis set problems and use it as a robustness metric in the presence of adversarially-perturbed and domain-shifted data. We show that PA outperforms standard accuracy-based metrics in both settings and possesses superior discriminative power and consistency across increasing levels of covariate shift.



# Notation

## Symbols

EHC	Conditional equation	[−]
$e$	Willans coefficient	[−]
$F, G$	Parts of the system equation	[K/s]

## Indicies

a	Ambient
air	Air

## Acronyms and Abbreviations

NEDC	New European Driving Cycle
ETH	Eidgenössische Technische Hochschule



# Chapter 1

## Introduction

This chapter aims to set the stage for the detailed analysis and discussion that will follow by providing a general overview of the problem of model robustness in machine learning and the current approaches to address it.

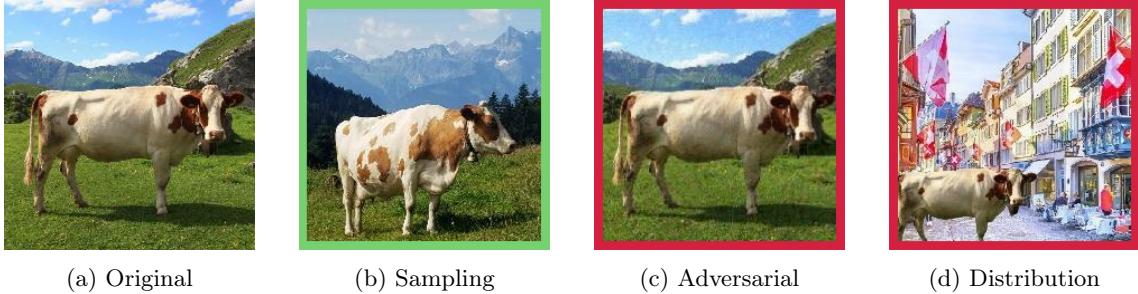
### 1.1 The robustness challenge

The field of machine learning revolves around the ambition of having computers perform certain tasks without being programmed with task-specific instructions, but instead by training them to find patterns in a data sample from which these instructions can be inferred. During training, algorithms learn from data and experience to incrementally improve their task performance. In the process, a set of assumptions are implicitly adopted about the statistical model generating the data and its connection with the desired task outcome. These assumptions constitute an inductive bias, and both architectures and training procedures are designed to learn optimal biases that generalize to unseen data samples and thus lead to robust implementations of the task at hand [45, 39]. This project will focus on the assessment of the robustness of deep learning models performing image classification tasks.

In a broad sense, robustness can be defined as the ability of a machine learning model to maintain its predictive power on observations that present some kind of transformation or variation with respect to the ones used for its training [48]. Overall, three sources of variability are relevant in the context of image classification, namely sampling randomness, adversarial perturbations and distribution shift, which are illustrated in Figure 1.1 [11].

Out of these, only sampling randomness is commonly accounted for by standard model validation techniques, in the sense that model selection and benchmarking are conducted using randomized subsets of unseen observations. In this way, the most generalizable features, and in turn the most generalizable models, are naturally selected. As it will be outlined in this chapter, this approach presents fundamental limitations that are rooted in the very nature of deep learning models and the data from which they learn.

First, the operative principles of neural networks make them vulnerable to small perturbations in the input space, which are often filtered out in human perception, that can lead to high-confidence incorrect predictions [55]. This issue is commonly known as adversarial vulnerability, and an ongoing arms race incentivizes the design of new ways of perturbing models and new ways of defending them against such attacks. Strategies that foster robustness to adversarial attacks are possible, but come at a price of hindering conventional generalization to sampling randomness in the original data [57].



(a) Original

(b) Sampling

(c) Adversarial

(d) Distribution

Figure 1.1: Illustrative example of the three expected sources of variability. A pre-trained MobileNetV2 model is shown to be vulnerable to adversarial perturbations as the one represented in (c), and also to distribution shifts as the one illustrated in (d), possibly because its inductive bias is influenced by the spurious correlation between cows and rural landscapes.

Second, the nature of the data used for training and selecting models is known to influence heavily the features that the model will learn to be the most predictive. The lack of representativity of certain aspects of the data or the presence of spurious correlations can lead to models that generalize well to sampling randomness within the same experiment but that fail to do so when those accidental relationships are not present. This is known as an out-of-distribution setting, given that samples in which the model is tested are not drawn from the same probability distribution that generated training samples [48].

At the core of the robustness challenge lies the poor understanding of how models construct their inductive bias and the nature of the transformations between the space of weights and the space of functions that they are able to represent [29]. Features learned by the optimal standard classifier can be completely different from those learned by a robust classifier, regardless of the amount of data provided, which results in a fundamental limitation for task performance in robust models [57, 71]. Besides, the feature space that deep learning models navigate is fundamentally different than that in which humans implicitly rely on, and we should therefore not expect models to be invariant to the same features humans are instinctively invariant to [27].

This thesis will encompass all these phenomena under the same theoretical framework, and devise a common approach to the measurement of the shift entailed by both adversarial and distribution variability sources. Robustness will be characterized from the space of outcomes of the model, by means of a (posterior) probability distribution that will rank models and algorithms according to the agreement in their predictions when facing different realizations of the same experiment.

### 1.1.1 Adversarial setting

As it was previously mentioned, certain perturbations on original test images, which can be almost imperceptible to the human eye, can lead to highly-confident but incorrect predictions by deep neural networks, even when their standard performance metrics are high. Adversarial examples have been shown to transfer across architectures and training procedures, and even across subsets of data, often yielding the same incorrect prediction in all of these cases [55].

These intriguing phenomena were initially hypothesized to arise from a lack of smoothness over the input space, a property commonly assumed in other learners, that derives from the non-linear nature of deep learning architectures. Nevertheless, extensive research on the field has elucidated that the root cause is instead the linearity of its learning units, which makes them vulnerable in certain directions of high-dimensional spaces where small effects can add up to significantly change the outcome [21].

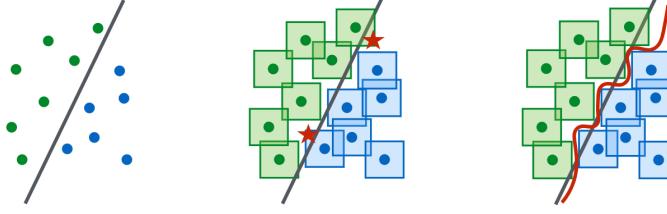


Figure 1.2: A conceptual illustration of standard vs. adversarial decision boundaries. (**left**) A set of linearly-separable points. (**middle**) Decision boundary learned via standard training. (**right**) Decision boundary learned via adversarial training. Both methods achieve zero training error, but only the robust model is able to generalize to  $\ell_\infty$  perturbations. [41]

Building on this intuition, numerous attacks have been proposed to evaluate the robustness of models against adversarial samples. A common strategy for inducing model failure involves identifying vulnerable directions in the feature space and adjusting perturbations to produce the desired misleading effect. Adversarial examples generated through these attacks can then be used to train robust models through regularization, thereby promoting generalization to the features present in worst-case examples and thus selecting models that are insensitivized to them [5].

Nevertheless, adversarial learning entails decision boundaries that are more complex than the ones derived via standard training (see Figure 1.2), intuitively demanding more data and more complex architectures, at the risk of overfitting to adversarial examples themselves [52]. These limitations express a fundamental trade-off that arises from the intrinsic disparity between robust and non-robust features [57, 71].

Features selected via standard training are the most predictive towards generalizing to sampling randomness within the same dataset, but they do not necessarily represent the features implicitly selected by humans and are not invariant to a human-based notion of similarity. Instead, features selected via adversarial training have been shown to better model this invariance, and thus align much better with human perception, as seen in Figure 1.3 [27]. Furthermore, adversarial perturbations of robust models display salient characteristics; that is, features that are perceived to belong to the class they are misclassified to, as illustrated in Figure 1.4 [57].

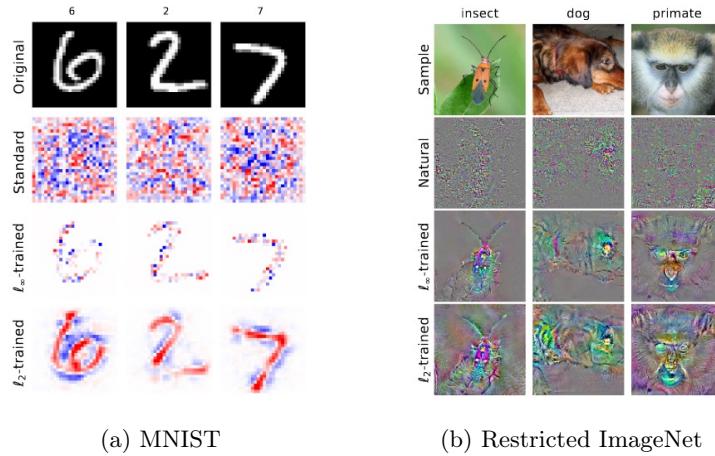


Figure 1.3: Scaled loss gradient with respect to input images. Input pixels yielding the highest predictive power are aligned with perceptually relevant features for the case of adversarial models, while appearing completely random in the case of standard models. [57]

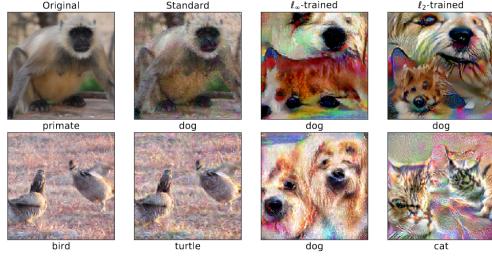


Figure 1.4: Adversarial examples for standard and adversarially-trained models. Perturbed images produced for robust models effectively capture salient data characteristics. [57]

Overall, these and other findings suggest that robustness in the adversarial setting is a fundamental property of the data features that are represented by models, rather than of the models themselves, and the phenomenon of transferability can be explained in these terms. Training strategies that manage to navigate the robustness-generalization trade-off will be the ones yielding the best results, provided that the data distribution is representative of the true underlying features.

### 1.1.2 Out-of-distribution setting

Most learning algorithms work under the fundamental assumption that a causal relationship exists between input and output spaces. The target function to learn represents that causality and must therefore remain invariant regardless of the available data, which implies that suitable approximations of this function can be obtained as long as data samples are independent and identically distributed in the input space [43, 48]. Nevertheless, this is not always the case, as often real-world data does not match the same statistical patterns of the data used for training. Ultimately, this phenomenon induces a distribution shift that leads to poor generalization performance [73, 60, 38].

	Train			Val (OOD)	Test (OOD)
	$d = \text{Hospital 1}$	$d = \text{Hospital 2}$	$d = \text{Hospital 3}$	$d = \text{Hospital 4}$	$d = \text{Hospital 5}$
$y = \text{Normal}$					
$y = \text{Tumor}$					

Figure 1.5: The `camelyon17` (WILDS) dataset comprises images of stained lymph node tissue patches sampled from different hospitals. [32]

A distribution shift can arise for various reasons, namely the unfeasibility of collecting diverse enough data, the changing or time-dependent nature of the data, or the implicit bias introduced during the data collection process. This last case is particularly relevant, as it can serve as a generalization of all the previous cases and raise epistemological questions about the learning framework itself. For instance, Figure 1.6 refers to a cross-generalization analysis in which popular machine learning datasets were shown to be biased towards specific representation of features. Considering the fact that all data is sampled from the same source (i.e. Internet), numerous human-induced biases are shown to determine the nature of representations, the most significant of all being negative bias, which arises when the negative subset<sup>1</sup> of the dataset is not representative of the input subspace excluding that particular class and results in a model that performs significantly worse in other datasets, even when trained with the same observations of that class.

<sup>1</sup>When certain observations in a dataset are labelled as belonging to a specific class, the remaining observations are implicitly assigned to not belong to that class, and therefore define a negative set in the model's feature space.

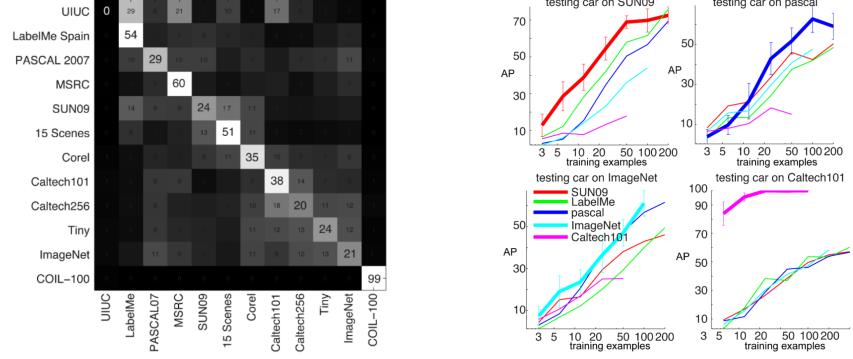


Figure 1.6: (**left**) Confusion matrix generated in a dataset identification task. A clearly pronounced diagonal indicates that each dataset possesses unique traits that make it distinguishable from the rest. (**right**) Cross-dataset generalization for **car** detection as function of training data. The vertical gap between lines represents the decrease in performance when training on a different dataset, and the horizontal shift corresponds to the increase in the amount of data needed to reach the same performance. [56]

Several approaches can be taken to address this issue, depending on the nature of the shift and the accessibility of the causal structure of the data (see Figure 3 and Table 2 in [60]). Nevertheless, the common goal is to push the model towards domain-invariant representations that foster robustness in the face of distribution shifts, sometimes relaxing the causality condition to an assumption of invariance or stability of the distribution in the output space [60, 38].

In general, every formulation considers a set of source domains encompassing data that is available for the training of the model, including any validation subsets used for model selection, regularization or hyperparameter tuning, and a set of target domains encompassing unseen data on which model performance will be evaluated. Within this framework, a straightforward approach to improving robustness is to directly sample target domains and adjust feature representations to be invariant between both, which is commonly known as domain adaptation.

In this work we will focus instead on domain generalization, which refers to the case in which target domains are not accessible and feature invariance can be only enforced from source domains [6]. In particular, two strategies will be considered, namely domain alignment and data augmentation/generalization.

On the one hand, domain alignment stems from the target invariance hypothesis, and can be formulated as a regularization problem that pushes towards the minimization of the dissimilarity of feature representations originated from different source environments. The feature space in which the alignment is performed (e.g. a kernel latent space, as Figure 1.7 illustrates) and the similarity metric considered will determine the peculiarity of the method [53, 37].

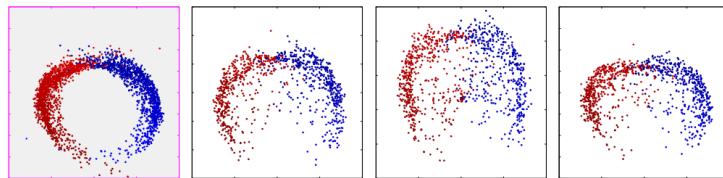


Figure 1.7: Projections of a binary synthetic dataset in the two principal DICA dimensions. The shaded box depicts the projection of training data, whereas the unshaded boxes show projections of unseen test datasets. [43]



(a) Original

(b) Mixup [72]

(c) CutMix [69]

Figure 1.8: Mixup and Cutmix strategies can be used to interpolate between different labels and/or domains by generating intermediate observations. [69]

On the other hand, data augmentation/generation strategies do not need to assume target invariance and instead achieve cross-domain generalization by generating artificial observations that diversify the original dataset with the hope of capturing the underlying causal structure of the data generation process. Augmented observations can either be randomizations of original observations (e.g. transformations such as rescaling or rotations) or new samples filling the distribution gaps between domains (e.g. through interpolation, as illustrated in Figure 1.8).

Unlike in the adversarial setting, there is no common way of measuring the shift in distribution between source domains, and current approaches are often constrained to specific datasets or training strategies. Robustness is instead quantified during (cross-)validation, either by reserving a subset of each domain, leaving one domain out or directly accessing target domains if they are available, which is known as the oracle approach. This last strategy is often used to provide an upper bound estimate of model robustness, as it usually provides over-confident performance estimates [73]. Numerous benchmark datasets, some of which will be considered in this work, are the current standard for robustness assessment even with the limitations they present [32].

## 1.2 Related work

In the adversarial front, early work [55] unveiled the nature of the susceptibility of deep learning models to adversarial examples and FGSM [21] was introduced as an intuitive approach for robust regularization. Since then, several gradient-based methods have been proven to enhance adversarial robustness, such as PGD [41], C&W [13], FMN [47] and many others (see [36] for reference). All of them ultimately entail a strategy to find a vulnerable direction and adjust the perturbation (e.g. minimum-norm, maximum-confidence, etc.) based on the location of the decision boundary, either via soft constraints (i.e. regularization), boundary attacks or gradient projections [5].

In general, the primary distinction among adversarial attacks lies in their knowledge of the model’s architecture and parameters. In that sense, white-box and black-box attacks can be distinguished, where the former have full access to the model and the latter only to the model’s predictions. In black-box settings, the loss gradient is unknown and other strategies such as score-based or decision-based attacks are used [36]. Regarding adversarial training (i.e. defenses), robustness can be achieved by a variety of methods, such as ensemble learning, defensive distillation, generative adversarial networks [66, 42], diffusion models [62, 26] and adaptive-boundary methods [16]. In this project, the Robustbench attack library [17] will be leveraged to evaluate adversarial robustness in the CIFAR10 dataset [33].

In the domain generalization front, the existing rich taxonomy of methods can be classified into two main groups, namely data manipulation and representation learning [60, 73, 38]. First, data manipulation strategies involve both augmentation and generation, as for example image randomization or adversarial augmentation [67, 72, 69]. Second, representation learning strategies are

primarily divided into domain-invariant methods (e.g. model-based [2], kernel-based [43, 3] or adversarial-based [46]) and feature disentanglement methods, which encompass causality-inspired approaches and general multi-component analysis. Other learning strategies include meta-learning, ensemble learning or self-supervised learning [35, 59].

Regarding robustness characterization, a wide range of metrics have been conceived (see [23]), but accuracy-based criteria are still the most common. As an alternative, some theoretically-grounded approaches have been proposed, such as CLEVER [64], ACTS [61] or PA [11], which is the one explored in this work. In general, robustness is often reported and compared using robustness benchmark datasets. Some of the most relevant for image classification tasks are MNIST [34] (and its multiple variations, such as DiagVib-6 [20]), PACS [68], VLCS [30] or WILDS [32].

### 1.3 Objectives

The main goal of this project is to assess the suitability of the posterior agreement framework in the context of deep learning model robustness for image classification tasks. For that, an operative version of posterior agreement for finite, discrete hypothesis classes must be derived and efficiently implemented, so that it can be used as a metric to evaluate and select models based on the robustness of their response to different sources and levels of randomness. The results obtained should be compared with the current state-of-the-art in robustness evaluation, namely Robustbench [17] and WILDS [32] benchmarks in the adversarial and out-of-distribution settings, respectively, and an overall analysis of the use of the metric as an early-stopping criterion should be provided.

In order to conduct a comprehensive assessment, a series of steps should be undertaken in a deductive manner, so that evidence is provided starting from first principles and culminating with performance evaluations on benchmark datasets. To begin with, the content of this work should be self-contained, and for that a formal statement of the learning problem should be given. Both learning algorithms and classification models should be defined, including the nuances of their parametrization via deep neural networks.

Next, the robustness challenge should be formulated within the framework of probability theory, providing a rigorous mathematical foundation for understanding the principal sources of randomness that are relevant in the context of this work. The generalization capabilities to specific sources of randomness will determine the robustness score of a model, which for PA is conceived as a measure of the stability of the probabilistic output of the model to the randomness of the data sampling process. In this sense, the fundamental generalization-complexity trade-off must be reframed within the context of information theory by relating complexity to the expected information content of the data. The estimated informativeness will determine the resolution of the hypothesis space and thus its stability under perturbations of various kinds. Once an operative PA metric is derived, its properties should be investigated and further assessed with artificial data, so that PA can be compared with baseline accuracy-based metrics at a fundamental level.

Experiments should be performed the adversarial setting first, for being adversarial perturbations intentionally designed to change the prediction of the model and thus entailing a performance-based notion of robustness that aligns with accuracy measures. After that, a customized implementation of the DiagVib-6 [20] synthetic data generation pipeline should be used to evaluate the suitability of PA in the domain generalization setting. This will allow for a detailed analysis of the model selection capabilities of PA under different experimental conditions, especially regarding the source, power and rate of the shift entailed by each dataset and the accessibility of target domains for model validation and selection.

All in all, both analytical and numerical tools should be used to provide a comprehensive analysis of the model-selection capabilities of PA and the source of its discriminative power.



# Chapter 2

## Theoretical background

### 2.1 The learning framework

Statistical learning theory encompasses the mathematical framework used to study generalization in machine learning [44]. In this formalism, the goal is to learn a target function  $f^* : \mathcal{X} \mapsto \mathcal{Y}$  by means of an approximated function  $f \in \mathcal{F}$  using a finite set of observations.

**Definition** (*Supervised dataset*). Let  $\mathcal{X}$  and  $\mathcal{Y}$  be the input and output spaces of the target function  $f^*$ , respectively. Let  $X$  be a random variable associated with a sampling experiment in the input space, thus defining a measure of probability  $P_X$  with support  $\mathcal{X}$ . Let  $\mathbf{X} = (X_1, \dots, X_N)$  be a (simple) random sample of  $X$  with size  $N$  [14]. A supervised dataset  $D$  is constructed from a realization  $\mathbf{x} \sim \mathbf{X}$  by pairing each observation  $x$  with its corresponding value  $f^*(x)$  under the target function:

$$D = \{(x_n, f^*(x_n))\}_{n \in [N]}.$$

The class of supervised datasets generated from  $\mathbf{X}$  will be denoted by  $\mathcal{D}$ .

**Definition** (*Empirical risk*). Let  $D$  be a supervised dataset generated from a sample  $\mathbf{x} \sim \mathbf{X}$ . Let  $L : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$  be a loss function measuring the discrepancy at observation  $x \in \mathbf{x}$  between the prediction of the model  $f(x)$  and its corresponding true value  $f^*(x)$ . The overall quality of the approximation  $f$  can be measured with its expected risk  $R(f)$ :

$$R(f) = \mathbb{E}_X L(f(x), f^*(x)).$$

We can approximate the expected risk with its empirical (plug-in) analogous if  $N$  is large enough (see Glivenko-Cantelli theorem [24]). The empirical risk of  $f \in \mathcal{F}$  computed on  $D$  is defined as

$$\hat{R}(f) = \frac{1}{N} \sum_{n=1}^N L(f(x_n), f^*(x_n)).$$

Therefore, learning amounts to minimizing  $\hat{R}$  over the function class  $\mathcal{F}$ . The empirical risk minimization problem (ERM) is thus defined as

$$f_{\text{ERM}} = \min_{f \in \mathcal{F}} \hat{R}(f).$$

Eventually, the selected model  $f$  will be that achieving the lowest generalization error, which is typically measured by evaluating the task performance of the model on a different realization of the sampling experiment  $\mathbf{x}' \sim \mathbf{X}$ . For that, a validation dataset  $D' \in \mathcal{D}$  is usually employed, or, in its absence, a cross-validation strategy that iteratively reserves different (disjoint) subsets of the training dataset  $D$  for model selection purposes.

It can be shown that the generalization error is ultimately linked to the complexity of the function class. The definition of complexity depends on the nature of the problem, but it is intuitively related to the cardinality of the subset of  $\mathcal{F}$  that the learning algorithm navigates. A complex or high-capacity algorithm will be able to represent a larger subset of  $\mathcal{F}$  and thus achieve a low empirical error, but will be also prone to overfitting to the specific sampling realization considered and yield a higher generalization error [44].

As a general principle, the inductive bias of the algorithm, which translates to a set of constraints imposed on the  $\mathcal{F}$  during learning, should be aligned with that of our target function [29]. In order to avoid overly expressive function classes to be selected during the loss minimization process, a common regularization strategy consists of including an additional term on the ERM objective that penalizes the complexity of the model under consideration.

**Definition** (*Regularized empirical risk*). Let  $\Omega : \mathcal{F} \mapsto \mathbb{R}$  be a functional quantifying the complexity of the elements of the function class. The regularized empirical risk of  $f \in \mathcal{F}$  is

$$\hat{R}_\Omega(f) = \hat{R}(f) + \lambda \Omega(f),$$

where  $\lambda \in \mathbb{R}^+$  expresses the trade-off between empirical risk and generalization error.

## 2.2 Learning with neural networks

Neural networks are biologically-inspired machine learning models that consist of a set of nodes (neurons) organized in layers and connected by weighted edges (synapses). Figure 2.1 illustrates the transformation performed within a single node [54, 44, 58].

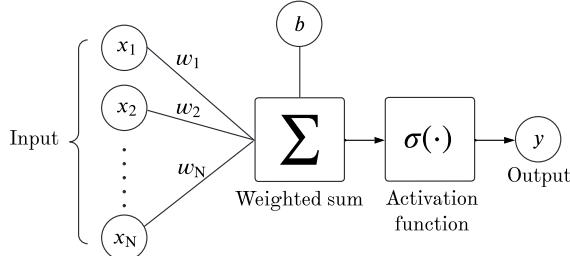


Figure 2.1: The output of a node is computed by applying a non-linear activation function  $\sigma$  to the weighted sum of its inputs  $x$  plus a bias term  $b$ .

**Definition** (*Neural network parametrization*). Let  $\mathbf{x}_k \in \mathbb{R}^{d_k}$  be the input to the layer  $k \leq M$ , and let  $\mathbf{W} \in \mathbb{R}^{d_k \times d_{k+1}}$  be the  $k$ -th weight matrix. The output of the layer can be expressed as

$$\mathbf{x}_{k+1} = \sigma_k(\mathbf{z}_{k+1}) = \sigma_k(\mathbf{W}_k^T \mathbf{x}_k + \mathbf{b}_k),$$

where  $\sigma_k$  is the non-linear activation function at layer  $k$ . We can therefore express the overall transformation of a neural network as the composition of its layers:

$$f_{\text{NN}}(\mathbf{x}_0; \gamma) = \mathbf{x}_{M+1} = \bigcirc_{k=0}^M \sigma_k(\mathbf{W}_k^T \mathbf{x}_k + \mathbf{b}_k),$$

where  $\gamma \in \Gamma \subset \mathbb{R}^{|\Gamma|}$  represents the set of parameters of the network. Therefore, the function class  $\mathcal{F}$  can be parametrized as  $\mathcal{F}_\Gamma$  through neural network architectures:

$$\begin{aligned} \text{NN} : \Gamma &\subseteq \mathbb{R}^{|\Gamma|} \mapsto \mathcal{F}_\Gamma \\ \gamma &\mapsto f_{\text{NN}}(\mathbf{x}; \gamma), \end{aligned}$$

where  $\Gamma$  is the parameter space navigated by this particular architecture. The function class  $\mathcal{F}_\Gamma$  consists of all mappings  $f(\cdot; \gamma) : \mathcal{X} \mapsto \mathcal{Y}$  that can be realized by some configuration  $\gamma \in \Gamma$ .

In order to solve the learning problem, the optimization algorithm must navigate the non-convex loss landscape towards the minimum of the empirical risk. This is computationally achieved by means of gradient-descent-based optimizers, which efficiently compute the gradient over the parameters via through backpropagation algorithm [50]. In practice, more efficient variations of gradient descent are used, such as stochastic gradient descent (SGD) [49] or Adam [31].

The universal approximation theorem states that arbitrarily wide architectures are able to represent virtually any function, but it is an open challenge to theoretically describe which complexity measure regulates generalization. A possible approach to this problem is to study the geometry of the loss landscape, especially in the vicinity of local minima. For instance, connected flat minima are often associated to better generalization capabilities, as they intuitively represent a robust region in the parametrization space and should be preferred over sharp minima [29].

In this work we will explore a different approach to the generalization problem, rooted on a measure of generalization error that accounts for the implicit randomness of the data generation process.

### 2.3 Posterior agreement

As mentioned at the start of the chapter, the input of learning algorithms are datasets containing observations of a random variable  $X$  with support  $\mathcal{X}$ . The implicit randomness embedded in the sampling process extends to the learning outcome of algorithms, even when performing a deterministic set of operations. An alternative intuition of generalization arises from this perspective, in the sense that a good algorithm should be expected to learn the same function when trained on different realizations of the same experiment (e.g.  $\mathbf{x}', \mathbf{x}'' \sim \mathbf{X}$ ). Each resulting dataset is drawn from the same probability distribution over the support, but entails a different instantiation of the randomness associated with the sampling process [9].

A regularization principle is derived from this intuition and can be formalized as a generalization-complexity trade-off by defining generalization as the robustness or stability of the learned function to sampling randomness. A suitable measure of complexity in this framework is the informativeness of the function, which represents its ability to learn the patterns in the data while filtering out the noise. The more expressive (i.e. complex) a function class is, the higher will be the estimated information content of the data. If the information content is underestimated, the approximated function will lack the capacity to learn some patterns in the data, whereas if informativeness is overestimated, it will overfit to the noise and thus not generalize to different realizations of the experiment [15, 12, 10].

The robustness-informativeness regularization principle can be enforced from the set of outputs of the learned model, when both the distribution of the data over the support and the sampling randomness associated to its measurement are accounted for. This section will formalize this principle and derive an expression for the minimization of the generalization error.

**Definition (Data distribution).** The (simple) random sample  $\mathbf{X} \stackrel{\text{iid}}{\sim} X$  entails a measure of probability described by the density function  $f_{\mathbf{X}}$ :

$$f_{\mathbf{X}} = \prod_{n=1}^N f_X,$$

where  $f_X$  is the density function of  $X$ . We will use  $\mathbf{P}_X$  to refer to the empirical approximation of this distribution; that is, to the distribution of samples in a dataset  $D$  drawn from  $\mathbf{X}$ :

$$\mathbf{P}_X = \frac{1}{N} \sum_{n=1}^N \delta(x - x_n).$$

**Definition (Sample).** Let  $X$  be a random variable associated with a measure of probability in  $\mathcal{X}$ . Let  $\tau \in \mathcal{T}$  be an instantiation of the randomness allowed by such measurement. The class of transformations  $\mathcal{T}$  is composed of the possible experimental conditions for the data sampling process. The dependency of measurement realizations  $\mathbf{x} \sim \mathbf{X}$  on experimental conditions will be captured by index  $\tau$ , and we will implicitly consider sample

$$\mathbf{x} := \tau \circ \mathbf{x}$$

to be a realization of the experiment  $\mathbf{X} \stackrel{\text{iid}}{\sim} X$  under conditions  $\tau$ .

The nature of  $\tau$  is subject to different interpretations depending on the experiment under consideration. In the context of this work, sampling is limited to a finite set of labeled images that have been pre-selected for the classification task. For instance, the class of possible transformations  $\mathcal{T}$  defining samples in MNIST [34] should a priori not include the full range of variations that each digit can encompass, since the data collection experiment has already been performed. Instead, it would contain the set of possible subsets of  $N$  images that can be drawn from the MNIST repository, as this reflects the true randomness to which the model is pushed to be invariant.

Nevertheless, the goal of the learning task is not to generalize within MNIST, but rather to distill those features in the data that allow the model to generalize to other similar digit observations that could be collected in the future. For that reason, we will interpret  $\tau$  as an instance of the randomness entailed by the data collection process, which is defined over the (infinite) support containing all possible images that can be generated under certain experimental conditions. Then, the true data distribution encapsulates the sampling probability of all image features that can be possibly generated, and as a consequence the information content of these features for the task at hand.

A dataset generated from  $\mathbf{x} \sim \mathbf{X}$  entails an instantiation  $\tau \in \mathcal{T}$  that remains unobserved, since the statistical model governing the data generation process is inaccessible. This implies that the information content of the features present in the sample must be approximated. The suitability of this approximation will ultimately determine the generalization capabilities of the model.

### 2.3.1 Posterior distribution

**Definition (Hypothesis class).** Let  $\mathcal{D}$  be the class of supervised datasets generated from  $\mathbf{X}$ . A data science algorithm learns a function  $f$  implementing the following mapping:

$$\begin{aligned} f : \mathcal{D} &\longmapsto \Theta \\ \mathbf{x} &\longmapsto (f(x_1), \dots, f(x_N)) = \theta. \end{aligned}$$

The hypothesis class  $\Theta$  is the output space of hypothesis representing all possible outcomes of a function  $f$  learned on a dataset sampled from  $\mathbf{X}$  [9].

Intuitively, this framework interprets complexity from the perspective of the set of possible outcomes of the function, rather than the function class itself. It can be argued that both perspectives are equivalent, in the sense that any function class can be ultimately mapped to a specific hypothesis space  $\Theta$ . Nevertheless, the underlying transformation is not homeomorphic in general, and more suitable generalization regularization constraints can be defined in  $\Theta$ , especially when dealing with intractable function classes  $\mathcal{F}_\Gamma$  represented by deep neural networks.

For instance, complexity in the hypothesis class can be associated to the nature of the randomness displayed by  $X$ . Ideally, too restrictive hypothesis classes that lack desirable hypothesis for some realization  $\mathbf{x} \sim \mathbf{X}$  should be avoided, and also those hypothesis classes containing unrealizable elements (i.e. hypothesis that are not outcome of any possible realization of the experiment). A richness condition for the construction of  $\Theta$  can thus be postulated following this intuition.

**Proposition (Richness condition).**  $\Theta$  should stem from a sufficiently rich set of experimental conditions  $\mathcal{T}$  such that every hypothesis  $\theta \in \Theta$  is the outcome of some realization  $\mathbf{x} \sim \mathbf{X}$ .

$$\forall \theta, \exists \tau \in \mathcal{T} \text{ such that } f(\mathbf{x}) = \theta$$

Definition 2.3 states that each datasets entails an instantiation of the sampling experiment, which can be formalized as an implicit index  $\tau \in \mathcal{T}$  associated with a specific realization of the data generation process. The cardinality of  $\mathcal{T}$  defines an upper bound on the resolution of the hypothesis class, in the sense that no more than  $|\mathcal{T}|$  different hypothesis will be navigated by a learner, regardless of the task at hand.

The richness condition formalizes this intuition by constructing  $\Theta$  as a surjective transformation of  $\mathcal{T}$ . As a consequence, the output of the model can be effectively interpreted as a random variable  $\theta$  associated with a measure of probability  $\mathbf{P}^f$  over  $\Theta$ .

**Definition (Posterior).** Let  $\mathfrak{P}^f$  be a family of distributions under consideration. A probability distribution over the hypothesis class can be defined as a conditional distribution given an realization  $\mathbf{x} \sim \mathbf{X}$ . We will refer to this distribution as the posterior over  $\Theta$  under  $f$ :

$$\begin{aligned} \mathbf{P}^f : \mathcal{D} \times \Theta &\longmapsto \mathbb{R} \\ (\mathbf{x}, \theta) &\longmapsto \mathbf{P}^f(\theta | \mathbf{x}). \end{aligned}$$

The posterior  $\mathbf{P}^f \in \mathfrak{P}^f$  establishes the stochastic relation between data realizations and hypotheses.

Using these definitions we can operate over  $\Theta$  within the framework of probability theory. For instance, we can obtain the (prior) probability of a hypothesis to be selected by  $f$  as

$$\Pi^f(\theta) = \mathbb{E}_{\mathcal{T}} \mathbf{P}^f(\theta | \tau) = \mathbb{E}_{\mathbf{X}} \mathbf{P}^f(\theta | \mathbf{x}),$$

from which we can derive a probabilistic version of the richness condition, where a limit case can be imposed with exactly one experiment per hypothesis, leading to a uniform prior

$$\Pi^f(\theta) = |\Theta|^{-1}$$

when the hypothesis class is finite. Within this framework, selecting suitable hypothesis classes amounts to selecting posterior distributions that yield a higher probability to the desired subset of hypothesis. This is the leading principle that will guide the derivations that follow.

### 2.3.2 Generalization error

In order to postulate an informativeness-based definition of generalization error, we will consider two samples  $\mathbf{x}', \mathbf{x}'' \sim \mathbf{X}$  generated from the same experiment. Samples are independent and the only differ in the implicit randomness entailed by the sampling experiment:

$$\mathbf{P}_{\mathbf{x}', \mathbf{x}''} = \mathbf{P}_{\mathbf{x}'} \mathbf{P}_{\mathbf{x}''}.$$

**Proposition (Posterior selection principles).** Let  $\mathcal{T}_{\mathbf{X}} \subseteq \mathcal{T}$  be the subset of sampling realizations that is effectively navigated by  $\mathbf{X}$ , in the sense that it encompasses only those instantiations that can be sampled from  $\mathbf{X}$ . Let  $\Theta_{\mathbf{X}} \subseteq \Theta$  be the realizable subset<sup>1</sup> of hypotheses, generated as a surjective transformation of  $\mathcal{T}_{\mathbf{X}}$ . Following the intuition introduced at the beginning of the chapter, two posterior selection principles are derived from the robustness-informativeness trade-off:

- Posteriors  $\mathbf{P}^f$  should be expressive enough to cover the realizable subset  $\Theta_{\mathbf{X}}$ .
- Equally likely inputs drawn from the same experiment should yield similar sets of hypothesis.

---

<sup>1</sup>In the context of supervised classification tasks,  $\mathcal{T}_{\mathbf{X}} \subseteq \mathcal{T}$  and therefore  $\Theta = \Theta_{\mathbf{X}}$ , since the number of classes is fixed beforehand and all possible combinations of labels are realizable a priori.

**Definition (Description length).** Let  $\mathcal{F}_\Gamma(\cdot)$  be the function class encompassing all functions represented by a parametrization  $\Gamma$ . Let  $\mathbf{P}_\Gamma$  be the universal distribution relative to  $\mathcal{F}_\Gamma$  fulfilling the minimum description length principle. The description length of a function  $f_\gamma \in \mathcal{F}_\Gamma$  is defined as the number of bits required to encode its parameters [22]. The code length of the argument of such distribution is

$$\text{DL}_{f_\gamma}(\cdot) = -\log f_\gamma(\cdot).$$

The quality of the represented function  $f$  will be measured by the description length of its posterior [9], and thus a loss function can be defined as follows:

$$\ell(\theta, \mathbf{x}) = -\log \mathbf{P}^f(\theta | \mathbf{x}).$$

Given that description length also accounts for the complexity of the hypothesis class and not only its generalization capabilities, loss values are normalized by dividing by the description length of the prior:

$$-\log \Pi^f(\theta) = -\log \mathbb{E}_{\mathbf{x}} \mathbf{P}^f(\theta | \mathbf{x}).$$

**Definition (Generalization error).** Let  $\mathbf{x}'$  and  $\mathbf{x}''$  be realizations of  $\mathbf{X}$ . Let  $\Theta$  be the (realizable) hypothesis class represented by  $f$  given  $\mathbf{X}$ . The generalization error is defined as the out-of-sample description length:

$$\mathcal{G}_{\mathcal{X}} = \mathbb{E}_{\mathbf{P}^f(\mathbf{x}', \mathbf{x}'')} \mathbb{E}_{\mathbf{P}^f(\theta | \mathbf{x}')} \left[ -\log \frac{\mathbf{P}^f(\theta | \mathbf{x}'')}{\Pi^f(\theta)} \right].$$

It amounts to the expected loss over the normalized posteriors on (validation) sample  $\mathbf{x}''$  weighted over the posterior distribution on (training) sample  $\mathbf{x}'$ . Intuitively, a lower generalization error is achieved when good quality hypothesis on  $\mathbf{x}''$  are likely to be drawn from  $\mathbf{x}'$ .

**Lemma 2.3.1 (Posterior agreement).** The generalization error  $\mathcal{G}_{\mathcal{X}}$  is non-negative and has a lower bound  $-\mathcal{J}$ . We define  $\mathcal{J}$  as the posterior agreement.

*Proof.*

$$\begin{aligned} \mathcal{G}_{\mathcal{X}} &\geq \mathbb{E}_{\mathbf{P}^f(\mathbf{x}', \mathbf{x}'')} \left[ -\log \left( \mathbb{E}_{\mathbf{P}^f(\theta | \mathbf{x}')} \frac{\mathbf{P}^f(\theta | \mathbf{x}'')}{\Pi^f(\theta)} \right) \right] \\ &= \boxed{\mathbb{E}_{\mathbf{P}^f(\mathbf{x}', \mathbf{x}'')} \left[ -\log \left( \sum_{\theta \in \Theta} \frac{\mathbf{P}^f(\theta | \mathbf{x}') \mathbf{P}^f(\theta | \mathbf{x}'')}{\Pi^f(\theta)} \right) \right]} = -\mathcal{J} \\ &\geq -\log \left( \mathbb{E}_{\mathbf{P}^f(\mathbf{x}', \mathbf{x}'')} \mathbb{E}_{\mathbf{P}^f(\theta | \mathbf{x}')} \mathbb{E}_{\mathbf{P}^f(\theta | \mathbf{x}'')} \frac{\mathbf{P}^f(\theta | \mathbf{x}'')}{\Pi^f(\theta)} \right) = 0, \end{aligned}$$

where Jensen's inequality has been applied twice to the convex function  $-\log(\cdot)$ .

□

### 2.3.3 Maximum posterior agreement

The maximum posterior agreement criterion, which follows from Lemma 2.3.1, can be formalized as an optimization problem over the function class.

**Definition (Kullback-Leibler divergence).** Let  $\mathbf{P}$  and  $\mathbf{Q}$  be two probability distributions over the same support  $\Theta$ . The Kullback-Leibler divergence of  $Q(\theta)$  relative to  $P(\theta)$  is defined as

$$\text{KL}(\mathbf{P}(\theta) \| \mathbf{Q}(\theta)) = \mathbb{E}_{\mathbf{P}(\theta)} \left[ \log \frac{\mathbf{P}(\theta)}{\mathbf{Q}(\theta)} \right].$$

**Definition (Cross-entropy).** Let  $\mathbf{P}$  and  $\mathbf{Q}$  be two probability distributions over the same support  $\Theta$ . The cross-entropy of  $\mathbf{Q}(\theta)$  relative to  $\mathbf{P}(\theta)$  is defined as

$$H_{\mathbf{P}, \mathbf{Q}} = -\mathbb{E}_{\mathbf{P}(\theta)} \log \mathbf{Q}(\theta)$$

**Proposition (Posterior agreement criterion).** The posterior agreement model-selection criterion is defined as follows.

$$\begin{aligned} & \sup_{\mathcal{F}} \mathcal{J} \\ \text{s.t. } & \text{KL}(\mathbf{\Pi}^f(\theta) \parallel |\Theta|^{-1}) \leq \xi, \end{aligned}$$

where  $\xi \in \mathbb{R}$  represents a small allowed deviation from uniformity in the prior.

**Theorem 2.3.1 (Maximum posterior agreement).** The optimal  $\mathbf{P}_*$  maximizing the posterior agreement criterion defines a lower bound in the generalization error  $\mathcal{G}_{\mathcal{X}}$  under the richness condition:

$$\inf_{\mathcal{F}} \mathcal{G}_{\mathcal{X}} \geq -\sup_{\mathcal{F}} \mathcal{J}.$$

*Proof.* We consider the lagrangian formulation of the generalization error minimization problem and apply Lemma 2.3.1.

$$\begin{aligned} & \inf_{\mathcal{F}} \{ \mathcal{G}_{\mathcal{X}} + \alpha \text{KL}(\mathbf{\Pi}^f(\theta)) \parallel |\Theta|^{-1} \} \\ &= \inf_{\mathcal{F}} \{ \mathcal{G}_{\mathcal{X}} + \alpha \mathbb{E}_{\mathbf{\Pi}^f(\theta)} \log \mathbf{\Pi}^f(\theta) + \alpha \mathbb{E}_{\mathbf{\Pi}^f(\theta)} \log |\Theta| \} \\ &\geq \alpha \log |\Theta| + \inf_{\mathcal{F}} \{ \alpha H_{\mathbf{\Pi}^f} \} - \sup_{\mathcal{F}} \{ \mathcal{J} \} \\ &\geq -\sup_{\mathcal{F}} \mathcal{J} \end{aligned}$$

The last inequality follows from the fact that the entropy does not exceed the log-cardinality of the hypothesis class:

$$H_{\mathbf{\Pi}^f}(\theta) \leq \log |\Theta|, \quad \forall \mathbf{\Pi}^f.$$

□



# Chapter 3

## Experimental setup

This chapter delineates the covariate shift setting within the supervised classification framework and introduces an operative formulation of posterior agreement. This formulation enables robustness-based model selection in discrete hypothesis class scenarios.

### 3.1 Problem formulation

Out of all the possible learning problems in which a distribution shift can be defined, this project will focus on the supervised classification of images. The function space to navigate is composed of parametrized classifiers.

**Definition (Classifier).** Let  $\mathcal{X}$  and  $\mathcal{Y} \subset \mathbb{N}$  be the input and output spaces of the target function, respectively. Let  $K \in \mathbb{N} < \infty$  be the cardinality of  $\mathcal{Y}$ .

- Let  $\Phi$  be a feature extractor, mapping the input space to a  $d$ -dimensional feature space.

$$\begin{aligned}\Phi : \mathcal{X} &\longmapsto \mathbb{R}^d \\ x &\longmapsto \Phi(x) = z\end{aligned}$$

- Let  $\mathbf{F}$  be a discriminant function, assigning a score to each of the  $K$  classes.

$$\begin{aligned}\mathbf{F} : \mathbb{R}^d &\longmapsto \mathbb{R}^K \\ z &\longmapsto (F_1(z), \dots, F_K(z)) = \mathbf{F}(z)\end{aligned}$$

- Let  $\eta$  be a decision rule, yielding the class label from a vector of scores. It will be set to the maximum a posteriori (MAP) rule.

$$\begin{aligned}\eta : \mathbb{R}^K &\longmapsto \mathcal{Y} = \{1, \dots, K\} \\ \mathbf{F}(z) &\longmapsto \hat{y} = \arg \max_k F_k(z)\end{aligned}$$

A  $K$ -class classifier can be defined as the composition of these three functions:

$$c = \eta \circ \mathbf{F} \circ \Phi.$$

The results presented in this work are limited to neural network classifiers, which entail a parametrization in  $\Gamma \subseteq \mathbb{R}^{|\Gamma|}$  that can be expressed as

$$\begin{aligned}c : \mathcal{X} \times \Gamma &\longmapsto \mathcal{Y} = \{1, \dots, K\} \\ (x, \gamma) &\longmapsto c(x; \gamma) = \hat{y},\end{aligned}$$

thus  $c(x; \gamma) = \eta \circ (\mathbf{F} \circ \Phi)(x; \gamma)$ .

The concepts introduced in the previous chapter provide the foundation for the formalization of the learning problem in which our robustness experiments will be conducted. We will refer to this problem as a  $K$ -class classification.

**Definition ( $K$ -class classification).** Let  $D \in \mathcal{D}$  be a supervised dataset. Let  $c$  be a neural network classifier, parametrized in  $\Gamma \subseteq \mathbb{R}^{|\Gamma|}$ . Let RRM be the regularized risk minimization problem for  $c$  on  $D$ . Let  $L$  be the cross-entropy loss function for classifier  $c$ , computed as

$$L(x, y) = -\log F_y(\Phi(x); \gamma).$$

The  $K$ -class classification is the RRM problem parametrized in  $\Gamma$  with cross-entropy loss  $L$ :

$$\gamma^* = \arg \min_{\gamma \in \Gamma} -\frac{1}{N} \sum_{n=1}^N \log F_{y_n}(x_n; \gamma) + \lambda \Omega(\gamma)$$

No further characterization of the regularization factor  $\lambda \Omega(\gamma)$  will be provided in this section, as the specific learning algorithms considered will be introduced later in the text.

## 3.2 Robustness in covariate shift settings

The concept of robustness, as defined in the previous chapter, entails a measure of the stability of the learner to the randomness of the data sampling process, but also requires an adequate characterization of such randomness. In the context of the  $K$ -class classification problem, sampling randomness can be formalized as a shift in the distribution of the input space, also known as covariate shift.

**Definition (Covariate shift).** Let  $\mathbf{x}'$  and  $\mathbf{x}''$  be two samples of  $\mathbf{X} \stackrel{\text{iid}}{\sim} X$  with size  $N$ . A covariate shift exists between  $\mathbf{x}'$  and  $\mathbf{x}''$  if their (empirical) distributions are significantly different<sup>1</sup> for  $N$  large enough. This is expressed as:

$$\mathbf{P}_{\mathbf{x}'} \not\sim \mathbf{P}_{\mathbf{x}''}.$$

It must be noted that, since the target function is assumed to be invariant (see Section 1.1.2), the true distribution over the output space remains the same [48].

The presence of covariate shift as defined above already leads to a non-zero generalization error, given that  $\mathbf{x}'$  and  $\mathbf{x}''$  represent different randomness instantiations and result in different learning outcomes. Nevertheless, this definition can be further expanded to encompass more practical sources of shift in the context of classification tasks.

**Definition (Distribution shift).** Let  $X'$  and  $X''$  be two random variables associated to different sampling experiments in  $\mathcal{X}$  such that  $P_{X'} \not\sim P_{X''}$ . The effective randomness entailed by their respective measurement process is also different in general (i.e.  $\mathcal{T}_{X'} \neq \mathcal{T}_{X''}$ ). In such case

$$\mathbf{x}' \sim \mathbf{X}' \stackrel{\text{iid}}{\sim} X' \text{ and } \mathbf{x}'' \sim \mathbf{X}'' \stackrel{\text{iid}}{\sim} X''$$

lead to a covariate shift known as out-of-distribution, given that the fundamental source of variability is the difference in the probability measure over the support induced by each experiment [48]. In simpler terms,  $X' \neq X'' \implies \mathbf{P}_{\mathbf{x}'} \not\sim \mathbf{P}_{\mathbf{x}''}$  in general.

In the out-of-distribution setting,  $\mathbf{x}'$  and  $\mathbf{x}''$  are drawn from different random variables, each with a distinct probability landscape over the support, namely source and target domains, that result in implicit and/or explicit differences (sometimes unbalanced) in the distribution of some features. Therefore, empirical distributions  $\mathbf{P}_{\mathbf{x}'}$  and  $\mathbf{P}_{\mathbf{x}''}$  will be different in general and induce a covariate shift that leads to a non-zero generalization error.

---

<sup>1</sup>The notion of difference relies on the nature of the data. Common measures include statistical distances such as the Kullback-Leibler divergence, Wasserstein distance, or even simpler metrics like the difference in means or variances. These methods help establish whether observed differences are statistically significant. [48]

**Definition (Adversarial shift).** Let  $\mathbf{x}' \sim \mathbf{X}$  be a sample drawn from experiment  $X$ . Let  $\Delta$  be a perturbation over the sample space. In this case,  $\mathbf{x}''$  is generated by perturbing  $\mathbf{x}'$  as

$$\mathbf{x}'' = \mathbf{x}' + \Delta,$$

which induces a covariate shift known as adversarial, given that perturbation  $\Delta$  is crafted ad-hoc to hinder the output of the model.

In the adversarial setting, sampling randomness is not the source of covariate shift, as both  $\mathbf{x}'$  and  $\mathbf{x}''$  arise from the same realization of the experiment.

Overall, this work must pursue a wider approach to generalization that does not only comprise the implicit randomness embedded in each realization  $\mathbf{x} \sim \mathbf{X}$  but also the explicit shift in the distribution of the input space generated by intentional or unintentional perturbations of the data generation process. This broader interpretation aligns practical covariate shift experiments with the robustness framework introduced in the first chapter.

Once the possible sources of randomness in the data generation process have been established and formalized, a general concept of robustness measure must be introduced accordingly, so that the suitability of posterior agreement as a robustness metric can be assessed.

**Definition (Robustness metric).** Let  $D', D'' \in \mathcal{D}$  be supervised datasets generated from realizations  $\mathbf{x}', \mathbf{x}'' \sim \mathbf{X}$ , respectively. A robustness metric is a function  $\Omega : \mathcal{D}'' \times \mathcal{F} \mapsto \mathbb{R}$  that quantifies the generalization capabilities of a model trained with  $D'$  to observations in  $D''$ .

The baseline robustness metric in supervised classification tasks is accuracy, defined as the proportion of correct predictions achieved by a pre-trained classifier  $\hat{c}$  over dataset  $D''$ :

$$\text{Accuracy} = \frac{1}{N} \sum_{n=1}^N \delta_{y_n''}(\hat{c}(x_n'')), \quad (x_n'', y_n'') \in D''.$$

As it was argued before, generalization will be interpreted from the perspective of the possible learning outcomes of an experiment. The ultimate goal of robustness measurement is thus the characterization of the resolution limit that can be achieved in the hypothesis space consistent with the intrinsic randomness entailed by each possible realization of the experiment.

The optimal resolution limit is not determined by the model but instead by nature of the randomness of the data generation process. Therefore, a robustness metric should evaluate how stable are hypothesis to different realizations of the same experiment in a model-agnostic way. As model complexity increases, so does the number of hypotheses navigated, but this also increases the risk of overfitting to sampling randomness, leading to unstable hypotheses. A regularization or model selection procedure derived from a robustness metric should therefore find the sweet spot between resolution and stability.

**Proposition (Properties of a robustness metric).** A suitable robustness metric should possess the following two properties:

**P1** (Non-increasing) The metric should be non-increasing with respect to the response of the model under increasing levels of covariate shift.

**P2** (Independent discriminability) The metric should discriminate models exclusively by their generalization capabilities against covariate shift. For instance, the metric should be independent of the task performance of the model.

The first property is commonly satisfied, but the second one entails a specific interpretation of stability that is not straightforward to quantify [11]. Let us consider the following example.

**Example 3.2.1.** Let  $\mathcal{D}$  be a class of balanced, binary, supervised datasets, each containing an equal number of observations from both classes. The following three classifiers will be evaluated on observations in  $D \in \mathcal{D}$ :

- A random classifier, returning a random prediction to each observation in the dataset. Overall performance tends to 50% accuracy as dataset size increases.
- A constant classifier, returning exactly the same prediction for each observation in the dataset. Overall performance is 50% accuracy, as the dataset is exactly balanced.
- A perfect classifier, returning the correct prediction to each observation in the dataset. Overall performance is 100% accuracy.

In terms of performance, the random and constant classifiers are equivalent when the dataset is large enough, and the perfect classifier would be selected as the best. Nevertheless, a robustness metric compliant with **P2** would evaluate the random classifier to be non-robust, while both perfect and constant classifiers would be considered equivalent and achieve maximum robustness since their output hypothesis remains constant for every dataset in  $\mathcal{D}$ .

In general, a performance-based robustness metric would discriminate the perfect and constant classifiers based on their accuracy, even if both are maximally robust by construction, and would even consider the latter to be as robust as a random classifier, which is unrobust by definition. It is now straightforward to see that accuracy or any task-dependent metric does not comply with **P2**.

This work will provide a **P2**-compliant robustness metric derived from the concept of posterior agreement. Before that, the statement of the problem must be completed with an extended characterization of adversarial and out-of-distribution shifts from a practical perspective; that is, the specific quantification of the shift magnitude that will be considered in the experiments.

### 3.3 Adversarial setting

The magnitude of adversarial shifts will be quantified by an aggregated measure of the perturbation applied to the each observation in the dataset.

**Definition (Perturbation).** Let  $\mathbf{x}'$  be a realization of  $\mathbf{X} \stackrel{\text{iid}}{\sim} X$  with support  $\mathcal{X} \subset \mathbb{R}^d$ . Let  $x \in \mathbf{x}'$  be an observation of the sample. Let  $\mathbf{B}_p^\epsilon(x)$  be the  $\ell_p$ -norm ball of radius  $\epsilon$  centered at  $x$ . A perturbation  $\Delta$  is defined as

$$\Delta \in \mathbb{R}^d \text{ s.t. } x + \Delta \in \mathbf{B}_p^\epsilon(x),$$

where  $\epsilon \in [0, 1]$  keeps it hard-box constrained due to the normalization of the input space. A perturbation set  $\Delta$  will be  $\epsilon_p$ -constrained if each of its components satisfies the previous definition. In such case,

$$\mathbf{x}'' = \mathbf{x}' + \Delta$$

defines an adversarial shift of magnitude  $\epsilon_p$ .

As it was previously outlined, the existence of adversarial examples was initially associated with their heavily non-linear nature and, as a consequence, to a lack of smoothness over the hypothesis space [55]. Nevertheless, it is instead the linearity of their units and the high dimensionality of inner representations that make them vulnerable to perturbations in certain directions [21].

**Example 3.3.1.** Let  $w \in \mathbb{R}^d$  be the weight vector of a neural network unit. The difference in activation responses between perturbed and original observations

$$w^\top (\mathbf{x}'' - \mathbf{x}') = w^\top \Delta$$

will be maximum when  $\Delta \propto \text{sign}(w)$ ; that is, when the perturbation is aligned with the weights.

Following this intuition, the most adversarial direction of perturbation can be obtained by maximizing the resulting loss.

**Attack (FGSM).** Perturbations are generated by alignment with the gradient of the loss with respect to the original observation:

$$\Delta = \epsilon_p \operatorname{sign}(\nabla_{x'} \mathcal{L}(x', y; \gamma)).$$

This is known as the fast gradient sign method attack [21].

An effective regularizer for adversarial training can be built by including the FGSM term on the objective that makes the model robust to  $\epsilon_p$ -constrained perturbations [21]. A multi-step version can be immediately derived that systematically perturbs observations in the most adversarial direction at each optimization step.

**Attack (PGD).** Perturbations are generated by iteratively applying the FGSM perturbation to each step and projecting the result back to the  $\epsilon_p$ -constrained ball:

$$x^{s+1} = \Pi_{\mathbf{B}_p^\epsilon(x)} (x^s + \epsilon_p \operatorname{sign}(\nabla_{x'} \mathcal{L}(x', y; \gamma))),$$

where  $\Pi$  is the projection operator. This is known as projected gradient descent attack [41].

It can be shown that a PGD regularizer for adversarial training navigates the loss landscape to minimize the model loss under the maximum adversarial perturbation:

$$\gamma^* = \arg \min_{\gamma \in \Gamma} \left\{ \mathbb{E} \left[ \max_{\Delta} \mathcal{L}(f(x + \Delta), y; \gamma) \right] \right\}.$$

The inherent complexity of this optimization problem requires making certain assumptions in order to solve it. For instance, it is commonly assumed that the loss landscape contains numerous local minima, but with very similar values. Then, the distribution of loss values attained with different starting points is well concentrated and has no outliers, which fosters robustness [41].

Our experimental setup will also consider a minimum-norm adversarial training method that works by iteratively finding the sample misclassified with maximum confidence within  $\mathbf{B}_p^\epsilon(x)$ , while adapting its radius to minimize the distance between the perturbed sample and the decision boundary.

**Attack (FMN).** Perturbations are generated as follows:

$$\begin{aligned} \Delta^* &= \arg \min_{\Delta} \|\Delta\|_p \\ \text{s.t. } &F_y(x; \gamma) - \max_{k \neq y} F_k(x; \gamma) < 0, \\ &x + \Delta \in \mathbf{B}_p^\epsilon(x). \end{aligned}$$

This is known as the fast minimum-norm attack [47].

### 3.4 Domain generalization setting

As described in the introductory chapter, domain generalization refers to a specific setting in which several instantiations of the data are shifted in the out-of-distribution sense, and only a subset of them are available. The problem can be formalized as follows:

**Definition (Domain generalization).** Let  $\mathbb{S} = \{X_s^1, \dots, X_s^{|\mathbb{S}|}\}$  and  $\mathbb{T} = \{X_t^1, \dots, X_t^{|\mathbb{T}|}\}$  be two sets of random variables associated with specific probability measures over the input space  $\mathcal{X}$ . The probability measure induced by each random variable implicitly selects a region of the support  $\mathcal{X}$ , so in this context they will be metonymically referred to as domains. The set  $\mathbb{S}$  encompasses source domains, and the set  $\mathbb{T}$  target domains (see Section 1.1.2) [38, 60].

According to Definition 3.2, datasets sampled from each domain exhibit an out-of-distribution shift resulting in a non-zero generalization error. The domain generalization problem involves selecting the model with the lowest generalization error between source target domains without having access to target domains at all.

Unlike the adversarial case, there is no standard way of quantifying the magnitude of the shift besides reporting model performance in benchmark datasets. These datasets encode specific variations in the causal structure that generates the data, and a robustness metric is expected to be sensitive to the intensity of these variations.

In this work, we will explore an epistemologically-grounded approach to robustness assessment that is agnostic to the nature of the process generating the images and in general to the concept of image itself. Even though some general-purpose metrics exist to evaluate structural similarity between pairs of image-representing tensors (see [23]), only the geometrical properties of the feature space and the resulting probability distribution over the output space will be used to quantify the shift. Even though this approach might seem biased towards the specific parameters of the classifier, the model selection problem will be reformulated so that any bias contribution can be accounted for under the expected randomness of the experiment.

Taking into account the magnitude of the existing covariate shift among source domains, the performance of the selected model will be reported for each of the target domains. In particular, average accuracy and worst-case accuracy will be provided [73].

### 3.5 Robust learners

This project will evaluate the suitability of posterior agreement as a robustness metric in the adversarial and out-of-distribution settings. In accordance with **P2** (see Properties 3.2), it must be eventually assessed whether the PA metric is able to differentiate between robust and non-robust models and, consequently, between learning algorithms selecting those models. For that reason, we will consider ERM as our baseline vanilla algorithm and compare its generalization performance under covariate shift with two algorithms representing two different robustness-fostering strategies.

As a first approach, NN architectures will be trained by means of IRM, a regularization method driven by feature alignment [2]. IRM follows a domain-invariant representation learning strategy emerging from the hypothesis of invariance of the causal structure of the input-output relation. The existence of representations encoding that causality in the feature space is assumed so that the invariance of such representations under different source domains can be enforced [38].

**Definition (IRM).** Let  $R^d$  be the risk of a classifier  $c$  over domain  $d \in \mathbb{S}$ . The IRM problem minimizes risk over all domains while enforcing the feature extractor to yield domain-invariant representations [2]:

$$\begin{aligned} c^* = \min_{c=\eta \circ \mathbf{F} \circ \Phi} & \sum_{d \in \mathbb{S}} R^d(c) \\ \text{s.t. } (\eta \circ \mathbf{F}) = \arg \min_{\bar{c}} & R^d(\bar{c}) \quad \forall d \in \mathbb{S}. \end{aligned}$$

A surrogate version of the problem simplifies its implementation:

$$c^* = \min_c \sum_{d \in \mathbb{S}} R^d(c) + \lambda \|\nabla_{w|w=1} R^d(w \cdot c)\|^2,$$

where  $w$  is a dummy classifier added to the problem to relax the invariance constraint and enforce instead that the optimal feature representation induces an optimal classifier that is the same in all domains (see [2] for details). The balance between the ERM term and the invariance predictor is controlled by the regularization hyperparameter  $\lambda \in [0, \infty)$ .

As a second approach, we will consider a data generation strategy that populates the gaps among source domain distributions with new observations obtained via interpolation. Learning invariant features via selective augmentation (LISA) is accomplished by interpolating original samples that either belong to the same class but a different source domain (LISA-D), or belong to the same domain but have different labels (LISA-L). The former helps the model learn domain-invariant features, while the latter fosters the learning of class-invariant features. Two interpolation strategies will be considered, namely Mixup [72] and CutMix [69].

**Definition (LISA).** Let  $D_1$  and  $D_2$  be datasets associated with two different source domains. A convex interpolation with weight  $\lambda \sim \text{Beta}(\alpha, \beta)$  generates a new sample that lies in the line segment connecting the two original samples.

(LISA-D) Let  $(x_1, y_1) \in D_1$  and  $(x_2, y_2) \in D_2$ , with  $y_1 = y_2$ ,

(LISA-L) Let  $(x_1, y_1), (x_2, y_2) \in D_1$ , with  $y_1 \neq y_2$ ,

$$\begin{aligned} x_{LISA} &= \lambda x_1 + (1 - \lambda)x_2 \\ y_{LISA} &= \lambda y_1 + (1 - \lambda)y_2, \end{aligned}$$

where a random value  $s \in \text{Bernoulli}(p)$  will determine the strategy to be applied, being  $p \in [0, 1]$  the probability of LISA-L [67].

## 3.6 Robustness assessment with posterior agreement

As it was argued in Section 3.2, the extended practice of quantifying robustness by reporting test accuracy in benchmark datasets does not offer any theoretical mechanism for the characterization of the model and is ultimately biased towards the nature of the data and the shift. In this section, a practical version of posterior agreement will be derived, so that it can be used in classification tasks to evaluate generalization capabilities under different sources of randomness.

From an information-theoretic perspective, a fundamental distinction between PA and accuracy can already be stated, namely the fact that PA is computed with the output of the discriminant function  $\mathbf{F}$ , which encodes information about the prediction confidence, whereas accuracy is limited to the decision  $\eta \circ \mathbf{F}$ . Primarily, confidence information increases the discriminative power of a robustness metric, particularly when comparing models with similar performance. Nevertheless, it also allows for a more consistent assessment of the true generalization capabilities, as it is less affected by low-information sources of randomness such as sampling variability within the same distribution or even random noise, regardless of their effect on the task at hand.

### 3.6.1 Posterior in classification tasks

Definition 2.3.1 established the posterior as a probability distribution over the hypothesis class encoding the stochastic nature of model outputs. The hypothesis class  $\Theta$  of a  $K$ -class classification problem is the set of all possible vectors of labels associating each of the  $N$  samples to one of the  $K$  classes. This amounts to

$$\Theta = \{1, \dots, K\}^N$$

with cardinality  $|\Theta| = K^N$ .

**Proposition (Classification confidence).** Let  $D \in \mathcal{D}$  be a supervised dataset generated from realization  $\mathbf{x} \sim \mathbf{X}$ . Let  $F_k$  be the  $k$ -th component of the prediction scores vector. The cost function driving posterior selection will be the negative prediction confidence:

$$R(\theta, \mathbf{x}; \gamma) = \sum_{n=1}^N -F_{\theta_n}(x_n; \gamma),$$

where  $\theta_n \in \mathcal{Y}$  is the prediction to observation  $x_n \in D$ .

**Theorem 3.6.1** (*Classification posterior*). Let  $\Theta$  be the hypothesis class associated with the  $K$ -class classification problem with classifier  $c$ . The family of posterior distributions  $\mathcal{P}^c$  to consider is the Gibbs distribution with inverse temperature parameter  $\beta$  [11]

$$\mathbf{P}^c(\theta | \mathbf{x}) = \frac{\exp(\beta R(\theta, \mathbf{x}; \gamma))}{\sum_{\theta \in \Theta} \exp(\beta R(\theta, \mathbf{x}; \gamma))}.$$

*Proof.* The proof is based on the maximum entropy principle (MEP), which states that given some prior testable information to be encoded by a probability distribution, the distribution that best encodes that information is the one minimizing additional assumptions besides the testable information; that is, the one maximizing information entropy within the testable space [28]. Testable information amounts to certain constraints on the MEP optimization problem over the non-negative, Lebesgue-integrable function class  $\mathcal{P}$ .

$$\begin{aligned} & \max_{\mathbf{P}^c(\theta | \mathbf{x}) \in \mathcal{P}} H_{\mathbf{P}^c}(\theta | \mathbf{x}) \\ \text{s.t. } & \sum_{\theta \in \Theta} \mathbf{P}^c(\theta | \mathbf{x}) = 1 \\ & \mathbb{E}_{\mathbf{P}^c(\theta | \mathbf{x})} R(\theta, \mathbf{x}) = \mu \quad \forall \theta \in \Theta \\ & (\mathbf{P}^c(\theta_i | \mathbf{x}) - \mathbf{P}^c(\theta_j | \mathbf{x})) (R(\theta_i, \mathbf{x}) - R(\theta_j, \mathbf{x})) \geq 0 \quad \forall \theta_i, \theta_j \in \Theta \end{aligned}$$

where  $\mu \in \mathbb{R}$  is a hyperparameter ensuring that the expected confidence is finite and the last constraint imposes a monotonic relationship between the confidence and the posterior. The lagrangian formulation of the problem with equality constraints is

$$\Lambda(\mathbf{P}^c, \alpha, \beta) = H_{\mathbf{P}^c}(\theta | \mathbf{x}) + \alpha \left( 1 - \sum_{\theta \in \Theta} \mathbf{P}^c(\theta | \mathbf{x}) \right) + \beta \left( \mathbb{E}_{\mathbf{P}^c(\theta | \mathbf{x})} [R(\theta, \mathbf{x})] - \mu \right).$$

Its derivative with respect to  $\mathbf{P}^c(\theta | \mathbf{x})$  can be computed as

$$\frac{\partial \Lambda}{\partial \mathbf{P}^c(\theta | \mathbf{x})} = -1 - \log \mathbf{P}^c(\theta | \mathbf{x}) - \alpha + \beta R(\theta, \mathbf{x}),$$

which has a unique solution

$$\mathbf{P}^c(\theta | \mathbf{x}) = \frac{\exp(\beta R(\theta, \mathbf{x}))}{\exp(1 + \alpha)}.$$

Setting  $\exp(1 + \alpha) = \sum_{\theta \in \Theta} \exp(\beta R(\theta, \mathbf{x}))$  and  $\beta \geq 0$  we ensure normalization and the fulfillment of the monotonic relationship constraint.  $\square$

From a statistical physics perspective, a dataset can be interpreted as system of  $N$  particles in thermal equilibrium with a thermal bath at  $T \propto 1/\beta$ . Under the Maxwell-Boltzmann statistics of ideal gases, hypotheses are considered states of the system, and the confidence in the prediction as the energy of each state. The normalization factor arises naturally from this perspective, as it corresponds to the partition function of the system. The posterior expression can be derived analogously by enforcing the MEP principle (in this case, the second law of thermodynamics) under the constraints of finite energy and number of particles [7].

### 3.6.2 The posterior agreement kernel

**Lemma 3.6.1** (*Exchangeability*). Let  $N, K \in \mathbb{N}$  and let  $\{\mathcal{E}_{ij} \mid i \leq N, j \leq K\}$  be an indexed set of values. Then,

$$\sum_{c \in \mathcal{C}} \prod_{n=1}^N \mathcal{E}_{i,c(n)} = \prod_{n=1}^N \sum_{k=1}^K \mathcal{E}_{ij}$$

*Proof.* See Appendix A.1.  $\square$

**Lemma 3.6.2** (*Posterior factorization*). The posterior distribution for a classification problem can be factorized as follows:

$$\mathbf{P}^c(\theta \mid \mathbf{x}) = \prod_{n=1}^N \mathbf{P}_n^c(\theta_n \mid \mathbf{x}) = \prod_{n=1}^N \frac{\exp(\beta F_{\theta_n}(x_n))}{\sum_{k=1}^K \exp(\beta F_k(x_n))}$$

*Proof.* See Appendix A.1.  $\square$

**Theorem 3.6.2** (*PA kernel for classification*). Let  $\mathbf{x}'$  and  $\mathbf{x}''$  be  $N$ -sized realizations of  $\mathbf{X}$ . Let  $\Theta$  be the hypothesis class represented by classifier  $c$  under  $\mathcal{D}$ . With no prior information about  $\Theta$ , the posterior agreement kernel for supervised  $K$ -class classification tasks has the following expression:

$$\text{PA}(\mathbf{x}', \mathbf{x}''; \beta) = \frac{1}{N} \sum_{n=1}^N \log \left\{ |\Theta| \sum_{k=1}^K \mathbf{P}^c(k \mid x'_n) \mathbf{P}^c(k \mid x''_n) \right\},$$

where  $\mathbf{P}^c(j \mid x_n)$  can be shown to be

$$\mathbf{P}^c(k \mid x_n) = \frac{\exp(\beta F_k(x_n))}{\sum_{q=1}^K \exp(\beta F_q(x_n))}.$$

*Proof.* The posterior agreement  $\mathcal{J}$  has the following expression, derived in Lemma 2.3.1:

$$\mathcal{J} = \mathbb{E}_{\mathbf{P}_{\mathbf{x}', \mathbf{x}''}} \left[ \log \left( \mathbb{E}_{\mathbf{P}^c(\theta \mid \mathbf{x}')} \frac{\mathbf{P}^c(\theta \mid \mathbf{x}'')}{\Pi^c(\theta)} \right) \right].$$

As defined previously,  $\Theta$  is a discrete, finite set of possible classification vectors of the  $N$  observations, and the sampling distribution  $\mathbf{P}_{\mathbf{x}', \mathbf{x}''}$  is assumed to be uniform. Therefore, the expectation operators amount to:

$$\begin{aligned} \mathbb{E}_{\mathbf{P}_{\mathbf{x}', \mathbf{x}''}} &= \frac{1}{N} \sum_{n=1}^N . \\ \mathbb{E}_{\mathbf{P}^c(\theta \mid \mathbf{x}')} &= \sum_{\theta \in \Theta} \mathbf{P}^c(\theta \mid \mathbf{x}') . \end{aligned}$$

A non-informative prior is assumed, thus enforcing the richness condition

$$\Pi^c(\theta) = |\Theta|^{-1}.$$

$\mathbf{P}^c(\theta \mid \mathbf{x})$  can be factorized on the terms expressed in Theorem 3.6.2.

$$\mathbf{P}^c(\theta \mid \mathbf{x}) = \prod_{n=1}^N \mathbf{P}_n^c(\theta_n \mid \mathbf{x}) = \prod_{n=1}^N \frac{\exp(\beta F_{\theta_n}(x_n))}{\sum_{k=1}^K \exp(\beta F_k(x_n))}.$$

Operating analogously for  $\mathbf{x}'$  and  $\mathbf{x}''$ , the expression for the PA kernel is obtained.

$$\begin{aligned} \text{PA}(\mathbf{x}', \mathbf{x}''; \beta) &= \frac{1}{N} \sum_{n=1}^N \left[ \log \left( \sum_{\theta \in \Theta} \mathbf{P}^c(\theta \mid \mathbf{x}') \frac{\mathbf{P}^c(\theta \mid \mathbf{x}'')}{|\Theta|^{-1}} \right) \right] \\ &= \frac{1}{N} \sum_{n=1}^N \left[ \log \left( |\Theta| \sum_{\theta \in \Theta} \mathbf{P}^c(\theta \mid \mathbf{x}') \mathbf{P}^c(\theta \mid \mathbf{x}'') \right) \right] \\ &= \frac{1}{N} \sum_{n=1}^N \left[ \log \left( |\Theta| \sum_{\theta \in \Theta} \prod_{n=1}^N \frac{\exp(\beta F_{\theta_n}(x'_n))}{\sum_{k=1}^K \exp(\beta F_k(x'_n))} \frac{\exp(\beta F_{\theta_n}(x''_n))}{\sum_{k=1}^K \exp(\beta F_k(x''_n))} \right) \right]. \end{aligned}$$

Finally, applying Lemma 3.6.1 to the product inside the logarithm, we reach the final expression.  $\square$

Once the expression of the posterior agreement for classification tasks has been reached, we can proceed to analyze its properties and its suitability as a robustness metric.

**Theorem 3.6.3** (*Properties of the PA kernel*). The posterior agreement kernel for  $K$ -classification problems  $\text{PA}(\mathbf{x}', \mathbf{x}''; \beta)$  has the following properties  $\forall \mathbf{x}', \mathbf{x}'' \sim \mathbf{X}$  and  $\beta \in \mathbb{R}^+$ .

**P1** (Boundedness)  $\text{PA}(\mathbf{x}', \mathbf{x}''; \beta) \leq N \log K$ . This makes sense from an information-theoretical perspective, as the minimum description length in the absence of generalization error amounts to that of the prior, which was set to be non-informative. It is clear that  $\log_2 K$  bits are needed to encode a uniform distribution over the classes for each observation.

**P2** (Symmetry)  $\text{PA}(\mathbf{x}', \mathbf{x}''; \beta) = \text{PA}(\mathbf{x}'', \mathbf{x}'; \beta)$ . This property is important from the robustness perspective, given that randomness instantiations are not indexed and no reference experiment can be performed.

**P3** (Concavity)  $\text{PA}(\mathbf{x}', \mathbf{x}''; \beta)$  is a concave function of  $\beta \in \mathbb{R}^+$ . This means that the kernel optimization problem will have a unique solution [8].

*Proof.* See Appendix A.2. □

**Theorem 3.6.4.** The posterior agreement kernel for classification tasks complies with the desired properties of a robustness metric, as defined in Properties 3.2.

*Proof.* See Appendix A.2. □

### 3.6.3 Implementation

Following the derivation in the previous section, an operative version of the posterior agreement kernel for the  $K$ -class classification problem is given by

$$\text{PA}(\mathbf{x}', \mathbf{x}''; \beta) = \sum_{n=1}^N \log \left\{ \sum_{k=1}^K \mathbf{P}_n^c(j | \mathbf{x}') \mathbf{P}_n^c(j | \mathbf{x}'') \right\},$$

where factors  $|\Theta|$  and  $1/N$  have not been considered as they are merely for scale. Kernel values are now bounded to the range  $(-\infty, 0]$ .

HERE TALK ABOUT THE RANGE OF OPTIMAL PA AND HOW IT WILL NOT BE INDICATED

The goal of this project is to use  $\text{PA}(\mathbf{x}', \mathbf{x}''; \beta^*)$  to assess the generalization capabilities of different models and eventually select the most robust one at the epoch level. Having to solve the optimization problem for each set of weights is a computationally demanding task, both in terms of memory and time, and hard-coding the kernel optimization inside the training loop for every task is also highly impractical.

For these reasons, a custom implementation<sup>2</sup> of the kernel optimization task has been developed with the purpose of this project. The implementation is wrapped under the the Torchmetrics<sup>3</sup> framework, which allows for a flawless integration in any machine learning project. Here is a small code snippet showcasing a simple evaluation:

---

<sup>2</sup><https://github.com/viictorjimenezzz/pa-metric>

<sup>3</sup><https://lightning.ai/docs/torchmetrics/stable/>

```
1  from pa metric import PosteriorAgreementBase, LogitsDataset
2
3  pa_metric = PosteriorAgreementBase(pa_epochs, beta_0)
4  results = pa_metric(
5      LogitsDataset([logits0, logits1], y)
6  )
7
8  logPA, beta = results["logPA"], results["beta"]
```

**Properties** (PA metric implementation). The most relevant features of the numerical implementation of the metric are the following:

- The metric launches a single optimization process with a pair of datasets. Nevertheless, additional datasets can be added for validation purposes. The metric also accepts different models, which can be useful in cross-validation settings.
- If data samples are not corresponding between datasets, the metric can implement several pairing strategies, namely label matching, nearest-neighbor or canonical correlation.
- Multi-device computation is supported. In particular, a distributed-data-processing (DDP) strategy on a CUDA-managed set of GPUs can be used to parallelize model evaluation.
- The output of the model is computed just once per training epoch, which reduces drastically the time required for the optimization of the kernel.
- Detailed information about the optimization process can be easily retrieved and logged, which helps tuning the optimization algorithm and other hyperparameters.

The complete code implementation along with the set of unit tests conducted to ensure consistency between training and data processing strategies can be found in the `pa-metric` repository<sup>4</sup>.

<sup>4</sup><https://github.com/viictorjimenezzz/pa-metric>



# Chapter 4

## Robustness assessment

The fundamental goal of this project is to assess the suitability of posterior agreement as a robust model selection criterion in the image classification setting. This chapter will explore the properties of the PA kernel as a robustness metric in the adversarial and out-of-distribution settings and will provide evidence supporting its suitability against baseline accuracy-based metrics.

### 4.1 In-distribution setting

Before addressing more relevant sources of covariate shift, an exploratory analysis of the behaviour of PA under random (noise) perturbations will be conducted. This setting is denoted as in-distribution because a single sampling experiment is considered and perturbations are generated randomly in a model-agnostic way, unlike in the adversarial setting.

The first experiment will explore the behavior of the metric under different levels of observation mismatch following Example 3.2.1. Perfect, constant and random classifiers will be considered under different conditions. The goal of this experiment is to empirically assess whether the implemented PA kernel satisfies the properties of a robustness metric established in Proposition 3.2.

**Experiment 1.** A binary sample  $\mathbf{y} \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p)$  of size 1000 is generated, with  $p = \mathbf{P}_Y(y = 1)$ . The nature of the covariates is irrelevant, as the predictive outcome of each classifier is obtained artificially. In particular, two classification outputs  $\hat{\mathbf{y}}', \hat{\mathbf{y}}''$  are considered.

- For a perfect classifier, predictions are  $\hat{\mathbf{y}}' = \hat{\mathbf{y}}'' = \mathbf{y}$ .
- For a constant classifier, predictions are  $\hat{\mathbf{y}}' = \hat{\mathbf{y}}'' = \mathbf{0}$ .
- For a random classifier, predictions are generated by randomly permuting  $\mathbf{y}$ , so that the number of mismatched observations depends on the value of  $p$ .

Furthermore, several levels of prediction confidence are simulated by artificially generating the output of the discriminant function of the classifier. In particular,

$$F_{\hat{y}} = (-1)^{\hat{y}} \frac{\Delta}{2}, \quad \text{and} \quad F_{1-\hat{y}} = (-1)^{(1-\hat{y})} \frac{\Delta}{2},$$

for  $\hat{y} = \{\hat{y}', \hat{y}''\}$ . In that way,  $\mathbf{F}$  discriminates the predicted and non-predicted class by  $\Delta \in \mathbb{R}$ .

Classifier	Accuracy	$\beta^*$	PA
Perfect	1.000	12.331	-0.0088218
Constant	0.525	12.331	-0.0088218
Random	0.516	0.000	-693.14

Table 4.1: Comparison of accuracy and PA values for the case  $p = 0.5$ .

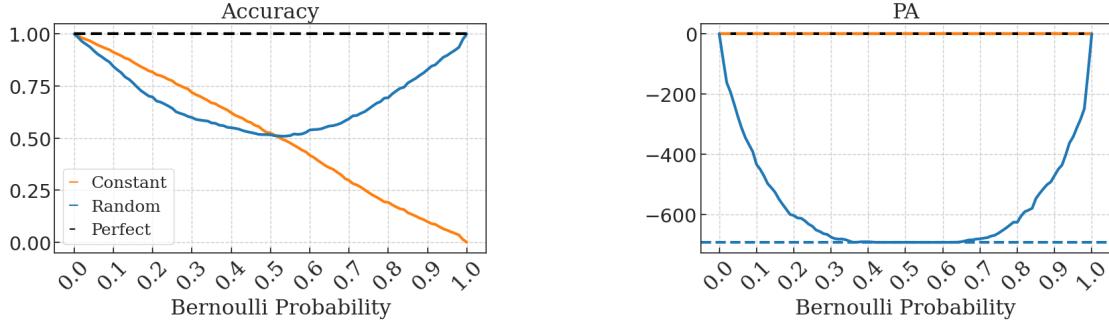


Figure 4.1: Evolution of performance and robustness for the three classifiers

For  $p = 0.5$ , the metric achieves its minimum value  $N \log 1/2$  for the random classifier, as  $\beta = 0$ , and tends to its maximum value of 0 as  $\beta \rightarrow \infty$ .

The random classifier sample was generated by permuting the original so that the number of mismatched observations depends on the bernoulli probability  $p$ . As we can see, the theoretical minimum value is obtained only after a certain perturbation threshold has been reached. This illustrates the trade-off navigated during the kernel optimization, in which matching samples penalize the metric value the lower the value of  $\beta$  is, whereas mismatching samples will penalize the overall value the further from zero  $\beta$  is. Given the highly non-linear nature of the logarithm in the interval  $[0, 1]$ , the metric will penalize disagreement much more than agreement, as shown in Figure 4.1. A truly random classifier (i.e. balanced on the two classes) would yield the minimum PA value for any possible original sample, as shown in the blue dashed line.

One of the most relevant differences between posterior agreement and any accuracy-based measure, as discussed in previous chapters, is the fact that its assessment is based on the whole probabilistic output of the model, and therefore can be used as a measure of confidence in the predictions. The prediction confidence, expressed as a difference in the unnormalized log-odds (commonly known as logits), is very informative with regard to the quality of the model, as we can intuitively infer that the latent space represented by a high-confidence model encodes a better set of features to discriminate observation classes than one with a lower prediction confidence.

For instance, when comparing two models of similar predictive power but at different confidence levels, maximum posterior agreement will be achieved with a higher  $\beta^*$  for the model that tends to yield more flattened distributions. This information is especially valuable in the covariate shift setting, given that robust models that rely on less accessible sets of features will most likely yield conservative predictions on in-distribution samples, but at the same time keep a high prediction power on out-of-distribution observations.

HERE INTERPRETATION OF INFORMATIVENESS AND BETA

ALSO HERE INTERPRETATION OF BETA CONVERGENCE AND WHY WE DONT SET IT MANUALLY.

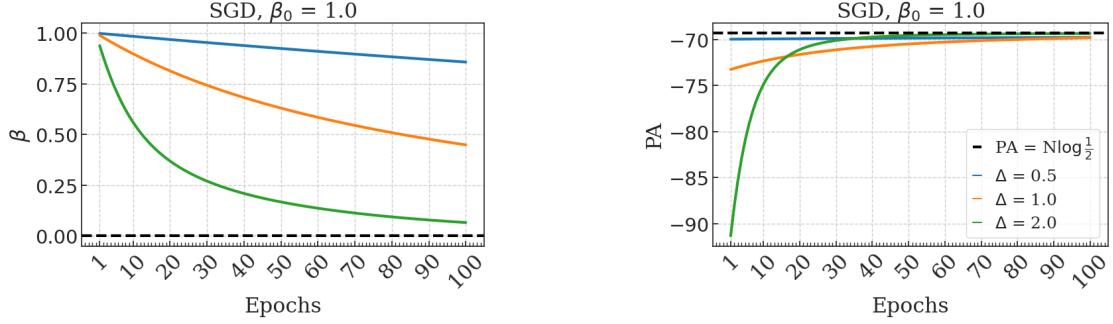


Figure 4.2: Evolution of PA kernel optimization under different levels of prediction confidence. An illustration of the original log-odds and its associated posterior distribution can be found in Appendix B.1.1.

These and other results (see Appendix B.1.1) indicate that the PA kernel behaves as expected and is highly informative of the generalization capabilities of the model, provided that the nature of the randomness existing between  $\mathbf{x}'$  and  $\mathbf{x}''$  is known.

The results obtained with artificial samples motivate the exploration of more realistic scenarios. In general, the PA metric is expected to capture the generalization capabilities of any model yielding probabilistic predictions, regardless of the task at hand. This already represents an incredible advantage from an epistemological perspective, as we can argue that the metric is agnostic of the underlying mechanism that generated the data and even to the nature of the data itself.

In order to verify this claim, we will start by evaluating the robustness of two different classifier models in two different domains under increasing levels of random noise. This particular setting, even if highly artificial, is relevant in any classification context, as it represents general measure of the quality of the features learned. The presence of noise, at least at low levels, does not perturb the set of features that define a particular class from a human perspective, and should therefore not perturb very significantly the predictions of robust models.

The first example considered is a sentiment classifier for IMDB, a popular NLP dataset containing 50,000 movie reviews labeled as either positive or negative [40]. A pre-trained DistilBERT-based architecture was adopted, with the appropriate tokenizer [51]. After a 100 epoch training using SGD with a learning rate of  $10^{-3}$ , the model with the highest validation accuracy was selected. Then, the original test dataset  $\mathbf{x}'$  was incrementally perturbed to generate  $\mathbf{x}''$ , by either adding, removing or replacing single characters. This perturbation amounts to increasing the Levenshtein distance  $L$  between identical observations of the dataset, and the attack power is defined to be  $2^L$ .

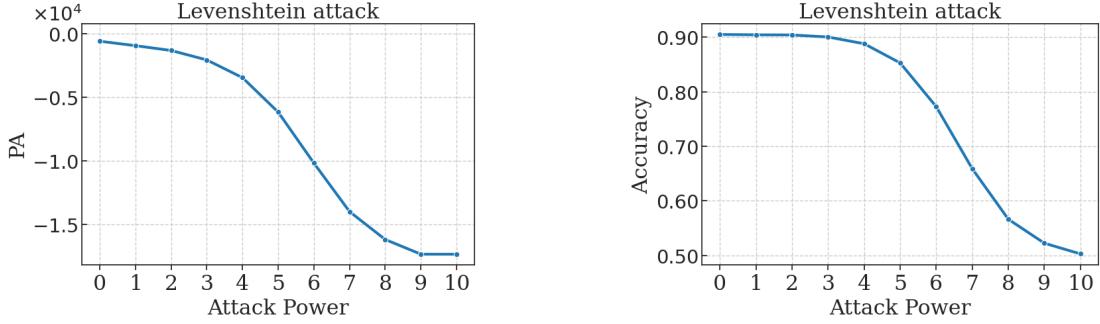


Figure 4.3: PA and accuracy for the IMDB sentiment classification under Levenshtein perturbations. The attack power is defined as  $2^L$ , being  $L$  the Levenshtein distance between pairs of observations in  $\mathbf{x}'$  and  $\mathbf{x}''$ , respectively.

Figure 4.3 shows that PA is highly sensitive to the presence of Levenshtein perturbations, as it is able to capture the change in the predictive outcome of the model even when performance remains unaltered. In particular, accuracy remains maximum even when 4 characters are perturbed in every observation, and begins to decrease only after that. In contrast, PA drops significantly after the first perturbation and reaches its minimum value after  $2^9$  perturbations have been made, indicating that the outcome is maximally unrobust. Further discrepancies between accuracy and PA can be observed if the attack is instead designed to directly influence the sentiment prediction. For instance, Figure B.6 shows that accuracy can be manipulated to increase or decrease by perturbing original samples with adjectives that reinforce or contradict, respectively, the sentiment expressed in the review. In both cases, PA measures the lack of robustness in the predictions that these perturbations represent, regardless of the task performance of the classifier.

As a second example, 10.000 CIFAR10 [33] images were perturbed with white noise of increasing intensity. A pre-trained, undefended WideResNet-28-10 [70] architecture was used to evaluate the effectiveness of the attack. The magnitude of the perturbations is expressed in the same terms as those of an adversarial attack for further reference, but translate to using  $\sigma = 3\ell_\infty$ , as 99.73% of the total mass of the gaussian distribution lies within the interval  $\pm 3\sigma$ .

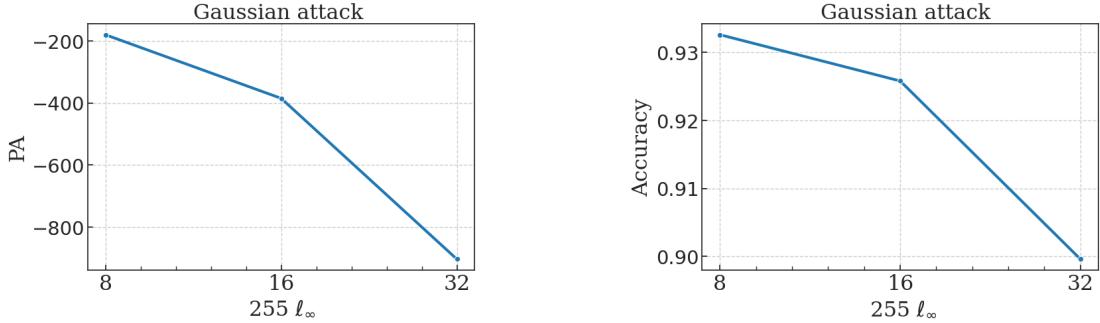


Figure 4.4: PA and accuracy of CIFAR10 classification for increasing levels of white noise.

Figure 4.4 shows that PA is highly sensitive to the presence of white noise and effectively captures the factor of increase in the magnitude of the perturbations. In general, gaussian perturbations genuinely simulate increasing levels of sampling randomness, as the perceptual features that should drive the inductive bias of the model are still present, even if in a proportionally dissimilar way as they were encoded. In contrast to adversarial attacks, these perturbations are not expected to distort the predictive outcome abruptly. As a result, any robustness score will be driven by the

standard generalization error, including validation accuracy.

We can obtain a higher understanding of the degree of sensitivity of the PA metric if we adjust the perturbation so that it only affects a certain ratio of observations.

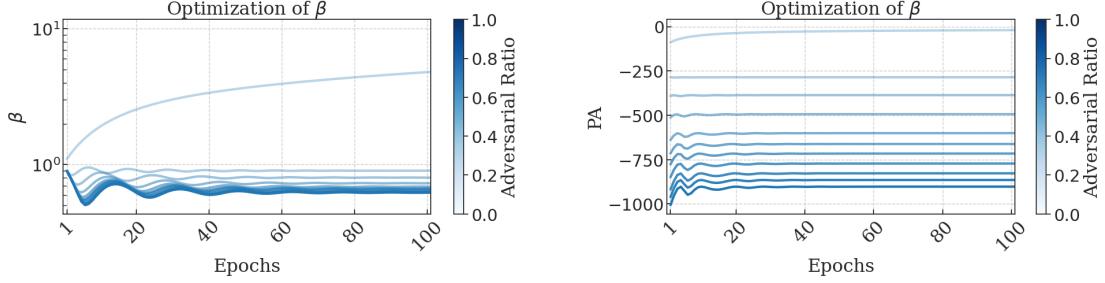


Figure 4.5: PA kernel optimization in the CIFAR10 gaussian noise setting for different ratio of perturbed samples. Perturbation magnitude is  $\ell_\infty = 32 / 255$ .

As expected,  $\beta$  tends to infinity in the unperturbed case, and converges quickly to its optimal value in the rest of cases, even if the sample size is considerably large and memory-intensive. The decay in the PA value is less pronounced the higher fraction of perturbed samples are there, which is consistent with the concept of robustness defined, as it already approaches the lower bound for these kinds of perturbation even when the whole sample has yet not been perturbed.

## 4.2 Adversarial setting

The first scenario in which covariate shift robustness will be tested is the adversarial setting. This setting serves as an archetypal use case for any robustness metric, given that adversarial perturbations are deliberately generated to mislead the model, and any robustness score will ultimately be driven by the effectiveness of the attack. In particular, PA should be highly informative about the defensive capabilities of models, as the posterior distribution over the hypothesis class will shift significantly in the presence of adversarial perturbations. This section aims to validate this claim and provide deeper insights into the nature of the metric.

It is important to note that adversarial perturbations constitute an intermediate instance between sampling randomness and distribution shift. On the one hand, they emulate a sampling variation that appears as an outlier under the model's representation of the true class, even if the source of variability is completely artificial. On the other hand, samples are known to contain the set of features that should align with the inductive bias of the model, and so the model's ability to distillate those features is in question. In practice, we are evaluating the quality of the complex discriminator function defining a basin of stability around original samples, and for that no deep understanding of the nature of the randomness of the samples or the features they encode is needed.

This interpretation is aligned with the measure provided by accuracy-based metrics, because adversarial samples are not expected to contain any relevant features of other classes or express any accountable source of randomness, but instead exploit specific vulnerabilities of models to alter the position of the maximum of the posterior distribution. A greater posterior overlap will still indicate higher robustness to attacks, regardless of the nature of the model or the attack, but optimal posteriors are expected to converge to very peaked gibbs distributions centered at the predicted class, reducing the interpretability of PA to that of accuracy.

In order to explore these claims, robustness and performance results will be provided through the adversarial fidelity ratio (AFR) value and compared to those yielded by PA. The AFR computed with the true class labels will be used as a baseline of model performance, whereas the AFR computed with the predicted class label will be a reference for robustness, as it aligns with the aforementioned interpretation.

**Definition** (Adversarial fidelity ratio). Let  $\hat{\mathbf{y}}', \hat{\mathbf{y}}'' \in \mathcal{Y}^N$  be the predicted class labels for  $\mathbf{x}'$  and  $\mathbf{x}''$ , respectively, and  $\mathbf{y} \in \mathcal{Y}^N$  the true labels. Let ACC be the standard accuracy metric, as defined in Section 3.2. The adversarial fidelity ratio (AFR) is expressed as

$$\begin{aligned} \text{AFR(T)} &= \text{ACC}(\hat{\mathbf{y}}'', \mathbf{y}), \\ \text{AFR(P)} &= \text{ACC}(\hat{\mathbf{y}}'', \hat{\mathbf{y}}'). \end{aligned}$$

The results provided in this section have been obtained using the CIFAR10 dataset [33], which is widely regarded as a standard benchmark for robustness evaluation. CIFAR10 is a balanced dataset containing 60.000 coloured  $32 \times 32$  pixel images belonging to 10 different classes. We will consider a pre-trained WideResNet-28-10 [70] as a baseline, undefended model and compare it to some state-of-the-art robust ResNet50 [25] models provided by the RobustBench [17] library under PGD [41] and FMN [47] attacks, both run for 1000 steps (see Section 3.3). The defenses applied are those proposed by Engstrom et al. [19], Athalye et al. [4], Wong et al. [65], Addepalli et al. [1] and Wang et al. [63]. The PGD attack power will be specified in terms of  $\ell_\infty$ , which corresponds to the maximum perturbation allowed for each pixel. This is consistent with the characterization of adversarial perturbation given in the previous chapter, as every perturbation will be bounded to the region defined by  $\mathbf{B}_\infty^{\ell_\infty}(x)$ .



Figure 4.6: Original and adversarially-perturbed CIFAR10 sample of class `horse`. Both perturbations succeed at misleading an undefended, pre-trained WideResNet-28-10 net.

Besides the maximum norm allowed for each perturbation, we are also interested in evaluating the sensitivity of our robustness measure to the ratio of perturbed samples in the dataset, also known as adversarial ratio (AR). The final adversarial dataset  $\mathbf{x}''$  will be generated as

$$\mathbf{x}'' := \text{AR } \mathbf{x}'' + (1 - \text{AR}) \mathbf{x}',$$

where  $\mathbf{x}'' = \mathbf{x}' + \Delta$ , as per Definition 3.3. This incremental expansion of the attack is particularly relevant for PA, as we would initially expect it to behave non-linearly with respect to AR and converge faster to the the AR = 1 robustness value than any accuracy-based metric, in light of the results obtained in the previous section. We can quantify the model discriminability over increasing AR by computing the adversarial ratio gap  $\Delta \text{AR}$ .

**Definition** (Adversarial ratio gap). Let  $\gamma_+$  and  $\gamma_-$  be two models, not necessarily different. Let  $\text{AR}_+$  and  $\text{AR}_-$  be the adversarial ratio values such that

$$\text{PA}^{\gamma+} \Big|_{\text{AR}_+} = \text{PA}^{\gamma-} \Big|_{\text{AR}_-}$$

the adversarial ratio gap ( $\Delta \text{AR}$ ) is obtained as

$$\Delta \text{AR} = \text{AR}_+ - \text{AR}_-.$$

Before delving into the results, it is worth exploring the immediate consequences of the previous claim, namely the fact that the maximum posterior agreement will be achieved when gibbs distributions are highly peaked on the predicted class, at least for moderately aggressive attacks. This is because most adversarial samples will not succeed at misleading the model and thus drive the inverse temperature to infinity. The divergence of  $\beta^*$  is only limited by the set of misleading adversarial samples, that for being perturbed from the original class are still expected to assign a significant confidence to the original prediction, even if not the maximum anymore. Table 4.2 illustrates this claim by showing that  $\beta^* > 1$  for all robust models, resulting in a substantial decrease of the entropy between initial and optimal posteriors.

Defense	$\beta_{\text{PGD}}^*$	$\Delta H_{\text{PGD}}$	$\beta_{\text{FMN}}^*$	$\Delta H_{\text{FMN}}$
<b>Undefended</b>	0.78	0.048	0.65	0.10
<b>Engstrom et al.</b>	15.63	-1.204	2.59	-0.71
<b>Athalye et al.</b>	35.48	-3.049	19.84	-2.13
<b>Wong et al.</b>	15.46	-1.229	4.59	-0.96
<b>Addepalli et al.</b>	15.89	-2.023	6.08	-1.71
<b>Wang et al.</b>	11.24	-1.833	2.53	-1.41

Table 4.2: Entropy difference  $\Delta H = H(\beta^*) - H(1)$  for different models, obtained for FMN and  $\ell_\infty = 8/255$  PGD attacks, both at  $\text{AR} = 1$ . Entropy values are estimated using the average posterior distribution over correctly classified samples, which constitute the largest proportion of the dataset. Figures ??-?? show the initial and optimal average posteriors from which these values were computed.

This realization allows us to break down the dataset into subsets of observations that contribute to the final PA value in different ways, and therefore improve the interpretation of the resulting robustness measurement. For a start, a robust model should be expected to correctly classify most of the original samples with high confidence, as they contain the discriminative features that define each class. Also in the original dataset, lack of generalization to sampling randomness should be penalized for lowering the confidence in the predicted class. Regarding adversarial samples, a clear distinction between robust and non-robust models should be made based on the success rate of perturbations and the confidence attributed to misleading predictions. Adversarial perturbations on samples originally misclassified will not be of much interest, as the effect on prediction confidence should not be as significant as in the correctly classified ones. An interpretable expression for PA in the adversarial setting can be obtained by approximating the optimal posterior for each of these groups of observations.

**Theorem 4.2.1** (Approximated PA in the adversarial setting). Let  $\Xi_{\text{ERR}}$ ,  $\Xi_{\text{MIS}}$  and  $\Xi_{\text{ADV}}$  be the approximated robustness contributions of correctly classified original samples, misclassified original samples, and misleading adversarial samples, respectively. Then, we can express

$$\text{PA} \approx \Xi_{\text{SAM}} + \Xi_{\text{ADV}} = \Xi_{\text{ERR}} + \Xi_{\text{MIS}} + \Xi_{\text{ADV}}$$

where

$$\begin{aligned}\Xi_{\text{ERR}} &= N \text{AFR}_T^0 \text{AFR}_P \log(1 - 2\delta_{\text{ERR}}), \\ \Xi_{\text{MIS}} &= N(1 - \text{AFR}_T^0) \text{AFR}_P \log(1 - 2\delta_{\text{MIS}}), \\ \Xi_{\text{ADV}} &= N \text{AFR}_T^0 (1 - \text{AFR}_P) \log \delta_{\text{ADV}},\end{aligned}$$

where  $\text{AFR}_T^0$  is the accuracy of the model in the original data and  $\text{AFR}_P \equiv \text{AFR}(\text{P})$ . Variables  $\delta_{\text{ERR}}$ ,  $\delta_{\text{MIS}}$  and  $\delta_{\text{ADV}}$  account for the average probability assigned to classes other than the predicted class for the three aforementioned cases (see illustration in Figure B.9).  $\Xi_{\text{SAM}}$  aggregates the first two terms and will be interpreted as the sampling randomness contribution.

*Proof.* See Appendix B.2. □

Figures B.10 and B.11 compare the true and approximated PA values under increasing adversarial ratio for PGD and FMN attacks, respectively. It is clear that penalizations are overestimated, given that the average posterior probability was used and differences by defect are more significant than those by excess due to the nonlinear nature of the logarithm in the range  $[0, 1]$ . Besides,  $\beta^*$  is fixed to its lowest possible value; that is, when  $\text{AR} = 1$ . This makes the approximation on the FMN attack less reliable for smaller adversarial ratio settings, as  $\beta^*$  decreases significantly due to the effectiveness of the attack.

Nevertheless, the relative differences in the approximated PA values are consistent with the true values, and the ranking of the models is largely preserved across different adversarial ratios, especially for  $\text{AR} = 1$ . For that reason, the interpretability provided by the approximated PA expression will illustrate the results, and will be used to better characterize the source of robust and unrobust behaviour observed in the different models.

#### 4.2.1 Adversarial robustness assessment with PA

The first results presented correspond to PGD attacks with different attack power  $\ell_\infty$ , namely 8/255, 16/255 and 32/255, for increasing ratio of perturbed samples in the CIFAR10 dataset.

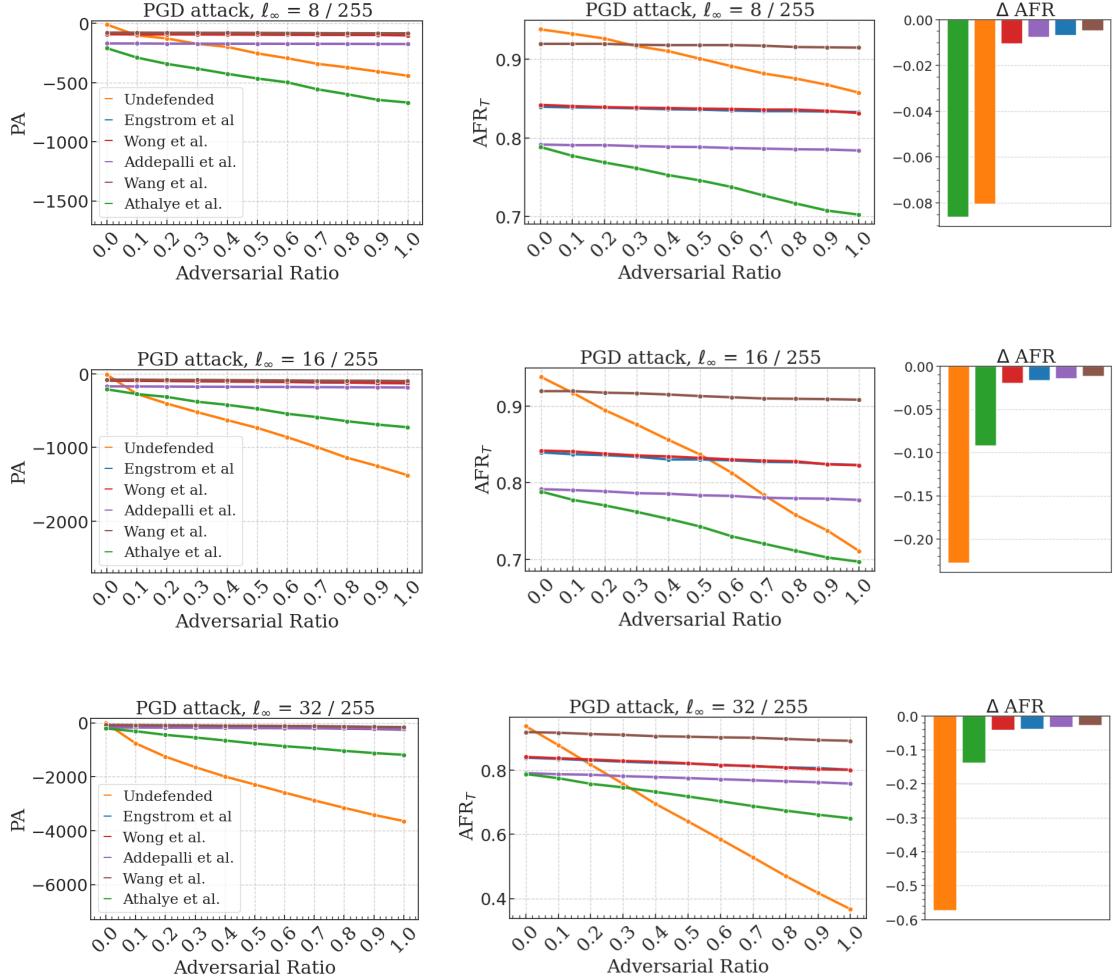


Figure 4.7: PA, AFR(T) and the AFR variation against increasing adversarial ratio at different perturbation norm bounds. The aforementioned undefended net and several RobustBench robust models are considered under a 1000 step PGD attack. When AR = 0, PA → 0 as  $\beta \rightarrow \infty$  in all cases, but the convergence rate depends on the prediction confidence (see Figure 4.2) and thus yields an assessment equivalent to that of AFR (T).

At first glance, it is clear that PA is able to discriminate robust models from the **Undefended** one, which is shown to significantly decrease its performance with increasing adversarial ratio and attack power. As expected, the rate at which its performance decreases is higher the more powerful the attack is, since a greater percentage of samples are able to mislead its predictions.

From both PA and AFR stems the fact that **Athalye et al.** is significantly less robust to PGD attacks than its RobustBench counterparts, as its performance decreases way more significantly with increasing AR. We observe that the rate at which its performance decreases is inversely proportional to the attack power, which is consistent with the behaviour expected from a JPEG compression defense mechanism as it is less effective for small perturbations [18].

A fundamental difference between these two models, which cannot be inferred from a purely performance-based metric, is the nature of the misalignment in the probabilistic output of the model, which is the source of the robust and non-robust behaviour observed. Figure 4.8 (right) shows the optimal  $\beta^*$  value for each model, which follows the entropy of the posterior distribution

and discriminates the two non-robust models from the rest and from each other. The **Undefended** model provides overconfident predictions that maximize disagreement in misleading and misclassified samples, whereas **Athalye et al.** provides uncertain predictions that minimize disagreement in adversarial samples but have the opposite effect in correctly classified ones.

Following the interpretation outlined in Chapter 2, we show that  $\beta^*$  effectively measures the quality of the informativeness estimation made by each model. The **Undefended** model overestimates the information content of the features it has learned, and thus overfits to these and is unable to generalize under adversarial perturbations. On the other hand, **Athalye et al.** underestimates the information content and thus provides overly uncertain predictions. The maximum mutual information between posteriors is lower than its robust counterparts for both models, the former because information is not mutual and the latter because there is less information. This insight clarifies the unintuitive behaviour observed earlier, by which **Athalye et al.** robustness value decreases at a lower rate with increasing attack power, despite maintaining a constant decrease in performance of  $\Delta \text{AFR} \sim 0.1$ .

#### EXPLAIN BETTER THE OPTIMIZATION OF PA. LOOK PAPER.

Figure 4.8 (**left**) illustrates the previous reasoning by displaying the average posterior probability assigned to the predicted class by each model, conditioned on the type of prediction assigned. This discrimination yields three groups of observations, namely original samples that are correctly classified by the robust model, original samples that are misclassified, and perturbed samples that, having their associated unperturbed sample been correctly classified, have been able to mislead the model. These three cases are relevant from the adversarial robustness perspective, as they illustrate the trade-off between high-confident original predictions and adversarial vulnerability, which has been already stated in previous chapters. **Wang et al.** acts as a reference for an ideal robust behaviour, in which original samples are predicted with high confidence and adversarially misleading predicted labels are only slightly more likely than the rest. Equivalent representations for the remaining models can be found in Figure B.12.

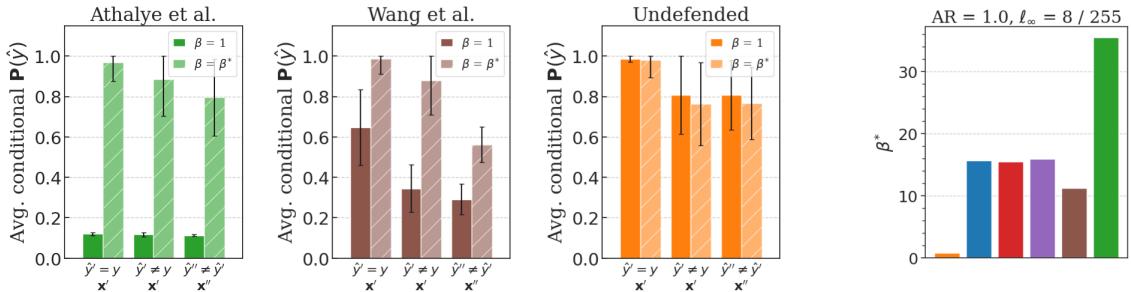


Figure 4.8: (**left**) Average posterior probability of the predicted class for correctly classified original samples, misclassified original samples, and misleading adversarial samples, respectively. (**right**) Optimal  $\beta^*$  value for each model. Results obtained through a PGD attack with  $\ell_\infty = 8/255$ .

With respect to robust models, we observe a significant difference in the discriminative power of PA and accuracy-based metrics that does not immediately derive from the informativeness of the optimal posterior. As remarked before, AFR (P) constitutes our baseline robustness metric, as by definition represents the ratio of predictions that remained constant under adversarial perturbations, and therefore ranks models by their predictive capabilities against these attacks. The value of  $\Delta \text{AFR}$  aligns with that definition, and discriminates robust models by a very thin margin, selecting **Wang et al.** as the best. Further analysis on PA is needed to understand the source of this discrepancy, as for instance why **Adddepalli et al.** model is attributed a significantly lower

value than the remaining robust models under a  $\ell_\infty = 8/255$  PGD attack, despite displaying a similar decrease in performance.

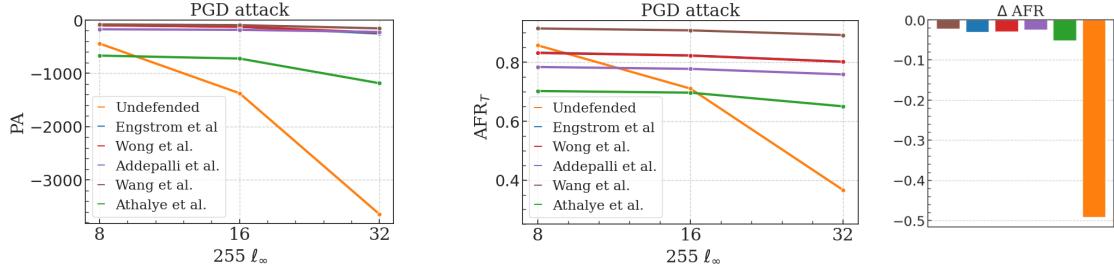


Figure 4.9: PA, AFR(T) and the AFR variation against increasing attack power for AR = 1. The aforementioned undefended net and several RobustBench robust models are considered under a 1000 step PGD attack.

Finally, Figure 4.9 shows that PA is also discriminative with respect to increasing attack power, expressed through the maximum allowed  $\ell_\infty$  norm. As mentioned earlier, PA values are heavily aligned with the performance decrease of the models under a specific attack power, but the observed decrease in PA under increasing  $\ell_\infty$  is much more significant than the decrease in performance. This can be explained by the fact that the metric is sensitive to the overall posterior shift and not only the position of the maximum. When increasing the attack power, confidence in the predicted class will decrease in general, even when the sample does not succeed at misleading the model, and therefore the overall overlap between posteriors will be reduced even at comparable performance levels. This observation further illustrates the independent discriminability power offered by PA (see Properties 3.2), which constitutes the cornerstone argument of this work.

In order to widen the scope of the analysis, analogous results will be obtained for FMN attacks, which are expected to be more effective than PGD attacks for being unbounded, which translates into an overall decrease in  $\beta^*$  (see Table 4.2). Figure 4.10 shows the evolution of PA against increasing adversarial ratio for the same models, and compares it with the assessment provided by AFR.

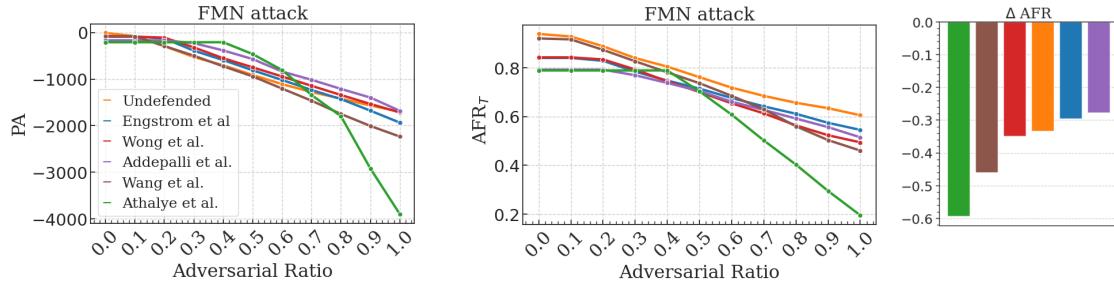


Figure 4.10: PA, AFR(T) and the AFR variation against increasing adversarial ratio. The aforementioned undefended net and several RobustBench robust models are considered under a 1000 step FMN attack.

As expected, the effectiveness of the FMN attack is superior to that of PGD attacks, as the decrease in performance is substantially more significant for all models, especially the ones previously considered robust. In this case, the assessment provided by PA is It is likely that these models have been defended with a compression strategy that succeeds at filtering out small perturbations, which are the ones employed by FMN, and for that reason main-

tain their performance at low adversarial ratio values [18]. In particular, **Athalye et al.** remains maximally robust until at least 40% of the samples are perturbed, at which point the defensive strategy is neutralized and a constant fraction of the additional perturbed samples succeeds at misleading the model, which translates into a linear decrease in performance and PA.

PA proves to be very discriminative among robust models and to represent the phase transition entailed by the collapse of the defense strategy better than AFR does, which can be observed in more detail in Figure B.8. A significative result is that PA is not so directly aligned with  $\Delta\text{AFR}$ , in contrast to the PGD case, which shows again that the decrease in performance is not the main driver of the robustness assessment provided by PA, but instead can be interpreted as a consequence of a misalignment in the posterior distributions of adversarial samples, which are the ones driving the metric after the AR thresold is reached.

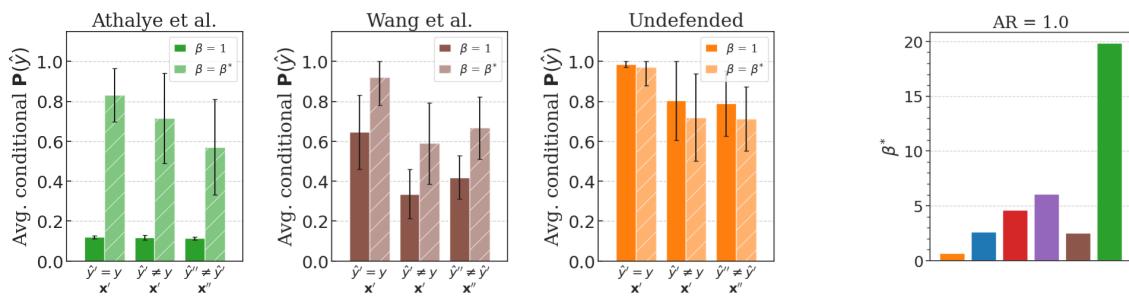


Figure 4.11: (left) Average posterior probability of the predicted class for correctly classified original samples, misclassified original samples, and misleading adversarial samples, respectively. (right) Optimal  $\beta^*$  value for each model. Results obtained through a FMN attack.

Figure 4.11 gives insight into the probabilistic output of the model and the informativeness of the optimal posterior for the **Undefended**, **Athalye et al.** and **Wang et al.** models, in analogous way to the PGD experiments. The first two models display a very similar behaviour for  $\beta = 1$ , but optimal posteriors are less informative due to the increased number of misleading samples, which translates into a smaller  $\beta^*$ . The response of the **Wang et al.** model further illustrates the higher effectiveness of FMN attacks, as adversarial perturbations are on average more misleading than outlier samples in the original dataset, which did not occur in the PGD case. Analogous representations for the remaining models can be found in Figure B.13, which show that the **Addepalli et al.** is the only robust model that maintains the same behaviour under both attacks.

Overall, we recognize that PA has a higher discriminative power than AFR, especially considering the evolution of each metric over increasing adversarial ratio, as seen in Figures 4.7 and 4.10. In particular, Figures B.7 and B.8 compare the evolution of PA with that of AFR (P), which is the baseline metric for robustness, and show the susceptibility of the latter to dataset variability. This is an important consideration, as PA not only improves the discriminability in terms of the scale of the differences between models, but also provides a more stable assessment across varying levels of perturbed sample presence, under which AFR (P) exhibits significant fluctuations that alter the ranking of the models at every step, as illustrated in Table 4.3.

These observations lead to the conclusion that PA offers a more reliable assessment of adversarial robustness, which in general aligns with the decrease in performance on perturbed samples, but that relies heavily on the informativeness of the posterior distribution and the confidence in the predictions for both original and perturbed samples.

PGD	AR = 0.2			AR = 0.4			AR = 0.6		
	PA	AFR <sub>P</sub>	AFR <sub>T</sub>	PA	AFR <sub>P</sub>	AFR <sub>T</sub>	PA	AFR <sub>P</sub>	AFR <sub>T</sub>
<b>Addepalli et al.</b>	<b>-172.5</b>	<b>0.995</b>	<b>0.785</b>	<b>-175.5</b>	0.992	<b>0.786</b>	<b>-177.6</b>	0.989	<b>0.783</b>
<b>Wong et al.</b>	-97.7	0.996	0.838	-102.9	0.992	0.834	-109.2	<b>0.987</b>	0.830
<b>Engstrom et al.</b>	-94.2	0.996	0.836	-104.6	<b>0.990</b>	0.830	-110.3	0.988	0.829
<b>Wang et al.</b>	-81.9	0.997	0.917	-84.6	0.996	0.915	-89.4	0.991	0.912
FMN	PA	AFR <sub>P</sub>	AFR <sub>T</sub>	PA	AFR <sub>P</sub>	AFR <sub>T</sub>	PA	AFR <sub>P</sub>	AFR <sub>T</sub>
<b>Addepalli et al.</b>	-169.4	1.0	<b>0.791</b>	-385.4	0.944	<b>0.737</b>	-838.9	0.867	0.660
<b>Wong et al.</b>	-111.2	0.991	0.834	-553.1	0.901	0.743	-944.8	0.810	<b>0.653</b>
<b>Engstrom et al.</b>	-128.5	0.988	0.828	-592.9	0.907	0.747	-1020	0.836	0.675
<b>Wang et al.</b>	<b>-291.6</b>	<b>0.952</b>	0.873	<b>-726.8</b>	<b>0.861</b>	0.781	<b>-1204</b>	<b>0.764</b>	0.684

Table 4.3: Comparison of PA, AFR<sub>P</sub> and AFR<sub>T</sub> scores for a PGD attack with  $\ell_\infty = 16 / 255$  and an FMN attack across different adversarial ratio values. The worst robustness score is emboldened for every case. PA displays higher consistency and discriminative power across varying AR with respect to accuracy-based metrics.

	AR = 0.1			AR = 0.5		
	PA	AFR <sub>P</sub>	AFR <sub>T</sub>	PA	AFR <sub>P</sub>	AFR <sub>T</sub>
PGD, $\ell_\infty = 8 / 255$	<b>1.0</b>	0.0	0.66	<b>1.0</b>	-0.66	0.66
PGD, $\ell_\infty = 16 / 255$	<b>0.66</b>	-0.33	0.66	0.66	<b>1.0</b>	0.66
PGD, $\ell_\infty = 32 / 255$	<b>0.66</b>	-0.33	0.66	0.66	<b>1.0</b>	0.66
FMN	0.33	<b>0.66</b>	0.0	<b>1.0</b>	1.0	-0.66

Table 4.4: Kendall rank correlation coefficient comparing the consistency of the model ranking provided by each metric at different adversarial ratio values. The ranking at a specific AR is compared with that with AR = 1.0. PA is shown to display the highest consistency, especially under low attack power settings.

#### 4.2.2 Interpretability of PA in the adversarial setting

In light of the results obtained, the suitability of PA in the adversarial setting has been demonstrated, but a deeper exploration of the reason of the discrepancies between PA and the baseline robustness measures is needed so that it can confidently be established as a model selection criterion. In particular, we will work with the approximated expression of PA derived in Theorem 4.2.1 and elucidate the source of the measured robust behaviour.

Table 4.5 shows the contribution of each subset of observations to the final approximated PA value for a PGD attack.  $N_{\text{ERR}}$ ,  $N_{\text{MIS}}$  and  $N_{\text{ADV}}$  are the number of (pairs of) contributing samples, and  $\Xi_{\text{ERR}}$ ,  $\Xi_{\text{MIS}}$  and  $\Xi_{\text{ADV}}$  are the total amount of the contribution. For reasons described earlier in this section, the PA approximation overestimates penalizations when compared to the true value, but relative discrepancies between models are still largely preserved and therefore the rationale behind the discriminative power of PA, as shown in Figures B.10 and B.11. The parameters  $2\delta_{\text{MIS}}$  and  $\delta_{\text{ADV}}$  account for the average probability assigned to classes other than the predicted class for misclassified original samples and misleading adversarial samples, respectively, and help interpret the informativeness of the distribution as well as the value of each individual penalization.

For instance, a large  $2\delta_{\text{MIS}}$  value indicates robustness to sampling randomness, as it represents higher average uncertainty in misclassified predictions. A model with a high performance on test data entails a more negative penalization  $\log(1 - 2\delta_{\text{MIS}})$ , for being misclassified samples more likely to be equivalently misclassified under adversarial perturbations, but at the same time makes misclassifications less likely, and therefore the number of terms added to  $\Xi_{\text{MIS}}$ . The existing trade-off

between standard and robust generalization arises when following this reasoning towards the minimization of  $\Xi_{\text{MIS}}$ , because reducing the number of misclassified samples will drive  $\beta^*$  to higher values and therefore decrease adversarial uncertainty  $\delta_{\text{ADV}}$ . As outlined before,  $\delta_{\text{ADV}}$  indicates robustness to adversarial perturbations, as it represents the average prediction uncertainty on adversarial misleading samples, and entails a penalization of  $\log(\delta_{\text{ADV}})$ .

The interpretation of these terms is vitally important for the purpose of this work, as it enables the identification of the different sources of robustness displayed by each model, and therefore the characterization of the randomness that we will demand models to generalize to. From a general perspective,  $\Xi_{\text{SAM}} = \Xi_{\text{ERR}} + \Xi_{\text{MIS}}$  can be understood as the lack of robustness to sampling randomness, and  $\Xi_{\text{ADV}}$  as the lack of robustness to adversarial perturbations.

Defense	$N_{\text{MIS}}$	$2\delta_{\text{MIS}}$	$\Xi_{\text{SAM}}$	$N_{\text{ADV}}$	$\delta_{\text{ADV}}$	$\Xi_{\text{ADV}}$
<b>Wang et al.</b>	799	0.24	-468.62	47	0.44	-39.44
<b>Engstrom et al.</b>	1591	0.17	-566.72	67	0.39	-63.43
<b>Wong et al.</b>	1562	0.17	-537.25	90	0.38	-88.98
<b>Addepalli et al.</b>	2063	0.21	-877.42	75	0.46	-58.92
<b>Undefended</b>	566	0.47	-736.63	810	0.24	-1173.55
<b>Athalye et al.</b>	1915	0.23	-963.85	747	0.21	-1183.96

Table 4.5: Approximated PA contributions for a PGD attack with  $\ell_\infty = 8/255$  and AR = 1.0. The number of originally misclassified and adversarially misleading samples is  $N_{\text{MIS}} = \lfloor N(1 - \text{AFR}_T^0) \text{AFR}_P \rfloor$  and  $N_{\text{ADV}} = \lfloor N \text{AFR}_T^0(1 - \text{AFR}_P) \rfloor$ , respectively. The penalization argument  $2\delta_{\text{ERR}}$  has not been included for being negligible in all cases.

As expected, the standard generalization error term  $\Xi_{\text{SAM}}$  is the one contributing most to the PA measure in robust models, as the selected PGD attack is not very effective and can only generate a small number of misleading samples  $N_{\text{ADV}}$ . The discrimination of models based exclusively on  $\Xi_{\text{SAM}}$  is very much aligned with that of AFR (T) in all cases with the exception of the **Undefended** model, which is penalized more heavily for providing overconfident predictions with a large number of misleading examples  $N_{\text{ADV}}$  and thus converging to a small  $\beta^*$ . This is an important realization, as it shows that even if standard and adversarial robustness contributions can be dissociated, they are mutually dependent and ultimately derive from the overall agreement in all predictions, regardless of the nature of the randomness they are bound to. The generalization error to sampling randomness will be exceedingly penalized the less robust a model is to other sources of randomness, because the optimal resolution of the hypothesis space is reduced and the less distinction can be made between adversarial samples and outliers from the original dataset. Further insights into this reasoning can be obtained by comparing these results with those of the **Athalye et al.** model, which has a similar accuracy on adversarial samples and a significantly worse accuracy on original samples. The fact that posterior distributions are profoundly uninformative increases agreement in between mismatching posterior and thus lowers penalization terms, even if more terms will be added as a consequence of the associated decrease in performance.

The same reasoning can be followed to explain the discrimination made by PA between **Addepalli et al.** and the other robust models. **Addepalli et al.** experiences a comparable drop in performance, and for displaying a reduced confidence in mismatching predictions is assigned a smaller  $\Xi_{\text{SAM}}$  contribution than some of these models. Nevertheless, such uncertainty is also observed for original samples, which lowers accuracy on the original dataset and thus increases random sampling penalization  $\Xi_{\text{SAM}}$ . In that sense, it can be argued that **Addepalli et al.** is more robust than **Wong et al.** and **Engstrom et al.** to adversarial perturbations, which also stems from the baseline AFR (P) and  $\Delta\text{AFR}$  values, but significantly less robust to sampling randomness. PA weights both contributions and yields an intermediate model selection criterion.

Regarding adversarial robustness, we observe that  $\Xi_{\text{ADV}}$  is driven by the decrease in performance under attack  $\Delta\text{AFR}$ , as  $N_{\text{ADV}}$  penalization terms are added. Nevertheless, the value of each of these terms is  $\log(\delta_{\text{ADV}})$ , which penalizes models that achieve maximum posterior agreement by increasing confidence on adversarially misleading examples. This is a clear distinctive trait with respect to accuracy-based metrics, whose penalizations are reduced to a binary decision.

Defense	$N_{\text{MIS}}$	$2\delta_{\text{MIS}}$	$\Xi_{\text{SAM}}$	$N_{\text{ADV}}$	$\delta_{\text{ADV}}$	$\Xi_{\text{ADV}}$
<b>Addepalli et al.</b>	1507	0.52	-1910.69	2187	0.28	-2788.89
<b>Wong et al.</b>	1032	0.46	-1125.53	2920	0.27	-3844.40
<b>Engstrom et al.</b>	1125	0.72	-2469.65	2505	0.32	-2847.99
<b>Wang et al.</b>	435	0.82	-1599.08	4215	0.33	-4637.45
<b>Undefended</b>	412	0.56	-704.58	3132	0.29	-3906.55
<b>Athalye et al.</b>	859	0.57	-2054.23	4679	0.43	-3955.43

Table 4.6: Approximated PA contributions for a FMN attack with  $\text{AR} = 1.0$ . The number of originally misclassified and adversarially misleading samples is  $N_{\text{MIS}} = \lfloor N(1 - \text{AFR}_T^0) \text{AFR}_P \rfloor$  and  $N_{\text{ADV}} = \lfloor N \text{AFR}_T^0(1 - \text{AFR}_P) \rfloor$ , respectively. The penalization argument  $2\delta_{\text{ERR}}$  has not been included for being negligible in all cases with the exception of the **Athalye et al.** model, which amounts to 0.36.

In contrast with the PGD case, the FMN attack is much more effective and  $N_{\text{ADV}} > N_{\text{MIS}}$  in all cases, which makes the adversarial contribution  $\Xi_{\text{ADV}}$  more relevant in the overall robustness assessment. The discrepancy observed between PA and accuracy-based metrics for the **Wong et al.** model can also be explained in these terms, as it is the least penalized by  $\Xi_{\text{SAM}}$  among robust models due to its superior accuracy. In the context of effective attacks, predictive certainty on original samples is highly rewarded because lower overall agreement makes sampling penalization terms  $\log(1 - 2\delta_{\text{MIS}})$  more negative.

Overall, this approximation shows that the final PA value stems from a combination of standard and adversarial generalization error, which are normally obtained independently through different accuracy-based metrics, and leads to the realization that PA provides an intermediate assessment in between AFR measures. An analogous combined metric weighting accuracy and  $\Delta\text{AFR}$  would not be equiparable to PA, given that these weights would be arbitrary and would not adjust to the particularities of each model. For instance, the **Undefended** model should be mostly penalized on the basis of its robustness to adversarial examples, whereas **Addepalli et al.** model should be mostly penalized for its lack of robustness to sampling randomness in the original data. These considerations are fundamental in the covariate shift setting, as different models with different defensive or invariant feature learning strategies will navigate the generalization-complexity trade-off in a different way.

As a conclusion to our analysis, other kinds of metrics have been contrasted to determine whether they are able to provide a reliable assessment of adversarial and sampling robustness that is comparable to the baseline accuracy-based measures. These include confidence-based measurements, such as the Kullback-Leibler (KL) divergence and the Wasserstein distance between posterior distributions, and typical distance measures on the feature space of the model, namely through dataset cosine similarity, distance between centroids, and other cluster-based approaches such as the maximum mean discrepancy or the Fréchet inception distance (FID). This last one is particularly interesting, as it is a widely used metric in the generative adversarial network literature and thus fits intuitively well into our setting.

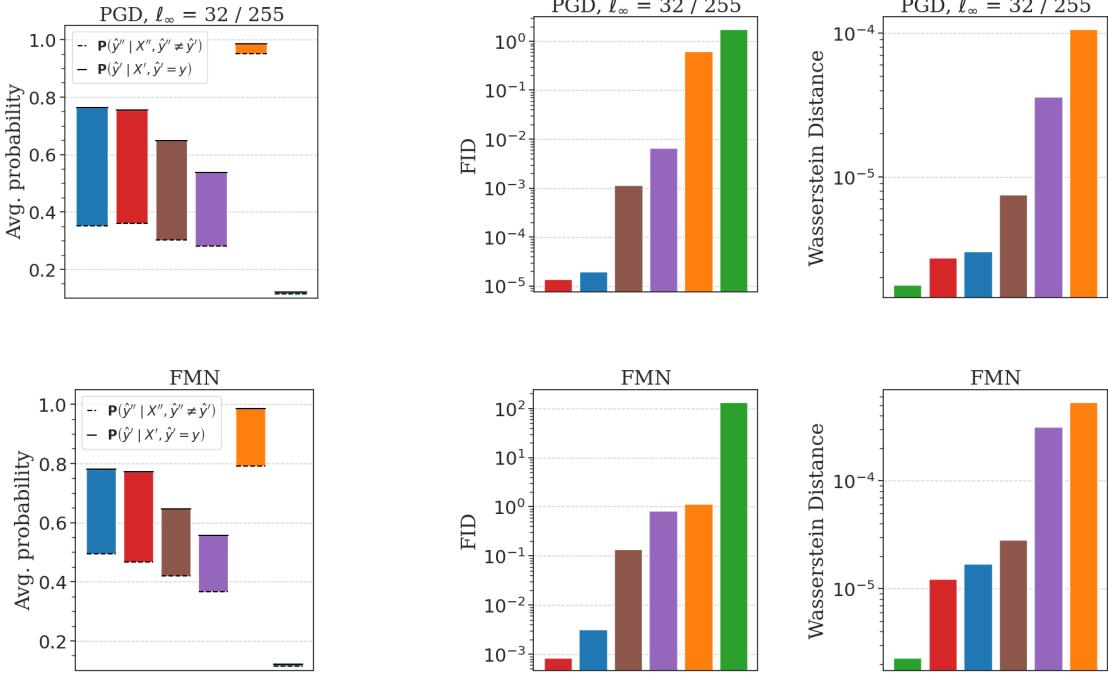


Figure 4.12: The two metrics considered are FID, which amounts to the group-based dissimilarity in the feature space, and Wasserstein distance, which measures the average distance between probability distributions.

Figure 4.12 displays some of these metrics for PGD and FMN attacks, together with a condensed visualization of the average probability gap existing amidst predictions on original and adversarial examples. As we can see, the discriminative power of these metrics stems from the overall difference between posteriors at  $\beta = 1$ . This result is not surprising, since the discriminant function of the classifier is constituted by a small subset of layers in the network, and is not expected to represent a complex discrimination that cannot be inferred from the distribution of samples in the feature space. As expected, probability-based metrics tend to coincide in their assessment in general terms, and feature-space-based metrics as well. Besides, it is also not surprising that feature space distances are more sensitive to the success rate of the attack, given that the inductive bias remains constant under any condition, whereas confidence-based metrics are susceptible to the overall shift in the posterior, which in general favours non-informative posteriors. A deeper insight into the evolution of these metrics for PGD and FMN attacks under increasing adversarial ratio can be found in Figures B.14 and B.15.

The results obtained in this section showcase the rationale behind the maximization of posterior agreement, which can be thought of a surrogate version of the maximization of the mutual information between original and perturbed datasets [9]. From an information-theoretic perspective, the maximization of mutual information effectively distillates the information contained in the posterior distributions to that which is relevant for robustness assessment. This property is the source of the discriminative power displayed by PA and constitutes a fundamental difference with respect to the baseline accuracy-based metrics considered.

### 4.3 Out-of-distribution setting

Following the analysis conducted in the previous section, the discriminability of PA will be now explored in the domain generalization setting, which is a priori more convenient for PA for being

accuracy-based metrics less informative in this context. This is because we are ultimately assessing the quality of the inductive bias of the model by its ability to generalize to target (i.e. unseen) domains. In this sense, the additional insight and discriminability exhibited by PA is expected to be more relevant for the selection of models that perform well not only on unseen data, but also on unseen data that shares limited features with the training data. Under these conditions, the overlap between posteriors is more informative than simply matching predictions (e.g. AFR<sub>P</sub>), because significant disagreement in the remaining classes indicates vulnerability to distribution shifts present in the source domains, which implies vulnerability to target domains as well. This is a fundamental difference with respect to the adversarial setting, in which the nature of the perturbation made posteriors less relevant for the robustness assessment.

This section will not address epoch-wise model selection, but will focus instead on the evaluation of the generalization capabilities of different learning algorithms under increasing levels of distribution shift, by computing the posterior agreement between source environments for models achieving maximum validation accuracy. More specifically, a baseline vanilla ERM algorithm will be used to train a ResNet18 model and will be compared with two robust learners, namely invariant risk minimization (IRM) and selective augmentation (LISA), both introduced in Section 3.5. Results should elucidate whether PA is able to discern datasets subjected to different levels of domain shift and whether models achieving highest PA scores perform better on new domains.

Experiments will be performed by means of the DiagViB-6 dataset framework [20], which comprises MNIST images of size 128x128 within an augmentation pipeline enabling the modification of six specific image factors: shape, hue, lightness, position, scale and texture. Several variations in the `diagvibsix` library<sup>1</sup> have been implemented with the purpose of this project so that datasets can be built with a specific configuration of factors for each sample, which allows for a wide range of experiments in the data shift assessment and model selection settings. In an analogous way to the adversarial case, datasets will be incrementally perturbed by including only a fraction of the shifted samples, which in this context we will call shift ratio (SR).

Following the notation introduced in Section 3.4, Definition 4.3 provides a characterization of source and target domains and the randomness entailed by each dataset. The control over these aspects is the rationale behind this experimental setup, since it is through synthetic image manipulation that we can maximize invariant feature learning possibilities during training while providing optimal robustness assessment conditions in validation and testing. Since changes in image factors can be independently introduced to each sample, the shifted dataset contains the same samples and in the same order as the original dataset, thus removing sampling randomness contributions from the robustness score. Table 4.7 stipulates the specific factors conforming each environment in this experiment and Figure 4.13 illustrates them with some examples.

CITE MORE INFO IN C

**Definition** (Shifted factors experiment). The classification task involves the prediction of the shape factor (i.e. the digit) of handwritten fours and nines from the MNIST dataset. In particular, source and target domains are generated as follows:

$$\begin{aligned}\mathcal{S} &= \{X_0, X_1\}, \\ \mathcal{T} &= \{X_1, X_2, X_3, X_4, X_5\},\end{aligned}$$

where  $X_j$  represents the random variable associated to domain  $j$ , being  $j$  the number of shifted factors with respect to domain  $X_0$ . Datasets are generated by considering four different realizations of the experiment, namely  $\tau_0^{\text{train}}$ ,  $\tau_1^{\text{train}}$ ,  $\tau^{\text{val}}$  and  $\tau^{\text{test}}$ , each sampling from disjoint subsets of

---

<sup>1</sup><https://github.com/victorjimenezzz/diagvibsix/tree/librarization>

MNIST. Following the notation introduced in Chapter 3, we can define:

$$\begin{aligned} D^{\text{train}} &= \{\mathbf{x}_0^{\text{train}}, \mathbf{x}_1^{\text{train}}\}, \text{ where } \mathbf{x}_j := \mathbf{x}_j \circ \tau_j^{\text{train}}, j = 0, 1 \\ D^{\text{val}} &= \{\mathbf{x}_0^{\text{val}}, \mathbf{x}_1^{\text{val}}\}, \text{ where } \mathbf{x}_j := \mathbf{x}_j \circ \tau_j^{\text{val}}, j = 0, 1 \\ D^{\text{test}} &= \{\mathbf{x}_j^{\text{test}}\}, \text{ where } \mathbf{x}_j^{\text{test}} := \mathbf{x}_j^{\text{test}} \circ \tau_j^{\text{test}}, j = 1, \dots, 5 \end{aligned}$$

In this way, only training data is subject to both sampling randomness ( $\tau_0^{\text{train}} \neq \tau_1^{\text{train}}$ ) and domain shift ( $X_0 \not\sim X_1$ ), emulating the conditions of real-world sampling experiments. In contrast, validation and testing datasets entail each a single noise instantiation, which means that distribution shift is the only accountable source of randomness. Overall, two sets of 40 000 images for training, two sets of 20 000 images for validation, and six sets of 10 000 images for testing are generated.

# Shift Factors	0	1	2	3	4	5
Hue	red	<b>blue</b>	blue	blue	blue	blue
Lightness	dark	dark	<b>bright</b>	bright	bright	bright
Position	CC	CC	CC	<b>LC</b>	LC	LC
Scale	normal	normal	normal	normal	<b>large</b>	large
Texture	blank	blank	blank	blank	blank	<b>tiles</b>
<i>Shape</i>	4,9	4,9	4,9	4,9	4,9	4,9

Table 4.7: Image factors associated to each of the environments considered in this experiment. CC and LC account for 'centered center' and 'centered low', respectively.

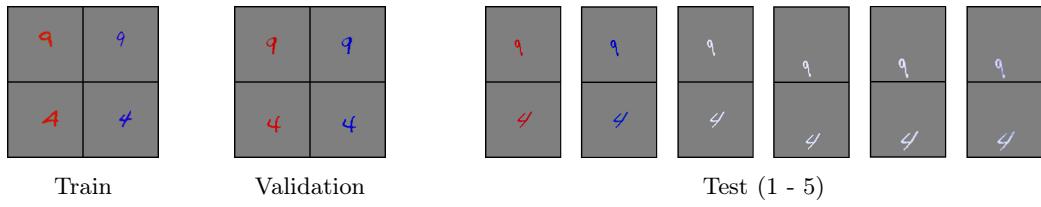


Figure 4.13: Illustration of the training, validation and test datasets. Samples for each training environment belong to different MNIST subsets, whereas samples of validation and test are corresponding.

Results obtained in this setup show that PA succeeds at discriminating the different models by their predictive response under increasing number of shifted samples and under increasing shift power. In particular, ERM can be identified to be non-robust by the fact that its score is maximum for the first shifted factor, but decays rapidly to the minimum value after the second factor. In contrast, IRM and LISA show a reduced rate of decay and even display a slight increase in performance for the last shift factor.

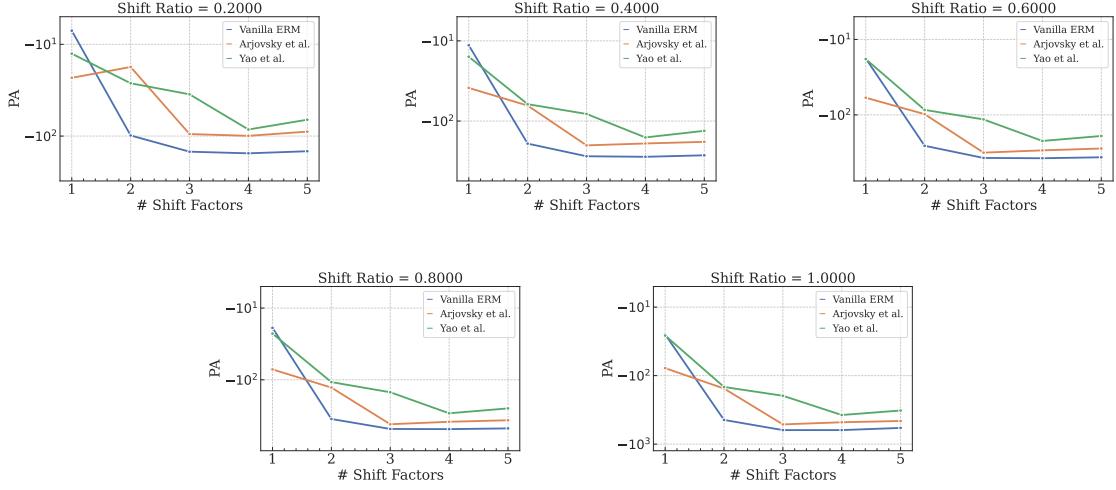


Figure 4.14: Evolution of PA under increasing levels of shift power. Weights maximizing validation accuracy were selected for ERM, IRM and LISA algorithms. Results for incremental presence of shifted samples indicate that PA is able to differentiate weak and robust models.

The unexpected increase in PA after four shifted factors seems inconsistent with the alleged non-increasing behaviour of PA, as per Section 3.2. Nevertheless, this phenomenon only highlights that robustness does not stem from the data generation process but instead from the latent representation of the model and the features selected for the construction of its inductive bias, as discussed in the introductory chapter. In this particular case, Table 4.8 shows that there is a clear discontinuity in the feature representation of the data when the texture factor is shifted, which results in a different discriminator function that ultimately leads to a predictive behaviour that aligns slightly better with the original predictions.

# Shift Factors	1	2	3	4	5
<b>Vanilla ERM</b>	0.9978	0.9303	0.9562	0.9561	<b>0.6661</b>
<b>Arjovsky et al.</b>	0.9967	0.9018	0.9296	0.9374	<b>0.5585</b>
<b>Yao et al.</b>	0.9980	0.9431	0.9431	0.9641	<b>0.7130</b>

Table 4.8: Pairwise cosine similarity between feature space representations of original and augmented images, for each of the shifted datasets. The abrupt decrease in similarity for the fifth environment indicates a discontinuity in the feature representation of images, which leads to non-comparable predictive outcomes.

	1 Shifted Factor			3 Shifted Factors			5 Shifted Factors		
	PA	AFR <sub>P</sub>	AFR <sub>T</sub>	PA	AFR <sub>P</sub>	AFR <sub>T</sub>	PA	AFR <sub>P</sub>	AFR <sub>T</sub>
<b>Vanilla ERM</b>	<b>-24.91</b>	0.999	0.993	-625.6	0.979	0.975	-579.38	0.976	0.873
<b>Arjovsky et al.</b>	-76.76	0.998	0.993	-514.2	0.976	0.978	-464.2	0.976	0.911
<b>Yao et al.</b>	-26.21	<b>0.999</b>	<b>0.994</b>	<b>-201.2</b>	<b>0.985</b>	<b>0.980</b>	<b>-324.4</b>	<b>0.988</b>	<b>0.945</b>

Table 4.9: Comparison of PA, AFR<sub>P</sub> and AFR<sub>T</sub> scores for ERM, IRM and LISA learning algorithms under different levels of shift power. The highest robustness score is emboldened for every case. PA is able to discriminate algorithms consistently and distinguish the first shifted factor, which is seen by the model during training, from the rest.



# Chapter 5

## Model selection

Chapter 4 explored the robustness assessment capabilities of the PA kernel in image classification tasks, and provided extensive evidence of its suitability as an algorithm selection criterion in covariate shift settings. This chapter extends our previous findings by investigating how the PA kernel can be leveraged for robust epoch-wise model selection with early stopping, potentially mitigating overfitting and enhancing generalization performance under distribution shifts.

### 5.1 Model selection under controlled experimental conditions

Building upon the exploratory results on the domain generalization setting previously obtained, we will assess the robust model selection capabilities of PA across a wide range of distribution shift settings in a controlled experimental setup. In particular, different shift factors will be considered for the source environments and also different learning targets. These findings will help identify the experimental conditions in which the discriminative rationale of PA is most effective.

Experiments have been conducted in a setting similar to that described in Section 4.3, with a reduced dataset size to avoid repetition of MNIST samples. This approach ensures that each training sample uniquely represents a specific instance of the "number drawing experiment", along with the corresponding domain shift perturbation. Given that shift perturbations are not entirely deterministic (see `diagvibsix` implementation), this approach prevents the model's inductive bias from being influenced by an implicit data augmentation process.

Both SGD and Adam optimizers under various learning rate values have been considered, so that the most informative results are displayed. In practice, Adam should be intuitively preferred in this setting, as it navigates the loss landscape in a less continuous way and explores a wider range of feature combinations, thus increasing the likelihood that domain-invariant features are considered for PA assessment. Nevertheless, SGD is more stable and can be used to analyze the convergence of PA across training epochs.

One of the key parameters analyzed in these experiments is the shape factor of the images, which serves as the learning objective for the classification task. First, ERM and IRM learners are trained for a binary classification task involving the digit pair (1, 7), which have close latent representations and thus entail higher variability in learning outcomes. In this setting, optimal posteriors are less informative and thus agreement in the non-predicted class is relevant for the PA penalization, as can be seen in Figure ???. This behaviour is relevant for domain generalization, as models with similar validation accuracy might display different confidence levels in their predictions and thus different generalization capabilities to unseen domains.

These results will be contrasted with those obtained through a 4-class classification involving the digits (1, 7, 4, 9). Given that (1, 7) and (4, 9) pairs are likely to be easily discriminated, the

behavior of validation accuracy should be similar to that of a binary classification task. However, PA is expected to discriminate both experiments and select the weights that better encode a domain-invariant representation by considering the whole posterior, not only the predicted class. An alternative experiment that considers binary classification of the digit pairs themselves was not pursued, as we seek a consistent inductive bias across all the experiments presented in this work, which should be constructed only from the features determining each digit and the implicit contribution of shifted factors.

HERE PLOT OF THE POSTERIORS.

YOU MUST ADD A CONTROL DATASET WITH ONLY RANDOMNESS !!!

The characterization of the inductive bias is outside of the scope of this work, but given the simplicity of the experimental setup and the learning task associated, it is reasonable to assume that it encompasses all the relevant features that are present in the data, including the noise instantiation, the nature of the shift, and their relative frequency in the dataset. The optimization process will iteratively navigate the loss landscape and implicitly balance these features in a different way, leading to different predictive outcomes.

In this regard, will examine two primary sources of inductive bias by varying the nature of the shift defining environments 0 and 1, namely based on the hue factor, as was the case in the previous chapter, and based on the position factor. These represent the two most significant sources of variability from an image representation perspective, and comparing the model selection capabilities across these settings will provide insight into the consistency of the metric.

The last variable to consider is the availability of target domains during validation; that is, for model selection purposes. The domain generalization challenge requires that target domains are inaccessible, which in our case helps select robustness-fostering algorithms from the vanilla ERM. Nevertheless, with the purpose of increasing the characterization of the robustness selection criterion, we will consider different degrees of incremental shift on the validation dataset, which should improve the effectiveness of the selection.

The last variable that will be considered is the availability of target domains during validation; that is, for model selection purposes. The domain generalization setting requires the assumption that target domains are entirely inaccessible, thus discriminating robustness-fostering learners from vanilla ERM. However, with the purpose of increasing the characterization of the robustness selection criterion, we will consider different degrees of incremental shift on the validation dataset, which should improve the effectiveness of the selection. Tables ?? and ?? describe the data composition of the experiments considered.

HERE DATASETS.

NOTE THAT FIRST ENVIRONMENT IS COMMON. TAKE IT INTO ACCOUNT WHEN GIVING RESULTS.

- Results to show: Table with accuracy and F1 of the selected models under all conditions. Maybe compare validation accuracy, etc
- Single plot. x axis is the configuration of the validation and model selection dataset, and y axis is the percentage of increase in performance of the accuracy-selected model With respect to the PA-selected model.

It is important to note that test datasets corresponding to the first shifted factor have the same configuration in both cases, which allows the comparison not only of different robustness configu-

rations but also of algorithms that are trained to perform the same task but are subject to shifts of different nature.

<b>ERM</b>	<b>0</b>		<b>1</b>		<b>2</b>		<b>3</b>		<b>4</b>		<b>5</b>	
	Acc.	$\Delta_{PA}$										
SD	99.4	$\pm 0.01$	68.0	<b>+10.1</b>	45.2	<b>-10.1</b>	42.1	<b>-6.0</b>	42.9	<b>-9.3</b>	32.0	<b>-3.8</b>
ID	<b>99.4</b>	$\pm 0.01$	67.1	<b>-7.9</b>	44.9	<b>+0.8</b>	42.7	<b>+1.3</b>	39.1	<b>+1.5</b>	30.3	<b>+1.4</b>
1F-MD	99.4	$\pm 0.01$	59.3	$\pm 0.01$	46.7	$\pm 0.01$	44.9	$\pm 0.01$	42.5	$\pm 0.01$	32.8	$\pm 0.01$
5F-MD	99.2	$\pm 0.01$	<b>74.5</b>	<b>-13.9</b>	<b>60.7</b>	<b>-19.7</b>	<b>52.9</b>	<b>-10.0</b>	<b>46.0</b>	<b>-3.7</b>	<b>38.1</b>	<b>-1.7</b>
OOD	99.2	$\pm 0.01$	74.5	$\pm 0.01$	60.7	$\pm 0.01$	52.9	$\pm 0.01$	46.0	$\pm 0.01$	38.1	$\pm 0.01$

<b>IRM</b>	<b>0</b>		<b>1</b>		<b>2</b>		<b>3</b>		<b>4</b>		<b>5</b>	
	Acc.	$\Delta_{PA}$										
SD	99.4	<b>-0.2</b>	77.0	<b>+3.2</b>	34.7	<b>+5.5</b>	39.5	<b>+2.3</b>	42.0	<b>+1.9</b>	29.1	<b>+2.4</b>
ID	<b>99.4</b>	<b>-0.2</b>	71.2	<b>+12.6</b>	<b>51.2</b>	<b>-0.8</b>	<b>55.1</b>	<b>+0.1</b>	<b>52.8</b>	<b>-1.5</b>	<b>38.4</b>	<b>-1.3</b>
1F-MD	99.2	$\pm 0.01$	<b>80.2</b>	$\pm 0.01$	40.2	$\pm 0.01$	41.8	$\pm 0.01$	43.9	$\pm 0.01$	31.5	$\pm 0.01$
5F-MD	99.4	<b>-0.7</b>	71.2	<b>-14.8</b>	51.2	<b>+4.8</b>	55.1	<b>-1.8</b>	52.8	<b>-6.8</b>	38.4	<b>-8.1</b>
OOD	99.2	$\pm 0.01$	69.1	$\pm 0.01$	46.3	$\pm 0.01$	49.4	$\pm 0.01$	46.6	$\pm 0.01$	31.8	$\pm 0.01$

Table 5.1: REMOVEopt=adam-lr=0.0001-mf=hue-npair=FalseREMOVE Test performance on increasingly shifted datasets for models selected during ERM and IRM procedures. Different validation datasets are used, and the selection capabilities of PA and validation accuracy are compared.

<b>ERM</b>	<b>0</b>		<b>1</b>		<b>2</b>		<b>3</b>		<b>4</b>		<b>5</b>	
	Acc.	$\Delta_{PA}$										
SD	<b>99.4</b>	<b>-0.1</b>	59.3	<b>+8.3</b>	46.7	<b>+1.5</b>	44.9	<b>+1.5</b>	42.5	<b>+0.9</b>	32.8	<b>-1.5</b>
ID	99.4	<b>-0.1</b>	59.3	<b>+8.3</b>	46.7	<b>+1.5</b>	44.9	<b>+1.5</b>	42.5	<b>+0.9</b>	32.8	<b>-1.5</b>
1F-MD	99.4	$\pm 0.01$	59.3	$\pm 0.01$	46.7	$\pm 0.01$	44.9	$\pm 0.01$	42.5	$\pm 0.01$	32.8	$\pm 0.01$
5F-MD	99.2	$\pm 0.01$	<b>74.5</b>	<b>-13.9</b>	<b>60.7</b>	<b>-19.7</b>	<b>52.9</b>	<b>-10.0</b>	<b>46.0</b>	<b>-3.7</b>	<b>38.1</b>	<b>-1.7</b>
OOD	99.2	$\pm 0.01$	74.5	$\pm 0.01$	60.7	$\pm 0.01$	52.9	$\pm 0.01$	46.0	$\pm 0.01$	38.1	$\pm 0.01$

<b>IRM</b>	<b>0</b>		<b>1</b>		<b>2</b>		<b>3</b>		<b>4</b>		<b>5</b>	
	Acc.	$\Delta_{PA}$										
SD	<b>99.4</b>	$\pm 0.01$	74.4	<b>-4.7</b>	41.4	<b>+14.2</b>	43.8	<b>+14.5</b>	41.6	<b>+14.9</b>	26.9	<b>+11.5</b>
ID	99.3	<b>-0.2</b>	81.9	<b>-12.9</b>	52.4	<b>-6.1</b>	54.5	<b>-5.1</b>	54.1	<b>-7.6</b>	36.5	<b>-4.7</b>
1F-MD	99.2	$\pm 0.01$	80.2	$\pm 0.01$	40.2	$\pm 0.01$	41.8	$\pm 0.01$	43.9	$\pm 0.01$	31.5	$\pm 0.01$
5F-MD	99.4	$\pm 0.01$	<b>84.1</b>	<b>-2.0</b>	<b>54.3</b>	<b>-3.2</b>	<b>56.3</b>	<b>-2.0</b>	<b>54.7</b>	<b>-0.8</b>	<b>38.3</b>	<b>-1.1</b>
OOD	99.2	$\pm 0.01$	69.1	$\pm 0.01$	46.3	$\pm 0.01$	49.4	$\pm 0.01$	46.6	$\pm 0.01$	31.8	$\pm 0.01$

Table 5.2: REMOVEopt=adam-lr=0.0001-mf=hue-npair=TrueREMOVE Test performance on increasingly shifted datasets for models selected during ERM and IRM procedures. Different validation datasets are used, and the selection capabilities of PA and validation accuracy are compared.

<b>ERM</b>	<b>0</b>		<b>1</b>		<b>2</b>		<b>3</b>		<b>4</b>		<b>5</b>	
	Acc.	$\Delta_{PA}$										
ID	99.3	+0.2	99.4	+0.2	86.6	-1.3	<b>68.5</b>	+5.7	50.8	+4.0	47.4	+8.2
1F-MD	<b>99.3</b>	-0.1	<b>99.5</b>	-0.1	84.9	+4.0	66.3	+1.8	52.0	+1.6	46.3	+7.4
5F-MD	99.2	$\pm 0.01$	99.3	+0.1	80.9	+3.8	64.5	-0.8	<b>66.8</b>	+2.8	<b>58.3</b>	-1.3
OOD	99.3	+0.1	99.4	$\pm 0.01$	<b>89.0</b>	-4.7	68.0	-7.6	53.6	-6.7	53.6	-8.7
<b>IRM</b>	<b>0</b>		<b>1</b>		<b>2</b>		<b>3</b>		<b>4</b>		<b>5</b>	
	Acc.	$\Delta_{PA}$										
ID	<b>99.4</b>	-0.2	<b>99.5</b>	-0.1	86.4	-3.5	50.4	+12.0	48.2	+22.2	39.8	+19.8
1F-MD	99.4	$\pm 0.01$	99.5	$\pm 0.01$	86.5	+0.1	50.7	-4.1	48.3	+0.6	39.0	-3.1
5F-MD	99.2	$\pm 0.01$	99.5	$\pm 0.01$	83.0	$\pm 0.01$	<b>62.4</b>	$\pm 0.01$	<b>70.4</b>	$\pm 0.01$	<b>59.6</b>	$\pm 0.01$
OOD	99.4	$\pm 0.01$	99.5	$\pm 0.01$	<b>86.6</b>	$\pm 0.01$	46.5	$\pm 0.01$	48.8	$\pm 0.01$	35.9	$\pm 0.01$

Table 5.3: REMOVEopt=adam-lr=0.0001-mf=pos-npair=FalseREMOVE Test performance on increasingly shifted datasets for models selected during ERM and IRM procedures. Different validation datasets are used, and the selection capabilities of PA and validation accuracy are compared.

<b>ERM</b>	<b>0</b>		<b>1</b>		<b>2</b>		<b>3</b>		<b>4</b>		<b>5</b>	
	Acc.	$\Delta_{PA}$										
SD	99.4	-0.2	<b>99.5</b>	-0.1	78.6	+8.1	<b>77.3</b>	-13.2	36.7	+13.3	38.1	+13.7
ID	99.3	+0.1	99.5	-0.1	84.9	+1.6	66.3	+2.7	52.0	-1.7	46.3	+1.8
1F-MD	99.3	+0.1	99.5	-0.1	84.9	+1.6	66.3	+2.7	52.0	-1.7	46.3	+1.8
5F-MD	99.4	$\pm 0.01$	99.5	$\pm 0.01$	<b>87.4</b>	$\pm 0.01$	74.5	$\pm 0.01$	56.4	-0.1	<b>57.6</b>	$\pm 0.01$
OOD	<b>99.5</b>	$\pm 0.01$	99.5	-0.1	87.0	-0.5	75.1	-6.1	<b>56.9</b>	-6.7	57.5	-9.4
<b>IRM</b>	<b>0</b>		<b>1</b>		<b>2</b>		<b>3</b>		<b>4</b>		<b>5</b>	
	Acc.	$\Delta_{PA}$										
SD	<b>99.4</b>	-0.1	99.4	+0.1	<b>86.2</b>	-8.5	51.1	+6.3	48.9	+7.0	38.6	+11.9
ID	99.3	$\pm 0.01$	<b>99.5</b>	$\pm 0.01$	77.7	$\pm 0.01$	57.4	$\pm 0.01$	55.9	$\pm 0.01$	50.4	$\pm 0.01$
1F-MD	99.3	-0.1	99.5	$\pm 0.01$	84.6	-6.9	<b>69.8</b>	-12.4	64.4	-8.5	57.5	-7.1
5F-MD	99.2	$\pm 0.01$	99.5	$\pm 0.01$	83.0	$\pm 0.01$	62.4	$\pm 0.01$	<b>70.4</b>	$\pm 0.01$	<b>59.6</b>	$\pm 0.01$
OOD	99.3	$\pm 0.01$	99.5	$\pm 0.01$	84.6	$\pm 0.01$	69.8	$\pm 0.01$	64.4	$\pm 0.01$	57.5	$\pm 0.01$

Table 5.4: REMOVEopt=adam-lr=0.0001-mf=pos-npair=TrueREMOVE Test performance on increasingly shifted datasets for models selected during ERM and IRM procedures. Different validation datasets are used, and the selection capabilities of PA and validation accuracy are compared.

## 5.2 ID model selection

In the preceding chapters, evidence was provided supporting PA as a suitable robustness metric that effectively captures generalization capabilities under both sampling randomness and covariate shift. So far, experiments in the domain generalization setting have been conducted under synthetic conditions in which distribution shift is the only accountable source of randomness between  $\mathbf{x}$  and  $\mathbf{x}''$ . These experiments have shown that PA successfully discriminates robust from non-robust learners and also provides increased early-stopping performance compared with current baseline metrics.

However, real-world datasets are subject to sampling randomness and often exhibit feature distributions that are severely misaligned with the true distribution in the sample space, which is commonly known as subpopulation shift. This section aims to reproduce these conditions by considering controlled environments where the presence of certain image factors is deliberately manipulated to induce an inductive bias towards suboptimal representations. These representations may generalize well to sampling variability within source environments but fail to adapt to distributional shifts in target environments, which poses an additional challenge to our domain

generalization problem.

Epoch-wise model selection under these conditions entails a fundamentally different approach, especially regarding experiments performed in the previous chapter. Evaluating PA on a model selected by performance standards (i.e. validation accuracy) and

- Here interpretation of the validation accuracy.
- The datasets should be explained maybe.

**Definition** (SO/GO in DiagVib-6). Let shape, hue, lightness, position, scale and texture be MNIST image factors that can be manipulated through the DiagViB-6 data generation pipeline. Let  $F^P$  be the factor to be predicted by the classifier, and  $F^L$  the factor taking different values in the training and validation datasets, thus defining source environments.

The inductive bias of the model will be influenced by spurious correlations between  $F^P$  and  $F^L$  if the co-occurrence of their instances is not uniform [20]. From that perspective, a shortcut opportunity (SO) will be induced when a specific instance of  $F^P$  is exclusively co-occurred with a specific instance of  $F^L$ . Conversely, a generalization opportunity (GO) will arise when such  $F^P$  and  $F^L$  instances are each additionally co-occurred with other instances of  $F^L$  and  $F^P$ , respectively, thus breaking the exclusivity. In this work, we will consider three particular settings:

- Zero generalization opportunities (ZGO): Each instance of  $F^P$  is exclusively co-occurred with a unique instance of  $F^L$ . The model is expected to overfit to spurious correlations and thus generalize poorly to datasets in which these are not present.
- Compositional generalization opportunities (CGO): Starting from the ZGO setting, the exclusive co-occurrence between some instances of  $F^P$  and  $F^L$  is broken by the presence of generalization opportunities. The model should be increasingly robust to unseen combinations the higher is the number of generalization opportunities.
- Zero shortcut opportunities (ZSO): All instances of  $F^P$  are uniformly co-occurred with all instances of  $F^L$ , so that all combinations of factors are present in source domains.

The setting for ID model selection requires that validation datasets contain the same configuration of factors than the training datasets. Experiments will be performed for ZGO, ZSO and single, double and triple CGO.

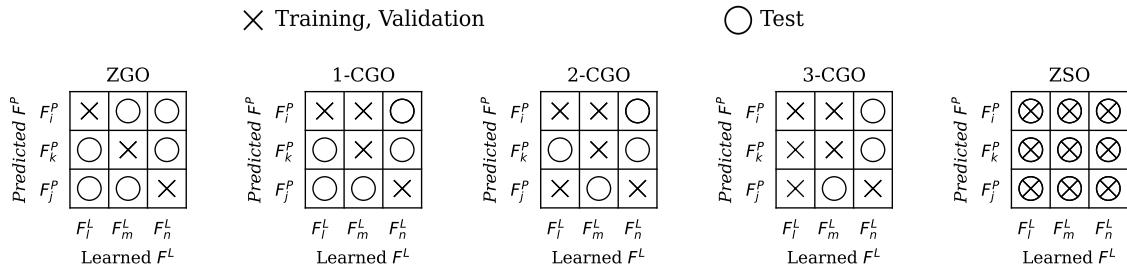


Figure 5.1: Blabla...

can be used as an early stopping criterion for model selection. Nevertheless, real world applications do not usually have a

More specifically, we aim at characterizing the response of the metric under different variations in the inductive bias of the model, which implicitly shifts according to the availability of certain features in the data. In this regard, two

1. Why epochwise is different than general model selection. KEY: Validation accuracy. The main problem in any robustness measurement is that we are not able to distinguish sampling randomness from robustness to adversarial shifts

2. We want to explore the difference between validation accuracy and PA for different inductive bias.

Best optimization hyperparameters were selected so that the best accuracy performance was obtained. No complex heuristic was needed given that the best hyperparameters were also the best for both learners and for all datasets with very few exceptions. Otherwise select the highest on the first domain, as the other domains are less significant. For that reason, models in the same table contain the same hyperparameters.

### REMOVE ENV 1 FOR ZSO

3. Results show that. What we do is paired because we want to make sure that the model overfits to the domains.

<b>ERM</b>	0		1		2		3		4		5	
	Acc.	$\Delta_{PA}$										
ZGO	99.9	<b>-45.6</b>	30.4	<b>+4.0</b>	33.3	<b>+2.4</b>	28.8	<b>-0.3</b>	33.6	<b>-0.3</b>	33.4	<b>+1.4</b>
1-CGO	99.4	<b>-51.8</b>	23.5	<b>+5.1</b>	25.1	<b>+4.0</b>	14.8	<b>+13.5</b>	32.7	<b>-2.4</b>	33.4	<b>-2.3</b>
2-CGO	98.5	$\pm 0.01$	57.5	$\pm 0.01$	58.3	$\pm 0.01$	37.1	$\pm 0.01$	38.9	$\pm 0.01$	38.5	$\pm 0.01$
3-CGO	98.3	$\pm 0.01$	65.7	$\pm 0.01$	66.0	$\pm 0.01$	42.8	$\pm 0.01$	41.8	$\pm 0.01$	40.3	$\pm 0.01$
ZSO	97.8	$\pm 0.01$	97.7	$\pm 0.01$	65.5	$\pm 0.01$	44.4	$\pm 0.01$	43.5	$\pm 0.01$	43.7	$\pm 0.01$

<b>IRM</b>	0		1		2		3		4		5	
	Acc.	$\Delta_{PA}$	Acc.	$\Delta_{PA}$	Acc.	$\Delta_{PA}$	Acc.	$\Delta_{PA}$	Acc.	$\Delta_{PA}$	Acc.	$\Delta_{PA}$
ZGO	100.0	<b>-32.9</b>	30.2	<b>+5.4</b>	33.9	<b>+1.9</b>	28.3	<b>+4.0</b>	32.6	<b>+1.0</b>	32.7	<b>+2.3</b>
1-CGO	99.6	<b>-40.0</b>	22.7	<b>+2.3</b>	24.8	<b>+1.7</b>	10.7	<b>+17.9</b>	32.9	<b>-3.9</b>	33.4	<b>-2.2</b>
2-CGO	98.9	<b>-0.1</b>	59.2	<b>-0.6</b>	57.7	<b>-0.2</b>	40.5	<b>-0.4</b>	39.2	<b>-0.3</b>	38.3	<b>-0.1</b>
3-CGO	98.5	$\pm 0.01$	67.7	$\pm 0.01$	67.7	$\pm 0.01$	45.5	$\pm 0.01$	43.1	$\pm 0.01$	41.9	$\pm 0.01$
ZSO	98.5	$\pm 0.01$	98.2	$\pm 0.01$	68.4	$\pm 0.01$	46.6	$\pm 0.01$	45.2	$\pm 0.01$	45.4	$\pm 0.01$

Table 5.5: REMOVEopt=sgd-lr=0.0001-mf=hue-npair=FalseREMOVE Test performance on increasingly shifted datasets for models selected during ERM and IRM procedures. Different validation datasets are used, and the selection capabilities of PA and validation accuracy are compared.

If the goal is to show that it performs better than accuracy, that can be easily shown, but the improvement is non-significant. But the principal challenge is theoretical: real-world covariate shift encompasses both distribution shift and sampling randomness, and PA (in the way we have been handling it) does not distinguish between them. When we compare the evolution of the training in between epochs, we will select the most robust features, but they won't necessarily be the most generalizable ones, they will depend on the nature of the data. => Here follows experiments on GO/SO. Explain the conclusions derived by those experiments. 1) Even when features are very similar, PA keeps increasing. 2) Model with more generalization opportunities is the least robust in high-shift scenarios. THIS IS THE KEY. The fact that robustness is so disentangled from accuracy (cite theoretical properties again) turns against us because what we want is actually to perform good. The reason is that SO entail a reduction in the effective feature space that the model navigates, which makes it more robust by definition. => Show the simple GO/SO experiment with paired samples. No sampling noise between  $X'$  and  $X''$ , only distribution shift. . Alternative way of thinking about it is forgetting about images. We are comparing different models trained on different instantiations of the features that represent our data. The randomness associated with the sampling process does not present homeostaticity, as noise magnitude increases with the distance between domains. The more generalizable, the more room for variation. In such case, PA would select the best model when the goal is to make features converge. In some sense it is, but overfitting to examples or subpopulation shifts would yield the same response.

For instance, I try to justify why PA does not always select the best performing model by refor-

mulating the problem so that the inputs are image features rather than images themselves. Every epoch represents an instantiation of the "feature extraction" experiment, and in such case a model overfitting to the shortcut opportunities in the data is also considered robust, since we don't control how much of the robustness "value" is attributed to sampling randomness and how much to the distribution shift. Sampling randomness can be modelled with the taus, and the key is that it presents heterocedasticity, in the sense that its "spread" is reduced when the model learns (also if it learns the wrong features) and that affects the PA value.

The most "elegant" way I have been able to devise (in the sense that it compiles both a measure of sampling noise and a measure of distribution shift) is comparing the first PCA direction in the feature space between both environments. The spread of each label in the PCA direction accounts for sampling randomness, whereas the MSE between components of the same sample in different environments accounts for distribution shift.

### 5.3 Model selection on benchmark datasets

In light of the results obtained in the previous sections, we will finally assess the model selection capabilities of PA on benchmark datasets. In particular, several WILDS [32] datasets will be considered, as they provide a comprehensive set of domain generalization tasks that are representative of real-world scenarios. Each of the datasets under consideration entails a specific configuration of learning opportunities that will shift the inductive bias towards suboptimal representations for out-of-distribution generalization. The performance displayed by PA in the previous chapter will be considered when evaluating its behaviour under these conditions.

- Show plot of the posteriors for each dataset.
- Show table with results on model selection.



# Chapter 6

## Conclusions

The objectives of this thesis centered on assessing the suitability of the posterior agreement framework as a robust model selection criterion within the context of deep learning for image classification tasks. The work entailed deriving and implementing an operative version of posterior agreement for discrete hypothesis classes, exploring its properties and comparing its discriminative power against that of baseline accuracy-based metrics. Experiments in adversarial and domain generalization settings were conducted to explore the source of the robustness measure provided by PA and its capabilities for selecting robustness-fostering algorithms. Additionally, results were provided on the effectiveness of PA as an early-stopping criterion for epoch-wise model selection, with experiments designed to assess its validity in both controlled synthetic environments and real-world scenarios, ultimately culminating in performance measures in benchmark out-of-distribution datasets. These objectives guided the research process, shaping the experiments and analyses that ultimately conform the conclusions of this work.

- Analyze the robustness challenge in image classification tasks. Mention the three different sources of randomness, and the trade-off existing when making models robust to them, which arises from a fundamental difference in the inductive bias of the model. This thesis introduces notation for a probability theory.
- The trade-off can be reformulated from an information theoretic perspective, as an estimation of the information content of the data, and the derived resolution over the hypothesis space. A robust learner is such that it correctly estimates the information content of the data and results in a hypothesis space that is stable to the randomness associated with the data generation process.
- The complexity of a model can be derived from the normalized description length value, and thus generalization error between pair of datasets can be defined by averaging it over the hypothesis space using the probability distribution of the first dataset.
- A lower bound for the generalization error can be obtained, leading to the expression of PA. An operative version of posterior agreement for finite hypothesis spaces is introduced and accessible using an efficient implementation that is easily to use by the ML community.
- Results were obtained

What you did: - Probability theory formulation of the covariate shift setting and the three sources of robustness. - Derivation of the posterior agreement kernel for finite hypothesis classes. Proof

of its properties, in particular convexity. - Efficient implementation of the PA metric. - Definition of a robustness metric, which is fundamentally different from a performance metric, and proof that PA does comply with the required properties. - Proved the suitability of PA as a robustness measure in the adversarial setting. In this setting, a model is considered to be more robust the better it performs in adversarial samples. In this sense, the robustness measure provided by PA was approximated and broken down to show that it accounts for both sampling randomness and randomness to perturbations, efficiently combining the usual standard accuracy-based metrics, but providing higher discriminability and consistency over different levels of attack power and samples ratio. - Proved that PA provides a consistent algorithm selection assessment based on its predictive power under distribution shifts.

Results have been outlined a deductive way, starting with an empirical exploration of the robustness properties of the PA kernel and culminating with performance results in benchmark datasets.

## Appendix A

# Theoretical Proofs and Derivations

We will define some notation shortcuts for the following proofs.

### A.1 Proof of problem formulation

**Lemma A.1.1.** Let  $N, K \in \mathbb{N}$  and let  $\{\mathcal{E}_{ij} \mid i \leq N, j \leq K\}$  be an indexed set of values. Then,

$$\sum_{c \in \mathcal{C}} \prod_{i=1}^N \mathcal{E}_{i,c(i)} = \prod_{i=1}^N \sum_{j=1}^K \mathcal{E}_{ij}$$

*Proof.* By induction on  $N$ . For the  $N = 1$  base case, observe that  $\mathcal{C}$  has only  $K$  elements, as there are only  $K$  functions mapping  $\{1\}$  to  $\{1, \dots, K\}$ . Then

$$\sum_{c \in \mathcal{C}} \prod_{i \leq N} \mathcal{E}_{i,c(i)} = \sum_{c \in \mathcal{C}} \mathcal{E}_{1,c(1)} = \sum_{j \leq K} \mathcal{E}_{1,j} = \prod_{i \leq N} \sum_{j \leq K} \mathcal{E}_{i,j}.$$

Assume now that the result holds for some  $N$ . We demonstrate then that it also holds for  $N + 1$ . Observe that there is a bijection between  $\mathcal{C}$  and  $\{1, \dots, K\}^N$ . Therefore, we identify every function  $c \in \mathcal{C}$  with the tuple  $(c(1), \dots, c(N))$ . Conversely, we identify every tuple  $(c_1, \dots, c_N) \in \{1, \dots, K\}^N$ , with the function  $c$  that maps  $i$  to  $c_i$ .

$$\begin{aligned}
& \sum_{c \in \mathcal{C}} \prod_{i \leq N+1} \mathcal{E}_{i,c(i)} = \\
&= \sum_{(c_1, \dots, c_{N+1}) \in \{1, \dots, K\}^{N+1}} \prod_{i \leq N+1} \mathcal{E}_{i,c_i} \\
&= \sum_{\substack{(c_1, \dots, c_N) \in \{1, \dots, K\}^N \\ c_{N+1} \leq K}} \prod_{i \leq N+1} \mathcal{E}_{i,c_i} \\
&= \sum_{(c_1, \dots, c_N) \in \{1, \dots, K\}^N} \sum_{c_{N+1} \leq K} \prod_{i \leq N+1} \mathcal{E}_{i,c_i} \\
&= \sum_{(c_1, \dots, c_N) \in \{1, \dots, K\}^N} \sum_{c_{N+1} \leq K} \left( \mathcal{E}_{N+1,c(N+1)} \prod_{i \leq N} \mathcal{E}_{i,c_i} \right) \\
&= \left( \sum_{c_{N+1} \leq K} \mathcal{E}_{N+1,c(N+1)} \right) \sum_{(c_1, \dots, c_N) \in \{1, \dots, K\}^N} \prod_{i \leq N} \mathcal{E}_{i,c_i} \\
&= \left( \sum_{c_{N+1} \leq K} \mathcal{E}_{N+1,c(N+1)} \right) \prod_{i \leq N} \sum_{j \leq K} \mathcal{E}_{i,j} \\
&= \left( \sum_{j \leq K} \mathcal{E}_{N+1,j} \right) \prod_{i \leq N} \sum_{j \leq K} \mathcal{E}_{i,j} \\
&= \prod_{i \leq N+1} \sum_{j \leq K} \mathcal{E}_{i,j}.
\end{aligned}$$

□

**Theorem A.1.1** (Posterior factorization). The posterior distribution for a classification problem can be factorized as follows:

$$\mathbf{P}^c(\theta | \mathbf{x}) = \prod_i^N \mathbf{P}^c(\theta_i | \mathbf{x}) = \prod_{i=1}^N \frac{\exp(\beta F_{\theta_i}(x_i))}{\sum_{j=1}^K \exp(\beta F_j(x_i))}$$

*Proof.* The posterior distribution solution to the MAP problem is the following:

$$\mathbf{P}^c(\theta | \mathbf{x}) \frac{\exp\left(\beta \sum_{i=1}^N F_{\theta_i}(x_i)\right)}{\sum_{\theta \in \Theta} \exp\left(\beta \sum_{i=1}^N F_{\theta_i}(x_i)\right)} = \frac{\prod_{i=1}^N \exp(\beta F_{\theta_i}(x_i))}{\sum_{\theta \in \Theta} \prod_{i=1}^N \exp(\beta F_{\theta_i}(x_i))}$$

Using Lemma 3.6.1 we can rewrite the denominator as:

$$\sum_{\theta \in \Theta} \prod_{i=1}^N \exp(\beta F_{\theta_i}(x_i)) = \prod_{i=1}^N \sum_{\theta \in \Theta} \exp(\beta F_{\theta_i}(x_i))$$

Therefore, the posterior distribution can be written as:

$$\mathbf{P}^c(\theta | \mathbf{x}) = \prod_{i=1}^N \mathbf{P}^c(\theta_i | \mathbf{x}) = \prod_{i=1}^N \frac{\exp(\beta F_{\theta_i}(x_i))}{\sum_{\theta \in \Theta} \exp(\beta R(\theta, \mathbf{x}))}$$

□

## A.2 Properties of the PA kernel

**Theorem A.2.1** (Symmetry of the PA kernel). The posterior agreement kernel is symmetric with respect to the definition of  $X'$  and  $X''$ .

$$PA(\mathbf{x}', \mathbf{x}'') = PA(\mathbf{x}'', \mathbf{x}')$$

*Proof.* Trivial, commutative property.  $\square$

**Theorem A.2.2** (Non-negativity of the PA kernel). The posterior agreement kernel is non-negative.

$$PA(\mathbf{x}', \mathbf{x}'') \leq 0$$

*Proof.* See Lemma 2.3.1.  $\square$

**Theorem A.2.3** (Concavity of the PA kernel). The posterior agreement kernel is concave in  $\mathbb{R}^+$ , and therefore has a unique maximum.

*Proof.* The posterior agreement kernel has been shown to have the following form:

$$PA(\mathbf{x}', \mathbf{x}'') \propto \sum_{n=1}^N \log \left[ \sum_{j=1}^K \mathbf{P}_n^c(\theta | x'_n) \mathbf{P}_n^c(\theta | x''_n) \right]$$

where the posteriors  $\mathbf{P}_n^c(\theta | x_n)$  are Gibbs distributions for each observation.

$$\mathbf{P}_n^c(\theta | x'_n) = \frac{e^{\beta F_j(x_n)}}{\sum_{k=1}^K e^{\beta F_k(x_n)}}$$

We will require three important results from optimization theory:

**T1** The minimum of  $G(\beta) = -PA(X', X'')$  over the convex set  $\mathbb{R}^+$  is unique  $\iff G(\beta)$  is convex.

**T2**  $G$  is absolutely convex  $\iff \frac{d^2}{d\beta^2} G(\beta) > 0$ .

**T3** The sum of convex functions is also convex.

To streamline the derivation, the following notation will be used:

$$F_j(x'_n) = F'_j$$

$$e^{\beta F_j(x'_n)} = e^{\beta F'_j} = e'_j$$

The observation index  $n$  will be omitted as it does not affect the convexity derivation (see **T3**). With that notation in mind, we can define  $G(\beta)$  properly:

$$G(\beta) = -k(\mathbf{x}', \mathbf{x}'') = \sum_{n=1}^N -\log \left[ \sum_{j=1}^K e'_j e''_j \right] + \sum_{n=1}^N \log \left[ \sum_{k=1}^K e'_k \sum_{p=1}^K e''_p \right]$$

We will focus on the first term:  $G_1^n(\beta) = G_1(\beta) = \log \left[ \sum_{j=1}^K e'_j e''_j \right]$ .

$$\frac{d}{d\beta} G_1(\beta) = \frac{\sum_{j=1}^K (F'_j + F''_j) e'_j e''_j}{\sum_{j=1}^K e'_j e''_j}$$

We will recurrently use the derivative  $\frac{d}{d\beta} e'_j e''_k$  in this proof:

$$\frac{d}{d\beta} e'_j e''_k = F'_j e'_j e''_k + e'_j F''_k e''_k = (F'_j + F''_k) e'_j e''_k$$

The second derivative is straightforward:

$$\frac{d^2}{d\beta^2} G_1(\beta) = \frac{\sum_{j=1}^K (F'_j + F''_j)^2 e'_j e''_j}{\sum_{j=1}^K e'_j e''_j} - \frac{\left( \sum_{j=1}^K (F'_j + F''_j) e'_j e''_j \right)^2}{\left( \sum_{j=1}^K e'_j e''_j \right)^2}$$

We impose the convexity condition and see whether it can be contradicted.

$$\frac{d^2}{d\beta^2} G_1(\beta) > 0 \iff \left( \sum_{j=1}^K e'_j e''_j \right) \left( \sum_{j=1}^K (F'_j + F''_j)^2 e'_j e''_j \right) - \left( \sum_{j=1}^K (F'_j + F''_j) e'_j e''_j \right)^2 > 0$$

Using the distributive property of the product over the sum, we can reindex our expression:

$$\begin{aligned} \sum_{k=1}^K \sum_{j=1}^K (F'_j + F''_j)^2 e'_j e''_j e'_k e''_k - \sum_{k=1}^K \sum_{j=1}^K (F'_j + F''_j) (F'_k + F''_k) e'_j e''_j e'_k e''_k &> 0 \iff \\ \sum_{k=1}^K \sum_{j=1}^K [(F'_j + F''_j) - (F'_k + F''_k)] (F'_j + F''_j) e'_j e''_j e'_k e''_k &> 0 \end{aligned}$$

As we can see,  $\Delta_{(jj),(kk)}$  corresponds to the difference in the cost attributed to reference class  $j$  and the cost attributed to class  $k$ , accumulated over  $\mathbf{x}', \mathbf{x}''$ . We can intuitively devise some symmetry in these terms, and we formalize it as follows:

$$E_{jk} = e'_j e''_j e'_k e''_k = E_{kj}$$

$$\Delta_{(jj),(kk)} = (F'_j + F''_j) - (F'_k + F''_k) = (F'_j - F'_k) + (F''_j - F''_k) = -\Delta_{(kk),(jj)}$$

Even if  $\Delta_{(jj),(jj)} = 0$ , we will still include this term to facilitate with the indexing. Overall, the sum can be expressed as:

$$\sum_{k=1}^K \sum_{j=1}^K [(F'_j + F''_j) - (F'_k + F''_k)] (F'_j + F''_j) e'_j e''_j e'_k e''_k = \sum_{k=1}^K \sum_{j=1}^K (F'_j + F''_j) E_{jk} \Delta_{(jj),(kk)} = \sum_{k=1}^K \sum_{j=1}^K S_{(jj),(kk)}$$

Then, the pairwise sum of symmetric combinations of indexes  $k$  and  $j$  yields

$$\begin{aligned} S_{(jj),(kk)} + S_{(kk),(jj)} &= (F'_j + F''_j) E_{jk} \Delta_{(jj),(kk)} + (F'_k + F''_k) E_{kj} \Delta_{(kk),(jj)} \\ &= E_{jk} \Delta_{(jj),(kk)} [(F'_j + F''_j) - (F'_k + F''_k)] = E_{jk} \Delta_{(jj),(kk)}^2 > 0 \end{aligned}$$

Given that the indexing sets in our nested sum are the same, it's straightforward to see that all the terms will be strictly positive, and the overall sum will be zero only if  $e_j = 0 \forall j = \{1, \dots, K\}$ , which is not possible in a classification setting since  $\beta \in \mathbb{R}^+$ . We end up with the following expression:

$$\frac{d^2}{d\beta^2} G_1(\beta) = \sum_{k=1}^K \sum_{j < k} E_{jk} \Delta_{(jj),(kk)}^2 > 0$$

Now we proceed analogously with the second term:

$$\begin{aligned} G_2^n(\beta) &= G_2(\beta) = \log \left[ \sum_{j=1}^K e'_j \sum_{k=1}^K e''_k \right] = \log \left[ \sum_{k=1}^K \sum_{j=1}^K e'_j e''_k \right] \\ \frac{d}{d\beta} G_2(\beta) &= \frac{\sum_{k=1}^K \sum_{j=1}^K (F'_j + F''_j) e'_j e''_k}{\sum_{k=1}^K \sum_{j=1}^K e'_j e''_k} \end{aligned}$$

$$\begin{aligned}
\frac{d^2}{d^2\beta}G_2(\beta) &= \frac{\sum_{k=1}^K \sum_{j=1}^K (F'_j + F''_k)^2 e'_j e''_k}{\sum_{k=1}^K \sum_{j=1}^K e'_j e''_k} - \frac{\left( \sum_{k=1}^K \sum_{j=1}^K (F'_j + F''_k) e'_j e''_k \right)^2}{\left( \sum_{k=1}^K \sum_{j=1}^K e'_j e''_k \right)^2} > 0 \iff \\
&\iff \left( \sum_{k=1}^K \sum_{j=1}^K e'_j e''_k \right) \left( \sum_{k=1}^K \sum_{j=1}^K (F'_j + F''_k)^2 e'_j e''_k \right) - \left( \sum_{k=1}^K \sum_{j=1}^K (F'_j + F''_k) e'_j e''_k \right)^2 > 0 \iff \\
&\iff \sum_{k=1}^K \sum_{q=1}^K \sum_{j=1}^K \sum_{i=1}^K (F'_j + F''_k)^2 e'_j e''_k e'_i e''_q - (F'_j + F''_k) e'_j e''_k (F'_i + F''_q) e'_i e''_q > 0 \iff \\
&\iff \sum_{k=1}^K \sum_{q=1}^K \sum_{j=1}^K \sum_{i=1}^K (F'_j + F''_k) e'_j e''_k e'_i e''_q [(F'_j + F''_k) - (F'_i + F''_q)] > 0
\end{aligned}$$

We can define as well:

$$\begin{aligned}
\frac{d^2}{d^2\beta}G_2(\beta) &= \sum_{k=1}^K \sum_{q=1}^K \sum_{j=1}^K \sum_{i=1}^K S_{(jk),(iq)} = \sum_{k=1}^K \sum_{q=1}^K \sum_{j=1}^K \sum_{i=1}^K (F'_j + F''_k) E_{(jk),(iq)} \Delta_{(jk),(iq)} \\
E_{(jk),(iq)} &= e'_j e''_k e'_i e''_q = E_{(ik),(jq)} = E_{(jq),(ik)} = E_{(iq),(jk)} \\
\Delta_{(jk),(iq)} &= (F'_j - F'_i) + (F''_k - F''_q) = -\Delta_{(iq),(jk)}
\end{aligned}$$

The symmetry arises when adding two elements that have mirror indexes in both  $\mathbf{x}'$  and  $\mathbf{x}''$ .

$$\begin{aligned}
S_{(jk),(iq)} + S_{(iq),(jk)} &= (F'_j + F''_k) E_{(jk),(iq)} \Delta_{(jk),(iq)} + (F'_i + F''_q) E_{(iq),(jk)} \Delta_{(iq),(jk)} \\
&= E_{(jk),(iq)} \Delta_{(jk),(iq)} [(F'_j + F''_k) - (F'_i + F''_q)] = E_{(jk),(iq)} \Delta_{(jk),(iq)}^2 > 0
\end{aligned}$$

Given that symmetries are independent for  $\mathbf{x}'$  and  $\mathbf{x}''$ , we end up with a similar expression:

$$\frac{d^2}{d\beta^2}G_2(\beta) = \sum_{k=1}^K \sum_{q<k}^K \sum_{j=1}^K \sum_{i<j}^K E_{(jk),(iq)} \Delta_{(jk),(iq)}^2 > 0$$

Even if a further simplified version can be obtained, this one will allow us to complete the proof. We can now define the function  $G(\beta)$  as the sum of the two terms:

$$\frac{d^2}{d\beta^2}G(\beta) = \sum_{n=1}^N \left[ \sum_{k=1}^K \sum_{q<k}^K \sum_{j=1}^K \sum_{i<j}^K E_{(jk),(iq)} \Delta_{(jk),(iq)}^2 - \sum_{k=1}^K \sum_{q<k}^K E_{(qq),(kk)} \Delta_{(qq),(kk)}^2 \right]$$

where we can clearly see that the particular case  $\{k = j, q = i\}$  cancels the negative terms:

$$\begin{aligned}
\frac{d^2}{d\beta^2}F^n(\beta) &= \sum_{k=1}^K \sum_{q<k}^K \sum_{j=\{1:K\}\setminus\{k\}} \sum_{i=\{1:K|i<j\}\setminus\{k\}} E_{(jk),(iq)} \Delta_{(jk),(iq)}^2 \\
&\quad + \sum_{k=1}^K \sum_{q<k}^K E_{(qq),(kk)} \Delta_{(qq),(kk)}^2 - \sum_{k=1}^K \sum_{q<k}^K E_{(qq),(kk)} \Delta_{(qq),(kk)}^2 = \\
&= \sum_{k=1}^K \sum_{q<k}^K \sum_{j=\{1:K\}\setminus\{k\}} \sum_{i=\{1:K|i<j\}\setminus\{k\}} E_{(jk),(iq)} \Delta_{(jk),(iq)}^2 > 0
\end{aligned}$$

Which proves that  $G(\beta)$  is absolutely convex in  $\mathbb{R}^+$ :

$$\frac{d^2}{d\beta^2}G(\beta) = \sum_{n=1}^N \frac{d^2}{d\beta^2}G^n(\beta) = \sum_{n=1}^N \left[ \sum_{k=1}^K \sum_{q < k} \sum_{j=\{1:K\} \setminus \{k\}} \sum_{i=\{1:K| i < j\} \setminus \{k\}} E_{(jk),(iq)} \Delta_{(jk),(iq)}^2 \right] > 0$$

We must note that on the limit  $\beta \rightarrow \infty$  the curvature is not defined, so it will be always a good practice to start the numerical procedure at a value  $\beta_0 = 0^+$ :

$$\lim_{\beta \rightarrow 0^+} \frac{d^2}{d\beta^2}G(\beta) > 0$$

□

**Properties.**  $\text{PA}(\mathbf{x}', \mathbf{x}''; \beta)$  has the expected behaviour in the artificial classifier examples.

**C1** In a random classifier, accuracy tends to 50% as sample size grows, but the posteriors are not necessarily paired. The highest agreement will be achieved when posteriors are completely flat; that is, when  $\beta = 0$ . In general:

$$\text{PA}(\mathbf{x}', \mathbf{x}''; \beta) \leq N \log \frac{1}{2} \quad \forall \beta \in \mathbb{R}^+, \quad \text{with } \lim_{\beta \rightarrow 0} \text{PA}(\mathbf{x}', \mathbf{x}''; \beta) = N \log \frac{1}{2}$$

**C2** In a perfect classifier, accuracy reaches 100% and the posteriors

$$\mathbf{P}_i^c(j | x'_i) = \mathbf{P}_i^c(j | x''_i) = \delta_j(y_i) \implies \lim_{\beta \rightarrow \beta^*=\infty} \text{PA}(\mathbf{x}', \mathbf{x}''; \beta) = 0$$

**C3** In a constant classifier, accuracy reaches 50%, and the posteriors

$$\mathbf{P}_i^c(j | x'_i) = \mathbf{P}_i^c(j | x''_i) = \delta_j(0) \implies \lim_{\beta \rightarrow \beta^*=\infty} \text{PA}(\mathbf{x}', \mathbf{x}''; \beta) = 0$$

## Appendix B

# Supplementary Results

### B.1 PA as a robustness metric

#### B.1.1 Empirical behaviour

Results on the empirical behaviour of the PA metric also serve as an exploration of the optimization landscape. As seen in Theorem 3.6.3, the posterior agreement kernel is concave for  $\beta \in \mathbb{R}^+$ , which implies that the optimization problem has a unique maximum.

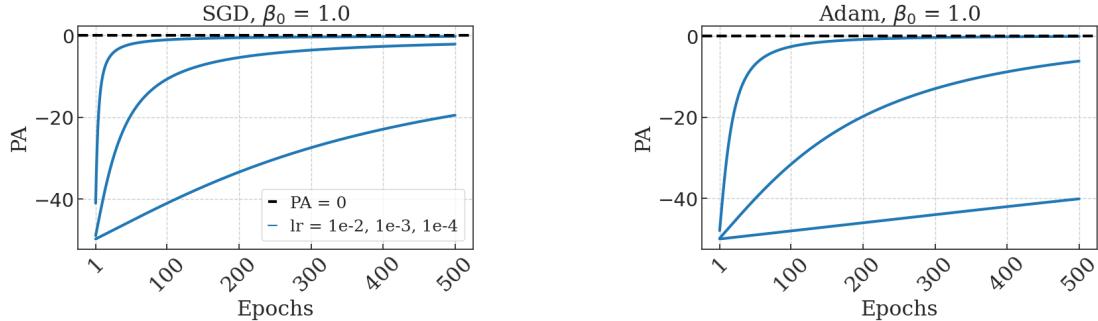


Figure B.1: Evolution of the  $\beta$  optimization for a robust sample.

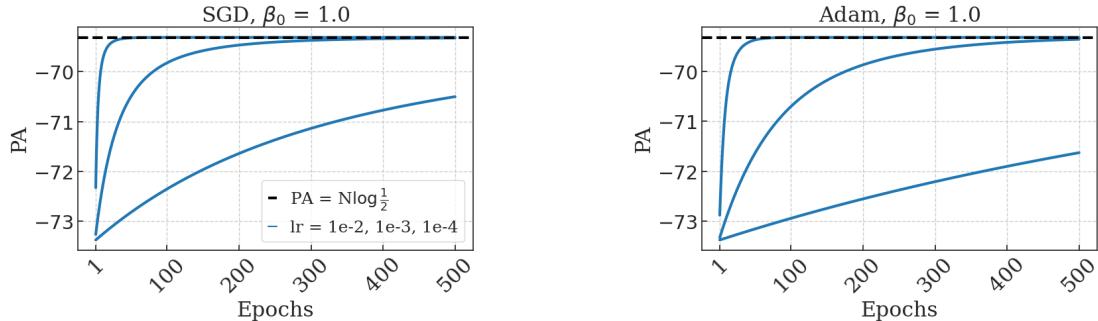


Figure B.2: Evolution of the  $\beta$  optimization for a non-robust sample.

The bernoulli sample simulation allows us also to assess the behaviour of the PA metric under different levels of prediction confidence. For instance, it was shown in Figure 4.2 that the value of

$\beta^*$  is highly informative of the nature of the model's output probability distribution, and could be an indication of possible underfitting or overfitting to specific features of the training set, which is highly valuable in the covariate shift setting.

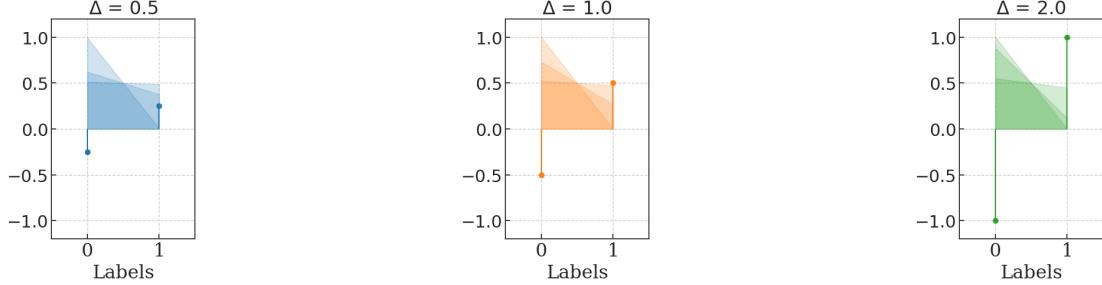


Figure B.3: Logit distributions associated with the behaviour observed in Figure 4.2.

We see that in the case of a non-robust model, the higher the beta the more pointy is the distribution. Given that many samples are completely misaligned, the highest  $\beta$  will be zero, which is when the distribution is completely flat. The smaller is the difference in the logits, the less pointy is the distribution for  $\beta > 0$ , which means that the overlap will be higher.

Finally, we can also check whether the optimization of the kernel is consistent with results on its concavity and the existence of a unique maximum. The following figures show that optimization converges for  $0 < \beta < M$ , with  $M$  large enough so that concavity is less and less defined.

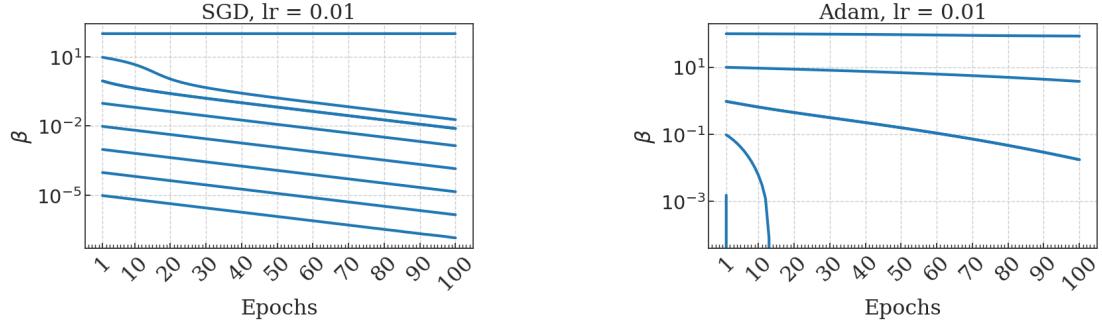


Figure B.4: Evolution of  $\beta$  optimization for different initial values for a non-robust classifier.

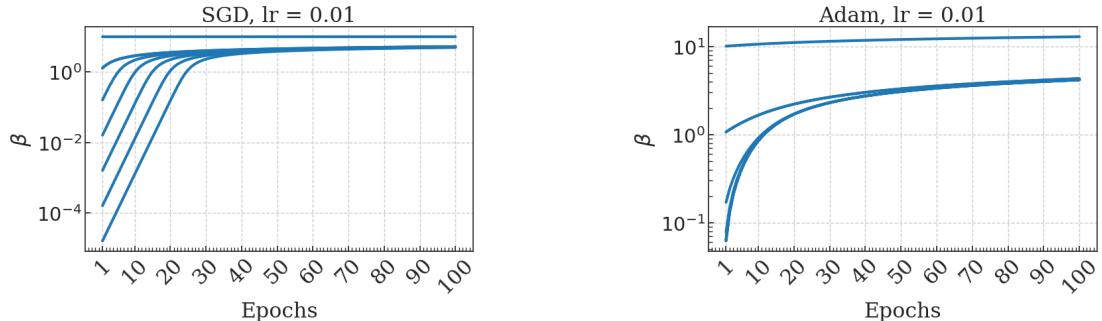


Figure B.5: Evolution of  $\beta$  optimization for different initial values for a robust classifier.

Some exploratory experiments in real scenarios were performed, comparing the evolution of PA and accuracy when datasets are perturbed with random noise. An even more significant difference can be observed when perturbations are not random, but instead are designed to mislead the model’s predictions. In these settings, PA arises as a reliable alternative.

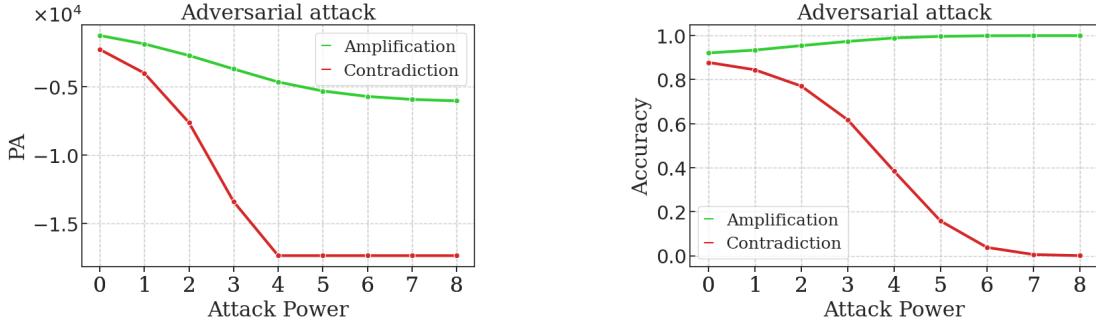


Figure B.6: PA and accuracy for the IMDB sentiment classification task under simple adversarial attacks. Observations are perturbed by replacing some words with positive or negative adjectives that either encourage (amplification) or discourage (contradiction) the true sentiment of the review. The attack power is defined as  $2^W$ , being  $W$  the number of words replaced.

## B.2 Adversarial setting

The first result provided in the adversarial setting is the entropy difference between initial (i.e.  $\beta = 1$ ) and optimal posterior distributions, which is shown to decrease significantly in robust models, due to the fact that few samples are misclassified and therefore maximum agreement is achieved with higher inverse temperature values. Entropy is computed for the average posterior distribution on correctly classified samples, which represent the largest portion of the dataset.

One of the claims made about PA is that its discriminative power is superior to that of AFR, because its value is not so much driven by the sampling randomness associated to a specific experiment (i.e. dataset), and therefore provides a more reliable assessment of the robustness capabilities of the model. We can observe that AFR(P), which is by definition the baseline measure of robustness in the adversarial setting, is way less discriminative and fluctuates its value significantly over different presence of adversarial samples.

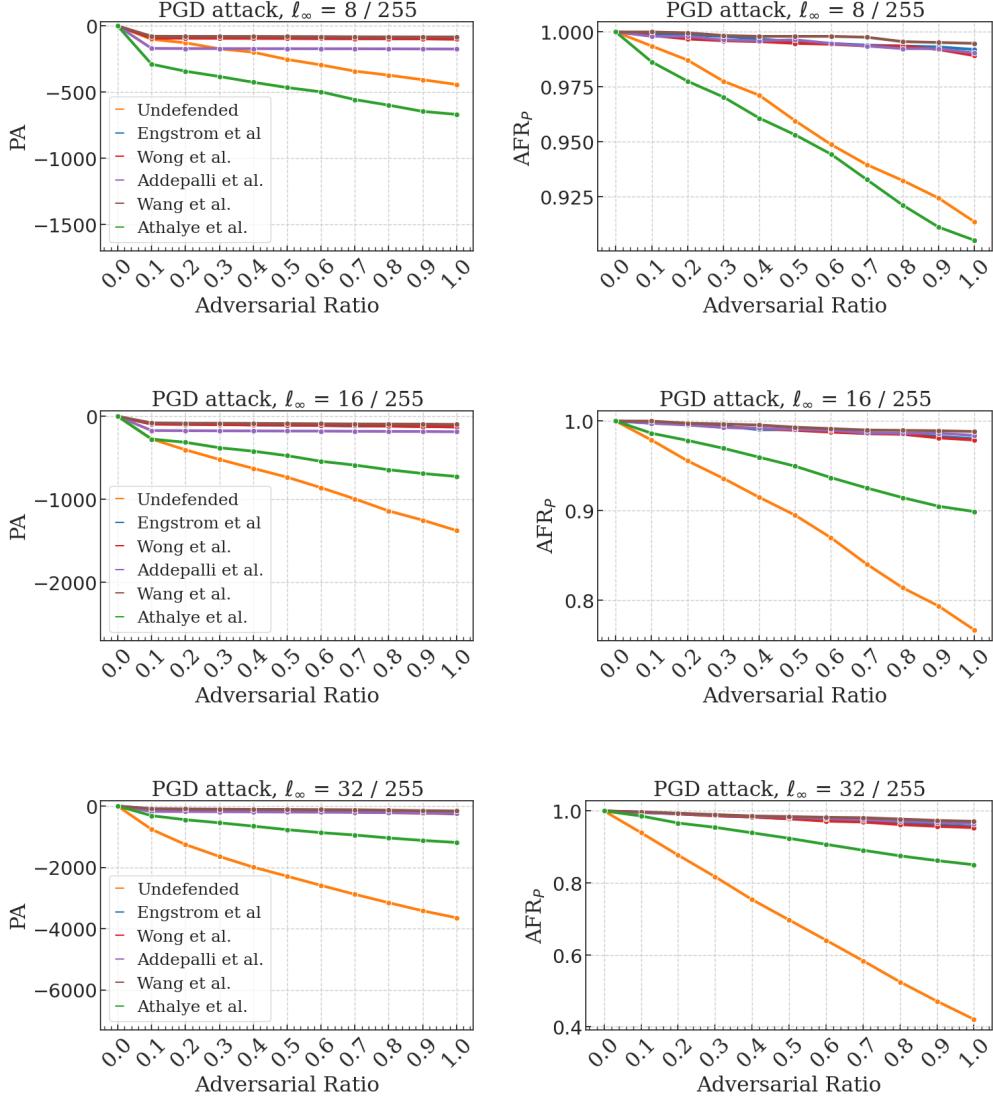


Figure B.7: PA and AFR(P) variation under increasing adversarial ratio at different perturbation norm bounds. The undefended net and several RobustBench robust models are considered against a 1000 step PGD attack.

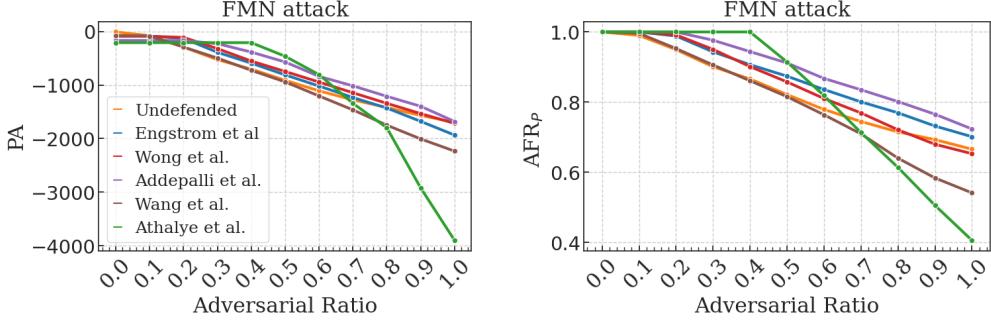


Figure B.8: PA and AFR(P) variation under increasing adversarial ratio. The undefended net and several RobustBench robust models are considered against a 1000 step FMN attack.

Another claim that was made is the fact that optimal posterior distributions are assumed to be highly peaked at the predicted class. This assumption allows us to break down the PA robustness score into a sum of terms that represent specific contributions to the robustness of the model, and that can be approximated analytically.

**Theorem B.2.1** (Approximated PA in the adversarial setting). The assumption of a peaked gibbs posterior allows approximate the maximum PA value as follows:

$$\text{PA} \approx N\tau\rho \log(1 - 2\delta_{\text{ERR}}) + N(1 - \tau)\rho \log(1 - 2\delta_{\text{MIS}}) + N\tau(1 - \rho) \log \delta_{\text{ADV}},$$

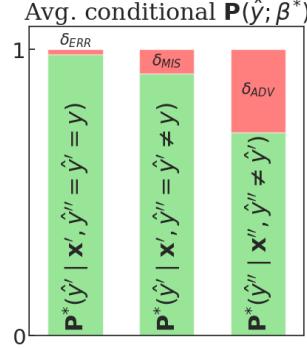


Figure B.9: Illustrative representation of the terms and posterior values constrained considered for the PA approximation.

*Proof.* Let  $\hat{y}'$  and  $\hat{y}''$  be the predicted class for an arbitrary original and an adversarial samples, respectively, and  $y_{\text{true}}$  the true class. The first and second most likely labels for a sample are obtained as:

$$\begin{aligned}\hat{y}_{\text{first}} &= \arg \max_y \mathbf{P}(y | \mathbf{x}) = \hat{y} \\ \hat{y}_{\text{next}} &= \arg \max_{y \setminus \{\hat{y}\}} \mathbf{P}(y | \mathbf{x})\end{aligned}$$

Let  $\mathbf{P}^*$  be the optimal posterior distribution over the classes for a specific sample; that is, the gibbs distribution with inverse temperature  $\beta^*$ . Following the PA kernel expression, the contribution of each pair of samples can be approximated as

$$\Xi = \log \left\{ \sum_{y \in \mathcal{Y}} \mathbf{P}^*(y | \mathbf{x}') \mathbf{P}^*(y | \mathbf{x}'') \right\} \approx \log \{ \mathbf{P}^*(\hat{y}' | \mathbf{x}') \mathbf{P}^*(\hat{y}'' | \mathbf{x}'') + \mathbf{P}^*(\hat{y}'_{\text{next}} | \mathbf{x}') \mathbf{P}^*(\hat{y}''_{\text{next}} | \mathbf{x}'') \}.$$

Let  $\mathfrak{f}$  be the fraction of samples that contribute to the PA value in a specific way for a given adversarial ratio, so that  $N\mathfrak{f}$  is the number of contributing terms. To avoid notation clutter, we will define:

$$\begin{aligned} \tau &= \text{ACC}(\hat{\mathbf{y}}', \mathbf{y}_{\text{true}}) = \text{AFR}(\text{T}) \Big|_{\text{AR}=0.0} \\ \rho &= \text{AFR}(\text{P}) \Big|_{\text{AR}} \end{aligned}$$

Given that optimal posteriors are expected to be peaked, these contributions can be approximated for three relevant cases.

**Case**  $y_{\text{true}} = \hat{y}'' = \hat{y}' = \hat{y}$ . Clearly  $\mathfrak{f} = \tau\rho$ . Then

$$\begin{aligned} \mathbf{P}^*(\hat{y} | \mathbf{x}') &\approx \mathbf{P}^*(\hat{y} | \mathbf{x}'') = 1 - \delta_{\text{ERR}}, \\ \mathbf{P}^*(\hat{y}'_{\text{next}} | \mathbf{x}') &\approx \mathbf{P}^*(\hat{y}''_{\text{next}} | \mathbf{x}'') \approx \delta_{\text{ERR}}, \end{aligned}$$

which yields

$$\Xi_{\text{ERR}}^i \approx \log \left\{ (1 - \delta_{\text{ERR}})^2 + \delta_{\text{ERR}}^2 \right\} \approx \log (1 - 2\delta_{\text{ERR}}),$$

where  $\delta_{\text{ERR}}$  represents the lack of confidence when successfully predicting original samples.

**Case**  $y_{\text{true}} \neq \hat{y}'' = \hat{y}' = \hat{y}$ . Clearly  $\mathfrak{f} \approx (1 - \tau)\rho$ . Then

$$\begin{aligned} \mathbf{P}^*(\hat{y} | \mathbf{x}') &\approx \mathbf{P}^*(\hat{y} | \mathbf{x}'') = 1 - \delta_{\text{MIS}}, \\ \mathbf{P}^*(\hat{y}'_{\text{next}} | \mathbf{x}') &\approx \mathbf{P}^*(\hat{y}''_{\text{next}} | \mathbf{x}'') \approx \delta_{\text{MIS}}, \end{aligned}$$

which yields

$$\Xi_{\text{MIS}}^i \approx \log \left\{ (1 - \delta_{\text{MIS}})^2 + \delta_{\text{MIS}}^2 \right\} \approx \log (1 - 2\delta_{\text{MIS}}),$$

where  $\delta_{\text{MIS}}$  represents the missing prediction confidence on misclassified original samples.

**Case**  $y_{\text{true}} = \hat{y}'' \neq \hat{y}' = \hat{y}$ . Clearly  $\mathfrak{f} = \tau(1 - \rho)$ . Then

$$\begin{aligned} \mathbf{P}^*(\hat{y}' | \mathbf{x}') &= 1 - \delta_{\text{ERR}}, \\ \mathbf{P}^*(\hat{y}'' | \mathbf{x}') &\approx \delta_{\text{ERR}}; \\ \mathbf{P}^*(\hat{y}'' | \mathbf{x}'') &= 1 - \delta_{\text{ADV}}, \\ \mathbf{P}^*(\hat{y}' | \mathbf{x}'') &\approx \delta_{\text{ADV}}, \end{aligned}$$

which yields

$$\Xi_{\text{ADV}}^i \approx \log \{ (1 - \delta_{\text{ERR}}) \delta_{\text{ADV}} + \delta_{\text{ERR}} (1 - \delta_{\text{ADV}}) \} \approx \log \delta_{\text{ADV}},$$

given that  $\delta_{\text{ERR}} \approx -2\delta_{\text{ERR}}\delta_{\text{ADV}}$ .  $\delta_{\text{ADV}}$  represents the missing confidence in the prediction of a misleading adversarial sample.

The approximated PA value amounts to the sum of all contributions.  $\square$

The previous expression has been validated empirically with both an FMN attack and an  $\ell_\infty=8/255$  PGD attack computed on the CIFAR10 data.

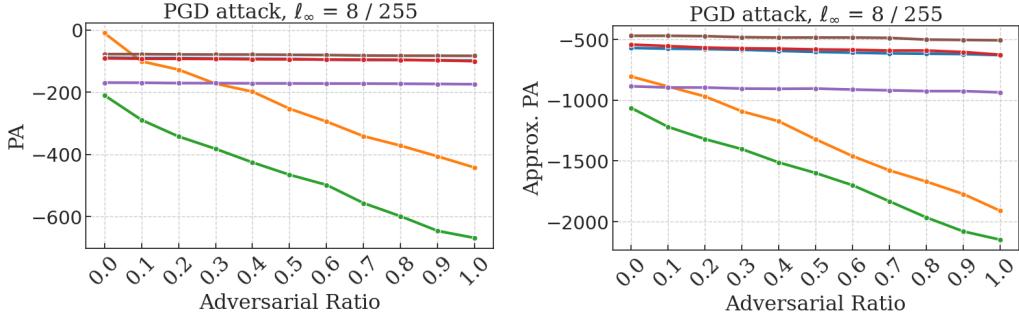


Figure B.10: True and approximated PA values under increasing adversarial ratio for a PGD attack with  $\ell_\infty=8/255$ .

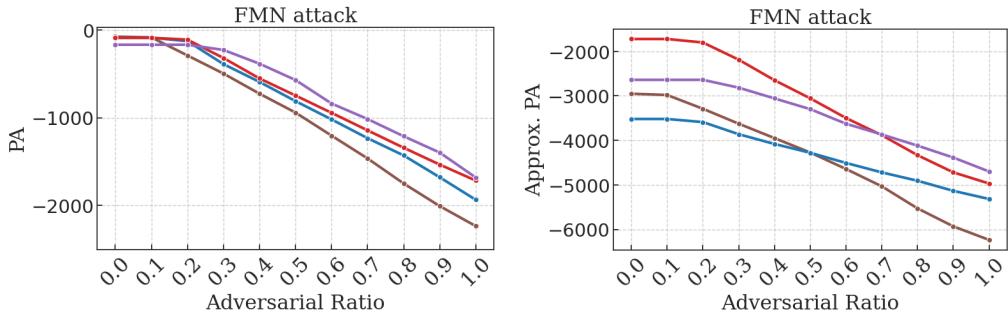


Figure B.11: True and approximated PA values under increasing adversarial ratio for a FMN attack.

Results obtained for this section illustrate the higher effectiveness of FMN with respect to PGD attacks in the adversarial dataset. Further insight into the response of the models in each case can be obtained by comparing the average posterior distributions on original samples, originally misclassified samples and adversarial samples.

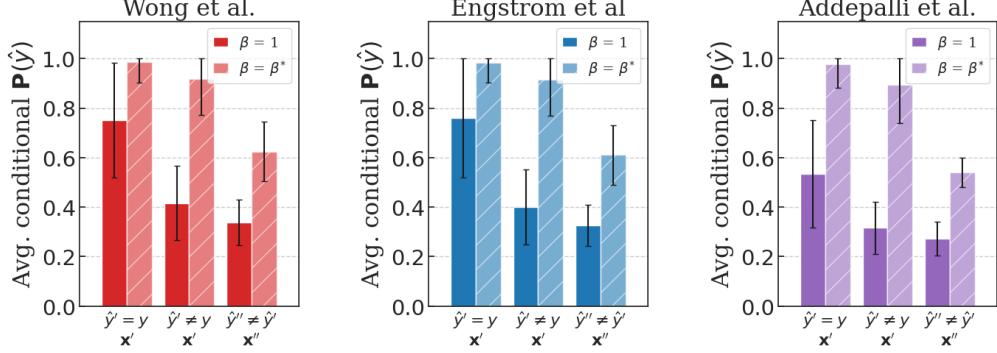


Figure B.12: Average posterior probability of the predicted class for correctly classified original samples, misclassified original samples, and misleading adversarial samples, respectively. Results have been obtained through a PGD attack with  $\ell_\infty=8/255$  and sorted by increasing  $\beta^*$ .

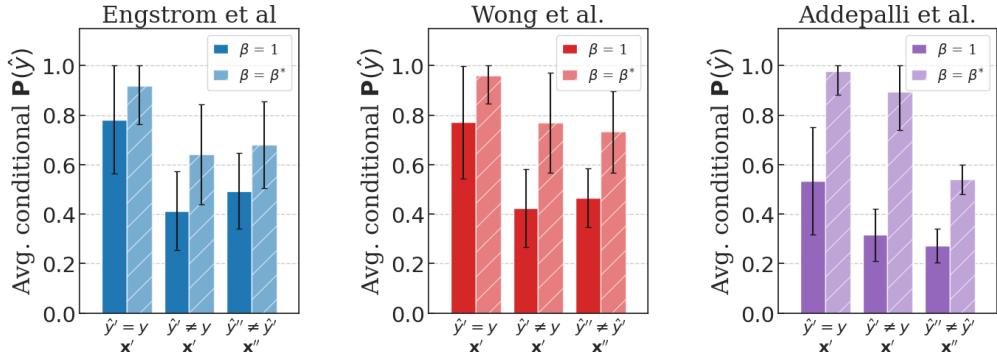


Figure B.13: Average posterior probability of the predicted class for correctly classified original samples, misclassified original samples, and misleading adversarial samples, respectively. Results have been obtained through a FMN attack and sorted by increasing  $\beta^*$ .

The results obtained in this section culminate with the comparison of PA with other metrics that are usually employed to assess dissimilarity in the response of models under different conditions. In particular, confidence-based metrics such as Kullback-Leibler divergence and Wasserstein distance are used to compare the probability output of original and perturbed datasets, and feature-space-based metrics assess the overall distribution of the latent representation of the samples before the discriminant function.

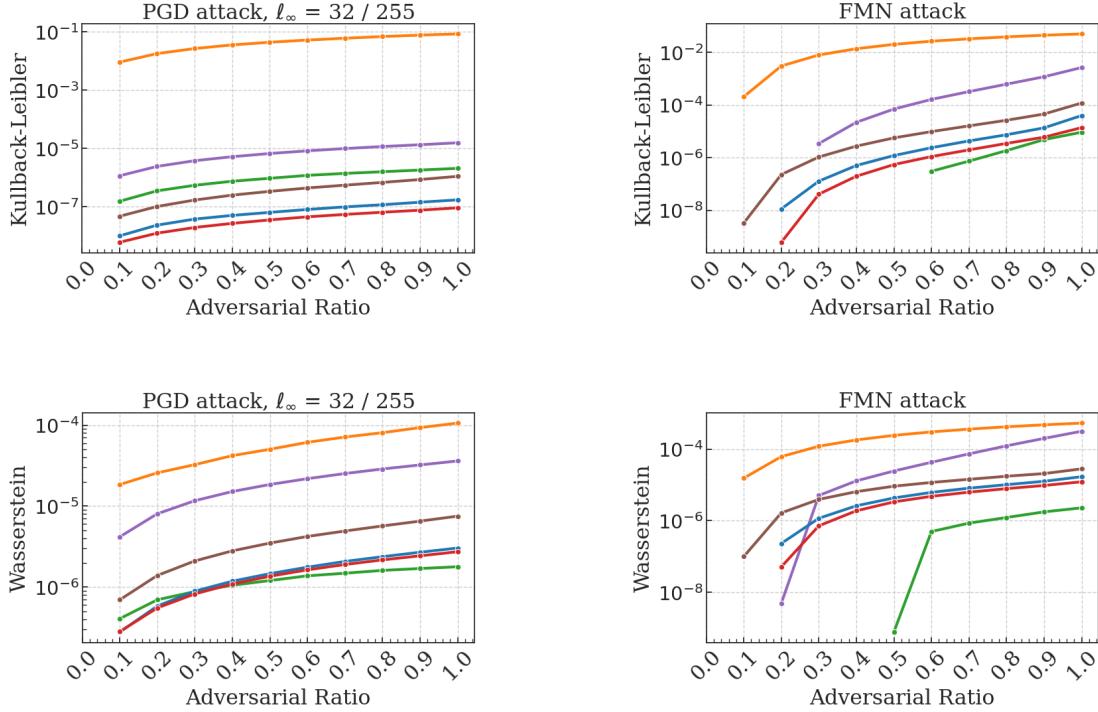


Figure B.14: Comparison of FMN and PGD attacks using probability-based distances.

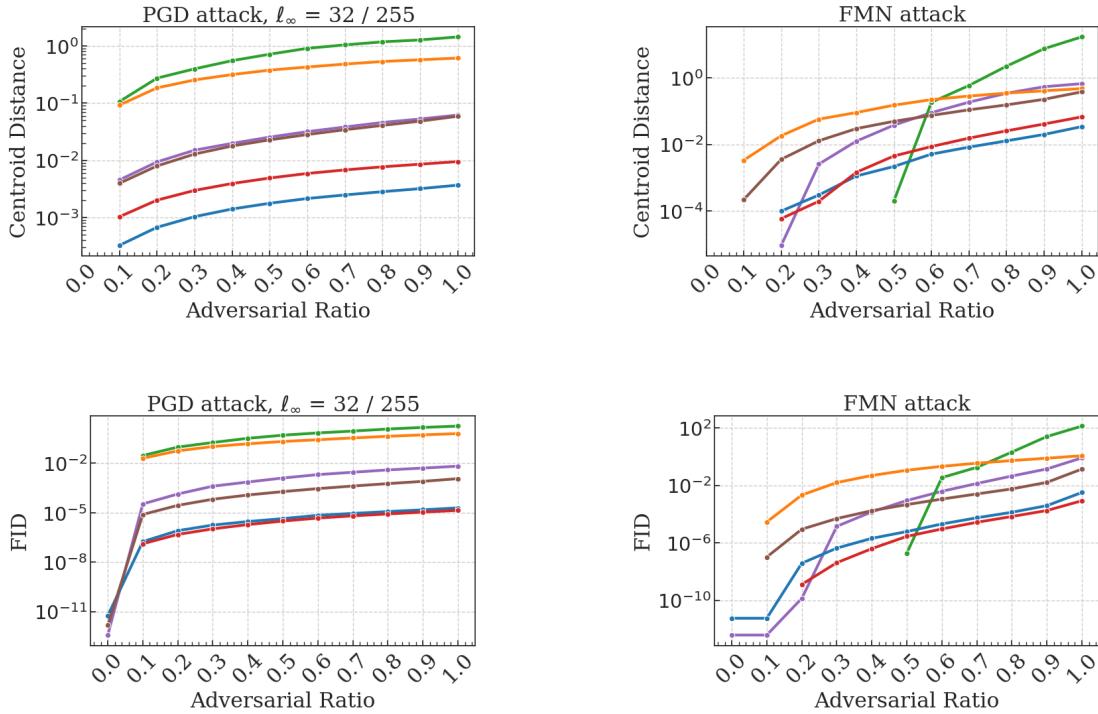


Figure B.15: Comparison of FMN and PGD attacks using feature-space-based distances.

## Appendix C

# Dataset reference

Experiments in the domain generalization setting for both robustness assessment and model selection purposes have been conducted with MNIST images under specific configurations of factors, by means of the DiagVib-6 data generation pipeline [20]. In general, two experimental frameworks can be defined, namely a control setting that replicates standard conditions in the absence of distribution shift and a covariate shift setting in which several image factors are manipulated. The synthetic nature of the dataset allows for a fine-grained control over image factors and thus the isolation of the different sources of randomness to which models are exposed.

In the first case, datasets  $\mathbf{x}'$  and  $\mathbf{x}''$  are exclusively subject to sampling randomness, in the sense that each observation in  $\mathbf{x}'$  is paired with an observation in  $\mathbf{x}''$  that belongs to the same class and has the same configuration of factors but represents a different MNIST instance. In the second case, a single MNIST instantiation is used to generate both  $\mathbf{x}'$  and  $\mathbf{x}''$ , each with a different configuration of factors, thereby removing any contribution of sampling randomness from the robustness score.

**Definition** (DiagVib-6 experiments). The classification task involves the prediction of the shape factor (i.e. the digit) of handwritten fours and nines from the MNIST dataset. Following the notation introduced in Chapter 3, we can define source and target domains as follows:

$$\begin{aligned}\mathcal{S} &= \{X_0, X_1\}, \\ \mathcal{T} &= \{X_1, X_2, X_3, X_4, X_5\},\end{aligned}$$

where  $X_j$  represents the random variable associated to domain  $j$ , being  $j$  the number of shifted factors with respect to domain  $X_0$ . Ideally, domain generalization requires datasets sampled from target domains to be used exclusively for testing, while training, validation and model selection are performed on source domains. Even if a more flexible approach has been considered for model selection, performance results are provided on test datasets only:

$$D_j^{\text{test}} = \{\mathbf{x}_j^{\text{test}}\}, \text{ where } \mathbf{x}_j^{\text{test}} := \mathbf{x}_j^{\text{test}} \circ \tau^{\text{test}}, j = 1, \dots, 5$$

and  $\tau^{\text{test}}$  represents a single sampling experiment. In this way, difference in performance can only be attributed to a lack of generalization capabilities under distribution shift. Regarding the generation of training and validation datasets, two main experimental settings have been considered.

- In the control setting, each dataset is generated from a different sampling realizations in disjoint subsets of MNIST, thereby avoiding repetition of instances.

$$\begin{aligned}D^{\text{train}} &= \{\mathbf{x}_0^{\text{train}}, \mathbf{x}_1^{\text{train}}\}, \text{ where } \mathbf{x}_j := \mathbf{x}_j \circ \tau_j^{\text{train}}, j = 0, 1 \\ D^{\text{val}} &= \{\mathbf{x}_0^{\text{val}}, \mathbf{x}_1^{\text{val}}\}, \text{ where } \mathbf{x}_j := \mathbf{x}_j \circ \tau_j^{\text{val}}, j = 0, 1\end{aligned}$$

where  $\tau_0^{\text{train}} \neq \tau_1^{\text{train}}$  and  $\tau_0^{\text{val}} \neq \tau_1^{\text{val}}$ .

- In the shifted setting, a single sampling realization is considered during validation, which allows robustness assessment to be performed under distribution shift only.

$$\begin{aligned} D^{\text{train}} &= \{\mathbf{x}_0^{\text{train}}, \mathbf{x}_1^{\text{train}}\}, \text{ where } \mathbf{x}_j := \mathbf{x}_j \circ \tau_j^{\text{train}}, j = 0, 1 \\ D^{\text{val}} &= \{\mathbf{x}_0^{\text{val}}, \mathbf{x}_1^{\text{val}}\}, \text{ where } \mathbf{x}_j := \mathbf{x}_j \circ \tau_j^{\text{val}}, j = 0, 1 \end{aligned}$$

where  $\tau_0^{\text{train}} \neq \tau_1^{\text{train}}$  and  $\tau_0^{\text{val}} = \tau_1^{\text{val}} = \tau^{\text{val}}$ .

four different realizations of the experiment, namely  $\tau_0^{\text{train}}$ ,  $\tau_1^{\text{train}}$ ,  $\tau^{\text{val}}$  and  $\tau^{\text{test}}$ , each sampling from disjoint subsets of MNIST.

This section of the appendix will serve as a reference for the specific dataset configurations that have been considered in the experiments.

## C.1 Robustness assessment

**Dataset** (Robustness assessment). The configuration of factors used for experiments in Section 4.3 considers only incrementally shifted observations.

	#0	#1	#2	#3	#4	#5
Hue	red	<b>blue</b>	blue	blue	blue	blue
Lightness	dark	dark	<b>bright</b>	bright	bright	bright
Position	CC	CC	CC	<b>LC</b>	LC	LC
Scale	normal	normal	normal	normal	<b>large</b>	large
Texture	blank	blank	blank	blank	blank	<b>tiles</b>
Shape	4,9	4,9	4,9	4,9	4,9	4,9

Table C.1: Image factors associated to each of the environments considered in this experiment. CC and LC account for 'centered center' and 'centered low', respectively.

## C.2 Model selection

### C.2.1 Controlled conditions

Section 5.1 reported test accuracy results for different configurations of pairs of datasets that were considered for robust model selection through early stopping. Test datasets entail a single sampling realization and represent incremental shifts of image factors.

**Dataset** (Test hue). Some experiments in Section 5.1 considered source domains to be shifted in the hue factor. Consequently, target domains entailed incremental shifts in the remaining factors.

	#0	#1	#2	#3	#4	#5
Hue	red	<b>green</b>	green	green	green	green
Lightness	dark	dark	<b>bright</b>	bright	bright	bright
Position	CC	CC	CC	CC	<b>UL</b>	UL
Scale	large	large	large	large	large	<b>small</b>
Texture	blank	blank	blank	<b>tiles</b>	tiles	tiles
Shape	1,4,7,9	1,4,7,9	1,4,7,9	1,4,7,9	1,4,7,9	1,4,7,9

Table C.2: Image factors associated to each of the environments considered in this experiment. CC and UL account for 'centered center' and 'upper left', respectively.

**Dataset (Test position).** Some experiments in Section 5.1 considered source domains to be shifted in the **position** factor. Consequently, target domains entailed incremental shifts in the remaining factors.

	#0	#1	#2	#3	#4	#5
Hue	green	green	green	<b>red</b>	green	green
Lightness	dark	dark	dark	dark	<b>bright</b>	bright
Position	UL	<b>CC</b>	CC	CC	CC	CC
Scale	large	large	<b>small</b>	small	small	small
Texture	blank	blank	blank	blank	blank	<b>tiles</b>
<i>Shape</i>	<i>1,4,7,9</i>	<i>1,4,7,9</i>	<i>1,4,7,9</i>	<i>1,4,7,9</i>	<i>1,4,7,9</i>	<i>1,4,7,9</i>

Table C.3: Image factors associated to each of the environments considered in this experiment. CC and UL account for ‘centered center’ and ‘upper left’, respectively.

**Dataset (Training and validation hue).** Some experiments in Section 5.1 considered source domains to be shifted in the **hue** factor.

	Env.	Hue	Lightness	Position	Scale	Texture	<i>Shape</i>
Training	0	red	dark	CC	large	blank	<i>1,4,7,9</i>
	1	<b>blue</b>	dark	CC	large	blank	<i>1,4,7,9</i>
Validation	0	red	dark	CC	large	blank	<i>1,4,7,9</i>
SD	1	red	dark	CC	large	blank	<i>1,4,7,9</i>
ID	1	<b>blue</b>	dark	CC	large	blank	<i>1,4,7,9</i>
1F-MD	1	<b>magenta</b>	dark	CC	large	blank	<i>1,4,7,9</i>
5F-MD	1	green	<b>bright</b>	<b>UL</b>	<b>small</b>	<b>tiles</b>	<i>1,4,7,9</i>
Validation OOD	0	<b>yellow</b>	dark	CC	large	blank	<i>1,4,7,9</i>
	1	<b>magenta</b>	dark	CC	large	blank	<i>1,4,7,9</i>

Table C.4: Image factors associated with each of the environments considered in this experiment. CC and UL account for ‘centered center’ and ‘upper left’, respectively.

**Dataset (Training and validation position).** Some experiments in Section 5.1 considered source domains to be shifted in the **position** factor.

	Env.	Hue	Lightness	Position	Scale	Texture	<i>Shape</i>
Training	0	green	dark	UL	large	blank	<i>1,4,7,9</i>
	0	green	dark	LR	large	blank	<i>1,4,7,9</i>
Validation	0	green	dark	UL	large	blank	<i>1,4,7,9</i>
SD	1	green	dark	UL	large	blank	<i>1,4,7,9</i>
ID	1	green	dark	<b>LR</b>	large	blank	<i>1,4,7,9</i>
1F-MD	1	green	dark	<b>UR</b>	large	blank	<i>1,4,7,9</i>
5F-MD	1	red	<b>bright</b>	<b>CC</b>	<b>small</b>	<b>tiles</b>	<i>1,4,7,9</i>
Validation OOD	0	green	dark	<b>UR</b>	large	blank	<i>1,4,7,9</i>
	1	green	dark	<b>LL</b>	large	blank	<i>1,4,7,9</i>

Table C.5: Image factors associated with each of the environments considered in this experiment. UL, LR, UR and CC account for ‘upper left’, ‘lower right’, ‘upper left’ and ‘centered center’, respectively.

### C.2.2 ID model selection

Model selection experiments have been conducted under controlled conditions, in which test performance under different robust model selection strategies was reported also for a **control** validation dataset pair, in which:

**Definition** (ID setting under ZGO). Let the predicted factor  $F^P$  be **shape** (i.e. the image digits), and let the learned factor  $F^L$  be **hue** and **position**, each for a different dataset. The zero shortcut opportunity datasets are obtained as follows:

$\times$ Training, Validation & Test #0			$\circ$ Test #1 - #5				
Source 1			Source 2				
Shape 7	$\times$	$\circ$	$\circ$	Shape 7	$\times$	$\circ$	$\circ$
4	$\circ$	$\times$	$\circ$	4	$\circ$	$\times$	$\circ$
9	$\circ$	$\circ$	$\times$	9	$\circ$	$\circ$	$\times$
B	R	G	Y	C	M		
Hue			Hue				
Position			Position				
UL	CC	LR	LC	UR	CL		

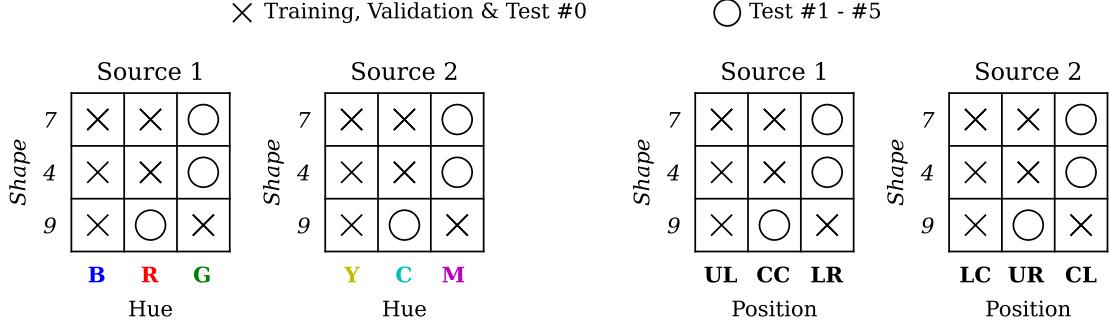
**Definition** (ID setting under 1-CGO). Let the predicted factor  $F^P$  be **shape** (i.e. the image digits), and let the learned factor  $F^L$  be **hue** and **position**, each for a different dataset. The single compositional generalization opportunity datasets are obtained as follows:

$\times$ Training, Validation & Test #0			$\circ$ Test #1 - #5				
Source 1			Source 2				
Shape 7	$\times$	$\times$	$\circ$	Shape 7	$\times$	$\times$	$\circ$
4	$\circ$	$\times$	$\circ$	4	$\circ$	$\times$	$\circ$
9	$\circ$	$\circ$	$\times$	9	$\circ$	$\circ$	$\times$
B	R	G	Y	C	M		
Hue			Hue				
Position			Position				
UL	CC	LR	LC	UR	CL		

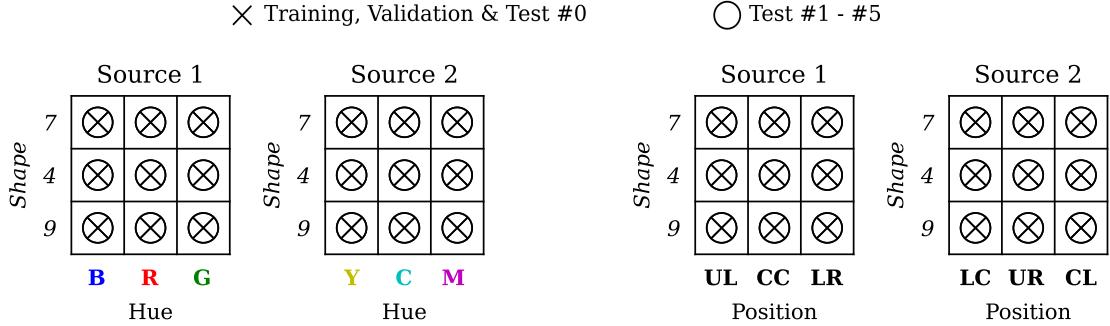
**Definition** (ID setting under 2-CGO). Let the predicted factor  $F^P$  be **shape** (i.e. the image digits), and let the learned factor  $F^L$  be **hue** and **position**, each for a different dataset. The double compositional generalization opportunity datasets are obtained as follows:

$\times$ Training, Validation & Test #0			$\circ$ Test #1 - #5				
Source 1			Source 2				
Shape 7	$\times$	$\times$	$\circ$	Shape 7	$\times$	$\times$	$\circ$
4	$\circ$	$\times$	$\circ$	4	$\circ$	$\times$	$\circ$
9	$\times$	$\circ$	$\times$	9	$\times$	$\circ$	$\times$
B	R	G	Y	C	M		
Hue			Hue				
Position			Position				
UL	CC	LR	LC	UR	CL		

**Definition** (ID setting under 3-CGO). Let the predicted factor  $F^P$  be **shape** (i.e. the image digits), and let the learned factor  $F^L$  be **hue** and **position**, each for a different dataset. The triple compositional generalization opportunity datasets are obtained as follows:



**Definition** (ID setting under ZSO). Let the predicted factor  $F^P$  be **shape** (i.e. the image digits), and let the learned factor  $F^L$  be **hue** and **position**, each for a different dataset. The zero shortcut opportunity datasets are obtained as follows:



Models trained and selected using each of the previous datasets will be evaluated on a test dataset replicating validation conditions (i.e. #0) and five additional test datasets exploiting the shortcut opportunities present in the data, so that the quality of the inductive bias of the selected model can be assessed. Each test dataset is shifted progressively in the factors that were not considered during training and validation.

hue	#0	#1	#2	#3	#4	#5
Position	CC	CC	<b>CR</b>	CR	CR	CR
Scale	large	large	large	<b>small</b>	small	small
Brightness	bright	bright	bright	bright	<b>dark</b>	dark
Texture	tiles	tiles	tiles	tiles	tiles	<b>wood</b>

position	#0	#1	#2	#3	#4	#5
Scale	large	large	<b>small</b>	small	small	small
Hue	blue	blue	blue	<b>red</b>	red	red
Brightness	bright	bright	bright	bright	<b>dark</b>	dark
Texture	tiles	tiles	tiles	tiles	tiles	<b>wood</b>

Table C.6: Image factors associated to each of the environments considered in test datasets, excluding the predicted and learned factors in which the shortcut or generalization opportunities are encoded. CC and CR stand for 'center center' and 'upper left', respectively.

# Bibliography

- [1] Sravanti Addepalli, Samyak Jain, Gaurang Sriramanan, and R. Venkatesh Babu. Scaling adversarial training to large perturbation bounds. In *European Conference on Computer Vision*, 2022.
- [2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization.
- [3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN.
- [4] Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, volume 80, pages 274–283. PMLR, 2018.
- [5] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent Advances in Adversarial Training for Adversarial Robustness.
- [6] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from Several Related Classification Tasks to a New Unlabeled Sample.
- [7] Anton Bovier. *Statistical Mechanics of Disordered Systems: A Mathematical Perspective*. Cambridge University Press.
- [8] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press.
- [9] Joachim M. Buhmann. Data Science Algorithms and the Rate-Distortion Tradeoff.
- [10] Joachim M. Buhmann. Information theoretic model validation for clustering.
- [11] Joachim M. Buhmann. Posterior Agreement for Model Robustness Assessment in Covariate Shift Scenarios.
- [12] Joachim M Buhmann, Morteza Haghir Chehreghani, Mario Frank, and Andreas P Streich. Information Theoretic Model Selection for Pattern Analysis.
- [13] Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks.
- [14] George Casella and Roger L. Berger. *Statistical Inference*. Wadsworth Group Duxbury, second edition.
- [15] Morteza Haghir Chehreghani, Alberto Giovanni Busetto, and Joachim M Buhmann. Information Theoretic Model Validation for Spectral Clustering.
- [16] Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified Adversarial Robustness via Randomized Smoothing.
- [17] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: A standardized adversarial robustness benchmark.

- [18] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E. Kounavis, and Duen Horng Chau. Keeping the Bad Guys Out: Protecting and Vaccinating Deep Learning with JPEG Compression.
- [19] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations, 2019.
- [20] Elias Eulig, Piyapat Saranrittichai, Chaithanya Kumar Mummadli, Kilian Rambach, William Beluch, Xiahan Shi, and Volker Fischer. DiagViB-6: A Diagnostic Benchmark Suite for Vision Models in the Presence of Shortcut and Generalization Opportunities.
- [21] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples.
- [22] Peter Grünwald and Teemu Roos. Minimum Description Length Revisited. 11(01):1930001.
- [23] Jun Guo, Wei Bao, Jiakai Wang, Yuqing Ma, Xinghai Gao, Gang Xiao, Aishan Liu, Jian Dong, Xianglong Liu, and Wenjun Wu. A comprehensive evaluation framework for deep model robustness. 137:109308.
- [24] Allan Gut. *An Intermediate Course on Probability*. Springer, second edition.
- [25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, pages 770–778. IEEE Computer Society, 2016.
- [26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models.
- [27] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial Examples Are Not Bugs, They Are Features.
- [28] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, first edition.
- [29] Ortiz Jimenez. The inductive bias of deep learning: Connecting weights and functions.
- [30] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A. Efros, and Antonio Torralba. Undoing the Damage of Dataset Bias. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, volume 7572, pages 158–171. Springer Berlin Heidelberg.
- [31] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization.
- [32] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A Benchmark of in-the-Wild Distribution Shifts.
- [33] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images.
- [34] Yann LeCun, Corinna Cortes, and Christopher J.C. Burges. The MNIST database of handwritten digits. <http://yann.lecun.com/exdb/mnist>, 1998. Accessed: 2024-09-07.
- [35] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to Generalize: Meta-Learning for Domain Generalization. 32(1).
- [36] Yao Li, Minhao Cheng, Cho-Jui Hsieh, and Thomas C. M. Lee. A Review of Adversarial Attack and Defense for Classification Methods. 76(4):329–345.

- [37] Jian Liang, Ran He, and Tieniu Tan. A Comprehensive Survey on Test-Time Adaptation under Distribution Shifts.
- [38] Jiashuo Liu, Zheyuan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards Out-Of-Distribution Generalization: A Survey.
- [39] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer.
- [40] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, 2011.
- [41] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks.
- [42] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning.
- [43] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain Generalization via Invariant Feature Representation.
- [44] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Second edition.
- [45] Kevin P. Murphy. *Probabilistic Machine Learning. An Introduction*. The MIT Press.
- [46] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-Adversarial Domain Adaptation.
- [47] Maura Pintor, Fabio Roli, Wieland Brendel, and Battista Biggio. Fast Minimum-norm Adversarial Attacks through Adaptive Norm Constraints.
- [48] Joaquin Quiñonero-Candela, editor. *Dataset Shift in Machine Learning*. Neural Information Processing Series. MIT Press.
- [49] Sebastian Ruder. An overview of gradient descent optimization algorithms.
- [50] David E Rumelhart, Geoffrey E Hintont, and Ronald J Williams. Learning representations by back-propagating errors.
- [51] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*. Association for Computational Linguistics, 2019.
- [52] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Mądry. Adversarially Robust Generalization Requires More Data.
- [53] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein Distance Guided Representation Learning for Domain Adaptation.
- [54] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition.
- [55] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks.
- [56] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE.
- [57] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness May Be at Odds with Accuracy.

- [58] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep Learning for Computer Vision: A Brief Review. 2018:1–13.
- [59] Chengyu Wang, Minghui Qiu, Jun Huang, and Xiaofeng He. Meta Fine-Tuning Neural Language Models for Multi-Domain Text Mining.
- [60] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. Generalizing to Unseen Domains: A Survey on Domain Generalization.
- [61] Yang Wang, Bo Dong, Ke Xu, Haiyin Piao, Yufei Ding, Baocai Yin, and Xin Yang. A Geometrical Approach to Evaluate the Adversarial Robustness of Deep Neural Networks. 19:1–17.
- [62] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better Diffusion Models Further Improve Adversarial Training.
- [63] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training, 2023.
- [64] Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach.
- [65] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020.
- [66] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating Adversarial Examples with Adversarial Networks.
- [67] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving Out-of-Distribution Robustness via Selective Augmentation.
- [68] Samuel Yu, Peter Wu, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. PACS: A Dataset for Physical Audiovisual CommonSense Reasoning.
- [69] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features.
- [70] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016.
- [71] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically Principled Trade-off between Robustness and Accuracy.
- [72] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond Empirical Risk Minimization.
- [73] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain Generalization: A Survey. pages 1–20.



Eidgenössische Technische Hochschule Zürich  
Swiss Federal Institute of Technology Zurich

Institute for Dynamic Systems and Control  
Prof. Dr. R. D'Andrea, Prof. Dr. L. Guzzella

**Title of work:**

Improved robustness of deep learning models through posterior agreement based model selection

**Thesis type and date:**

Master Thesis, September 2024

**Supervision:**

**Student:**

Name: ~~enzoRJ~~ Jimenez  
E-mail: vjimenez@student.ethz.ch  
Legi-Nr.: 97-906-739  
Semester: 5

**Statement regarding plagiarism:**

By signing this statement, I affirm that I have read and signed the Declaration of Originality, independently produced this paper, and adhered to the general practice of source citation in this subject-area.

Declaration of Originality:

[http://www.ethz.ch/faculty/exams/plagiarism/confirmation\\_en.pdf](http://www.ethz.ch/faculty/exams/plagiarism/confirmation_en.pdf)

Zurich, 8. 9. 2024: \_\_\_\_\_