ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

DINFK
Institute for Machine Learning

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 work 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.

# Acknowledgements

First and foremost, I would like to express my deepest gratitude to Prof. Joachim M. Buhmann for granting me the opportunity to join the ISE group. I must specifically acknowledge his original work on Posterior Agreement, which constitutes the cornerstone of the ideas and methods explored in this thesis. The first two chapters rely heavily on his foundational contributions, and it is impossible to fully do justice to the breadth of his work within the text. I would also like to thank him for his guidance, as well as for the many interesting insights and conversations we shared. As he embarks on his well-deserved retirement, I wish him all the best for the future.

I am equally grateful to Dr. João Carvalho, whose mentorship has greatly contributed to my growth, both academically and personally. I sincerely appreciate the opportunity he gave me to join this project, as well as for involving me in other exciting and interesting endeavors. His valuable insights have had a profound impact on the work presented here and beyond. I am truly thankful for everything he has provided, and I will carry the lessons learned from him throughout my life.

My sincere thanks also go to Dr. Alessandro Torcinovich, whose advice and experience have consistently helped refine and improve my research skills. His thoughtful feedback has greatly enhanced the quality of this thesis and my future work as well. I truly appreciate the time he dedicated to guiding me and offering insightful ideas that helped further develop this project.

Finally, I would like to extend my heartfelt thanks to all the members of the ISE group, namely Eugene, Fabian, Ivan, Robin, Xia, Lukas, Ami, Alina and Rita. Your warm acceptance and collaboration made my time here not only productive but truly enjoyable. In one way or another, each of you helped create a stimulating research environment, and I am grateful for the many engaging conversations we shared. I leave with cherished memories of our time together and wish you all continued success in your future endeavors.

# Contents

# 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.

# 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 [47, 41]. 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 [50]. 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 [57]. 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 [59].

(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 [50].

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 [30]. 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 [59, 73]. 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 [28].

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 [57].

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 [22].

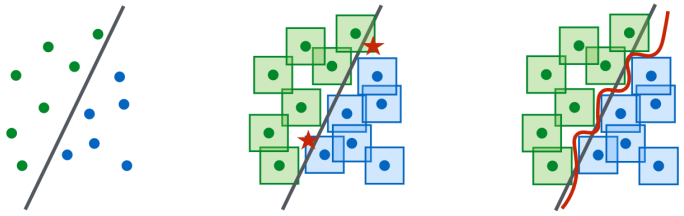
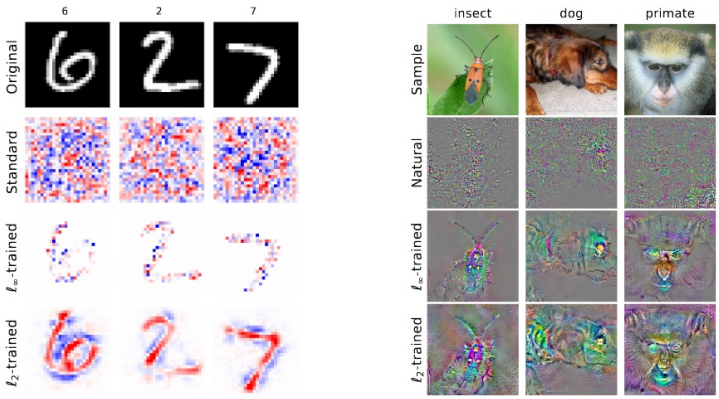


Figure 1.2: A conceptual illustration of standard vs. adversarial decision boundaries. (**left**) A set of linearly-separable points. (**middle**) Decision boundary learned via standard training. (**right**) Decision boundary learned via adversarial training. Both methods achieve zero training error, but only the robust model is able to generalize to $\ell_\infty$ perturbations. [43]

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 insensitive 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 [54]. These limitations express a fundamental trade-off that arises from the intrinsic disparity between robust and non-robust features [59, 73].

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 [28]. 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 [59].




(a) MNIST (b) Restricted ImageNet

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. [59]

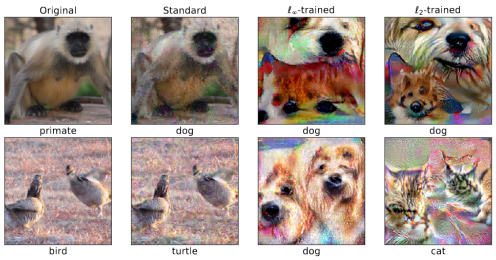



Figure 1.4: Adversarial examples for standard and adversarially-trained models. Perturbed images produced for robust models effectively capture salient data characteristics. [59]

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 [45, 50]. 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 [75, 62, 40].

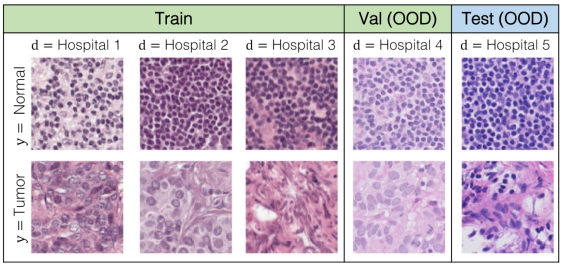



Figure 1.5: The `camelyon17` (WILDS) dataset comprises images of stained lymph node tissue patches sampled from different hospitals. [34]

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[1] 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.

[1]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.

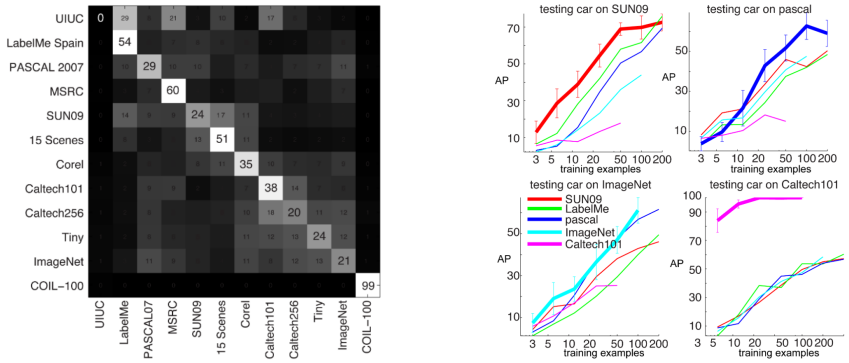



Figure 1.6: **(left)** Confusion matrix generated in a dataset identification task. A clearly pronounced diagonal indicates that each dataset possesses unique traits that make it distinguishable from the rest. **(right)** Cross-dataset generalization for `car` detection as function of training data. 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. [58]

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 [62]). 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 [62, 40].

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/generation.

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 [55, 39].

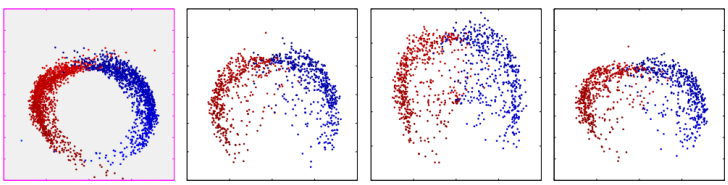


Figure 1.7: Projections of a binary synthetic dataset in the two principal DICA dimensions. The shaded box depicts the projection of training data, whereas the unshaded boxes show projections of unseen test datasets. [45]

(a) Original (b) Mixup [74] (c) CutMix [71]

Figure 1.8: Mixup and Cutmix strategies can be used to interpolate between different labels and/or domains by generating intermediate observations. [71]

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 [75]. 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 [34].

## 1.2 Related work

In the adversarial front, early work [57] unveiled the nature of the susceptibility of deep learning models to adversarial examples and FGSM [22] 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 [43], C&W [13], FMN [49] and many others (see [38] 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 [38]. Regarding adversarial training (i.e. defenses), robustness can be achieved by a variety of methods, such as ensemble learning, defensive distillation, generative adversarial networks [68, 44], diffusion models [64, 27] and adaptive-boundary methods [16]. In this project, the RobustBench attack library [17] will be leveraged to evaluate adversarial robustness in the CIFAR10 dataset [35].

In the domain generalization front, the existing rich taxonomy of methods can be classified into two main groups, namely data manipulation and representation learning [62, 75, 40]. First, data manipulation strategies involve both augmentation and generation, as for example image randomization or adversarial augmentation [69, 74, 71]. Second, representation learning strategies are

primarily divided into domain-invariant methods (e.g. model-based [2], kernel-based [45, 3] or adversarial-based [48]) 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 [37, 61].

Regarding robustness characterization, a wide range of metrics have been conceived (see [24]), but accuracy-based criteria are still the most common. As an alternative, some theoretically-grounded approaches have been proposed, such as CLEVER [66], ACTS [63] 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 [36] (and its multiple variations, such as DiagVib-6 [20]), PACS [70], VLCS [32] or WILDS [34].

## 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 [34] 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 [46]. In this formalism, the goal is to learn a target function $f^* : \mathcal{X} \longmapsto \mathcal{Y}$ 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 $\boldsymbol{X} = (X_1, ..., X_N)$ be a (simple) random sample of $X$ with size $N$ [14]. A supervised dataset $D$ is constructed from a realization $\boldsymbol{x} \sim \boldsymbol{X}$ 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 $\boldsymbol{X}$ will be denoted by $\mathcal{D}$.

**Definition** (*Empirical risk*)**.** Let $D$ be a supervised dataset generated from a sample $\boldsymbol{x} \sim \boldsymbol{X}$. Let $L : \mathcal{Y} \times \mathcal{Y} \longmapsto \mathbb{R}$ be a loss function measuring the discrepancy at observation $x \in \boldsymbol{x}$ 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 $R(f)$:

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

We can approximate the expected risk with its empirical (plug-in) analogous if $N$ is large enough (see Glivenko-Cantelli theorem [25]). The empirical risk of $f \in \mathcal{F}$ computed on $D$ is defined as

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

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 $\boldsymbol{x'} \sim \boldsymbol{X}$. 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 [46].

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 [30]. 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 $\Omega : \mathcal{F} \longmapsto \mathbb{R}$ 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 [56, 46, 60].

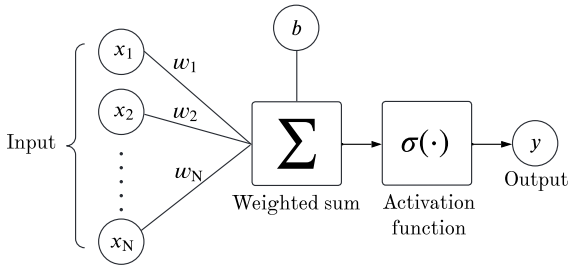



Figure 2.1: The output of a node is computed by applying a non-linear activation function $\sigma$ to the weighted sum of its inputs $\boldsymbol{x}$ plus a bias term $b$.

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

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

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:

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

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 arquitectures:

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

where $\Gamma$ is the parameter space navigated by this particular architecture. The function class $\mathcal{F}_\Gamma$ consists of all mappings $f(\cdot; \gamma) : \mathcal{X} \longmapsto \mathcal{Y}$ that 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 [52]. In practice, more efficient variations of gradient descent are used, such as stochastic gradient descent (SGD) [51] or Adam [33].

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 [30].

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. $\boldsymbol{x'}, \boldsymbol{x''} \sim \boldsymbol{X}$). 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 $\boldsymbol{X} \overset{\text{iid}}{\sim} X$ entails a measure of probability described by the density function $f_{\boldsymbol{X}}$:

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

where $f_X$ is the density function of $X$. We will use $\mathbf{P}_{\boldsymbol{x}}$ to refer to the empirical approximation of this distribution; that is, to the distribution of samples $\boldsymbol{x} \sim \boldsymbol{X}$:

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

**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 $\boldsymbol{x} \sim \boldsymbol{X}$ on experimental conditions will be captured by index $\tau$, and we will implicitly consider sample

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

to be a realization of the experiment $\boldsymbol{X} \overset{\text{iid}}{\sim} X$ 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 [36] 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 $\boldsymbol{x} \sim \boldsymbol{X}$ 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 $\boldsymbol{X}$. A data science algorithm learns a function $f$ implementing the following mapping:

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

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

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

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 $\boldsymbol{x} \sim \boldsymbol{X}$. We will refer to this distribution as the posterior over $\Theta$ under $f$:

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

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

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

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 $\boldsymbol{x}', \boldsymbol{x}'' \sim \boldsymbol{X}$ generated from the same experiment. Samples are independent in general and they only differ in the implicit randomness entailed by the sampling experiment:

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

**Proposition** (*Posterior selection principles*)**.** Let $\mathcal{T}_{\boldsymbol{X}} \subseteq \mathcal{T}$ be the subset of sampling realizations that is effectively navigated by $\boldsymbol{X}$, in the sense that it encompasses only those instantiations that can be sampled from $\boldsymbol{X}$. Let $\Theta_{\boldsymbol{X}} \subseteq \Theta$ be the realizable subset[1] of hypotheses, generated as a surjective transformation of $\mathcal{T}_{\boldsymbol{X}}$. 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 $\Theta_{\boldsymbol{X}}$.
- Equally likely inputs drawn from the same experiment should yield similar sets of hypothesis.

[1] In the context of supervised classification tasks, $\mathcal{T}_{\boldsymbol{X}} \subseteq \mathcal{T}$ and therefore $\Theta = \Theta_{\boldsymbol{X}}$, 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 [23]. The code length of the argument of such distribution is

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

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:

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

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:

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

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

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

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

**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.*

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

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

□

### 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. See illustration in Figure 2.2.

**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

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

**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.

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

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

**Theorem 2.3.1** (*Maximum posterior agreement*)**.** The optimal $\mathbf{P}_*^f$ 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.

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

The last inequality follows from the fact that the entropy does not exceed the log-cardinality of the hypothesis class, $H_{\mathbf{\Pi}^f}(\theta) \leq \log |\Theta|, \ \ \forall \mathbf{\Pi}^f$. □

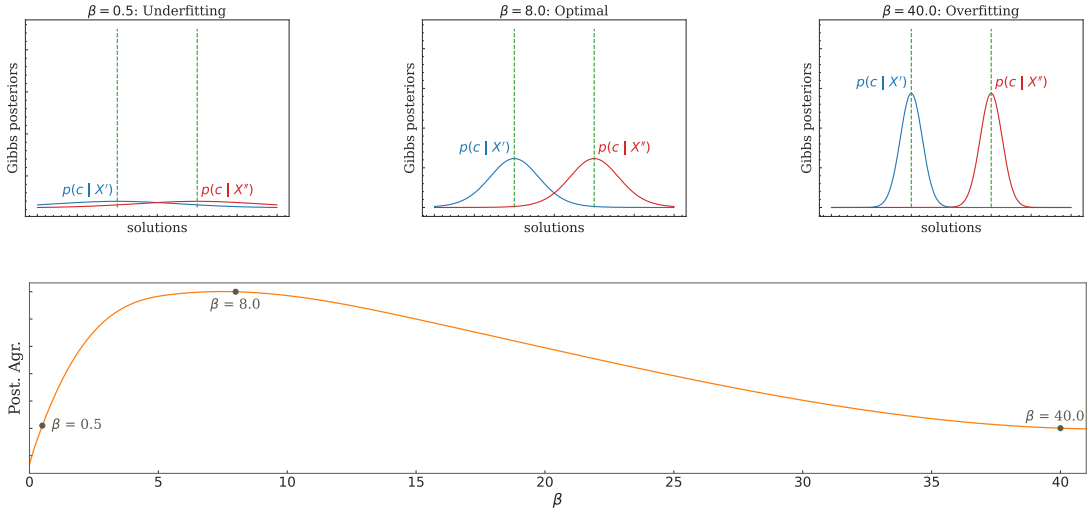



Figure 2.2: Qualitative illustration of the optimization over the inverse temperature parameter $\beta$. When $\beta \longrightarrow 0$, the informativeness of the posterior is reduced, eventually converging to a uniform distribution. When $\beta \longrightarrow \infty$, the informativeness of the posterior grows, leading to an increasingly peaked distribution. Posterior Agreement is maximum at a value $\beta^*$ in which hypothesis selected from the posterior over $\theta \mid \boldsymbol{x}'$ are assigned a high probability by the posterior over $\theta \mid \boldsymbol{x}''$, which in this case aligns with the maximum posterior overlap. [9]

# 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 $\boldsymbol{F}$ be a discriminant function, assigning a score to each of the $K$ classes.

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

- 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.

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

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

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

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

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 under covariate shift

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 $\boldsymbol{x}'$ and $\boldsymbol{x}''$ be two samples of $\boldsymbol{X} \overset{\text{iid}}{\sim} X$ with size $N$. A covariate shift exists between $\boldsymbol{x}'$ and $\boldsymbol{x}''$ if their (empirical) distributions are significantly different[1] for $N$ large enough. This is expressed as:

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

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 [50].

The presence of covariate shift as defined above already leads to a non-zero generalization error, given that $\boldsymbol{x}'$ and $\boldsymbol{x}''$ 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. $\mathcal{T}_{\mathbf{X}'} \neq \mathcal{T}_{\mathbf{X}''}$). In such case

$$\boldsymbol{x}' \sim \boldsymbol{X}' \overset{\text{iid}}{\sim} X' \text{ and } \boldsymbol{x}'' \sim \boldsymbol{X}'' \overset{\text{iid}}{\sim} X''$$

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 [50]. In simpler terms, $X' \neq X'' \implies \mathbf{P}_{\boldsymbol{x}'} \not\sim \mathbf{P}_{\boldsymbol{x}''}$ in general.

In the out-of-distribution setting, $\boldsymbol{x}'$ and $\boldsymbol{x}''$ 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 $\mathbf{P}_{\boldsymbol{x}'}$ and $\mathbf{P}_{\boldsymbol{x}''}$ will be different in general and induce a covariate shift that leads to a non-zero generalization error.

[1] 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. [50]

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

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

which induces a covariate shift known as adversarial, given that perturbation $\boldsymbol{\Delta}$ 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 $\boldsymbol{x}'$ and $\boldsymbol{x}''$ 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 $\boldsymbol{x} \sim \boldsymbol{X}$ 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 $\boldsymbol{x}', \boldsymbol{x}'' \sim \boldsymbol{X}$, respectively. A robustness metric is a function $\Omega : \mathcal{D}'' \times \mathcal{F} \longmapsto \mathbb{R}$ 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 $D''$:

$$\text{Accuracy} = [100\times] \, \frac{1}{N} \sum_{n=1}^{N} \delta_{y_n''} \left( \hat{c}(x_n'') \right), \; (x_n'', y_n'') \in D''.$$

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 every observation in the dataset.

**Definition** (*Perturbation*)**.** Let $\boldsymbol{x}'$ be a realization of $\boldsymbol{X} \overset{\text{iid}}{\sim} X$ with support $\mathcal{X} \subset \mathbb{R}^d$. Let $x \in \boldsymbol{x}'$ 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 $\boldsymbol{\Delta}$ will be $\epsilon_p$-constrained if each of its components satisfies the previous definition. In such case,

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

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 [57]. 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 [22].

**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

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

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 [22].

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 [22]. 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:

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

where $\Pi$ is the projection operator. This is known as projected gradient descent attack [43].

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 [43].

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:

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

This is known as the fast minimum-norm attack [49].

## 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 $\mathbb{S} = \{X_s^1, \cdots, X_s^{|\mathbb{S}|}\}$ and $\mathbb{T} = \{X_t^1, \cdots, X_t^{|\mathbb{T}|}\}$ 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) [40, 62].

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 [24]), 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 [75].

## 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 [40].

**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]:

$$c^* = \min_{c=\eta\circ\boldsymbol{F}\circ\Phi} \sum_{d\in\mathbb{S}} R^d(c)$$
$$\text{s.t. } (\eta \circ \boldsymbol{F}) = \arg\min_{\bar{c}} R^d(\bar{c}) \; \forall d \in \mathbb{S}.$$

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 [74] and CutMix [71].

**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 [69].

## 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 $N$ observations 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 $\boldsymbol{x} \sim \boldsymbol{X}$. 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:

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

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 $\mathfrak{P}^c$ to consider is the Gibbs distribution with inverse temperature parameter $\beta$ [11]

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

*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 [29]. Testable information amounts to certain constraints on the MEP optimization problem over the non-negative, Lebesgue-integrable function class $\mathcal{P}$.

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

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

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

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

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

which has a unique solution

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

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

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,

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

*Proof.* See Appendix A.1. □

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

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

*Proof.* See Appendix A.1. □

**Theorem 3.6.2** (*PA kernel for classification*)**.** Let $\boldsymbol{x}'$ and $\boldsymbol{x}''$ be $N$-sized realizations of $\boldsymbol{X}$. 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:

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

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

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

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

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

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

$$\mathbb{E}_{\mathbf{P}_{\boldsymbol{x}',\boldsymbol{x}''}} = \frac{1}{N} \sum_{n=1}^{N} \cdot$$

$$\mathbb{E}_{\mathbf{P}^c(\theta|\boldsymbol{x}')} = \sum_{\theta \in \Theta} \mathbf{P}^c(\theta \mid \boldsymbol{x}') \cdot$$

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

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

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

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

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

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

Finally, applying Lemma 3.6.1 to the product inside the logarithm, we reach the final expression. □

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 $\mathrm{PA}(\boldsymbol{x}', \boldsymbol{x}''; \beta)$ has the following properties $\forall \boldsymbol{x}', \boldsymbol{x}'' \sim \boldsymbol{X}$ and $\beta \in \mathbb{R}^+$.

**P1** (Boundedness) $\mathrm{PA}(\boldsymbol{x}', \boldsymbol{x}''; \beta) \leq N \log K$. 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) $\mathrm{PA}(\boldsymbol{x}', \boldsymbol{x}''; \beta) = \mathrm{PA}(\boldsymbol{x}'', \boldsymbol{x}'; \beta)$. This property is important from the robustness perspective, given that randomness instantiations are not indexed and no reference experiment can be performed.

**P3** (Concavity) $\mathrm{PA}(\boldsymbol{x}', \boldsymbol{x}''; \beta)$ 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. □

Stemming from the derived expression, an operative version of the posterior agreement kernel for the $K$-class classification problem is given by

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

where factors $|\Theta|$ and $1/N$ have not been considered as they are merely for scale, making kernel values now range within $(-\infty, 0]$. Robustness assessment and model selection results will be reported through the maximum posterior agreement, which is denoted as

$$\mathrm{PA} \equiv \mathrm{PA}(\boldsymbol{x}', \boldsymbol{x}''; \beta^*)$$

to avoid notation clutter, where $\beta^*$ is obtained via Adam optimization. The optimization process considers a learning rate of 0.1, with the initial value set at $\beta_0 = 1$. Since the cost function driving posterior selection is the classifier's discriminant function itself, the initial posterior considered during optimization is the actual predictive output of the model. The evolution of $\beta$ will re-weight these predictions to maximize the mutual information for $\boldsymbol{x}'$ and $\boldsymbol{x}''$. These optimization hyperparameters have been found to be reliable in any setting, as the optimal $\beta^*$ converges to the same value in all cases (see for instance Figure 4.4), which is consistent with the concavity property of the PA kernel. In the adversarial setting, optimization will be carried out over 500 epochs, while in the domain generalization and model selection settings, it will extend to 1000 epochs.

As an intuitive approach, the PA kernel can be interpreted as an aggregation of robustness contributions across all observations. During the optimization of $\beta$, the informativeness-stability trade-off is navigated towards the maximum, at which the optimal posterior for each observation encodes its estimated information content, consistent with the observed stability of $\theta \mid \boldsymbol{x}'$ against $\theta \mid \boldsymbol{x}''$. When predictions match, their robustness contribution increases with the informativeness of the posterior, because a higher probability is assigned to the predicted class. Conversely, when predictions do not match, their contribution is maximized as the posterior becomes more spread-out and less informative. Given that the match or mismatch in predicted label is the same for all $0 < \beta < \infty$, as it corresponds to the maximum of the posterior, the increase or decrease of contributing terms is monotonically increasing or decreasing with $\beta$, respectively. Two clear extreme situations emerge, namely when none of the predictions match, and when all predictions match.

**Theorem 3.6.5** (*Boundedness of* PA)**.** The PA value is bounded by the two extreme robust and unrobust situations.

$$-N \log K \leq \text{PA} \leq 0$$

*Proof.* When all predictions match, $\beta^* \longrightarrow \infty$ and posteriors converge to a delta distribution centered at the predicted class $\hat{y}_n$, which is the same for both samples.

$$\text{PA} = \sum_{n=1}^{N} \log \left\{ 1 + \sum_{k \neq \hat{y}_n} 0 \right\} = 0.$$

When none of the predictions match, $\beta^* \longrightarrow 0$ and posteriors converge to a uniform distribution.

$$\text{PA} = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} \frac{1}{K^2} \right\} = -N \log K.$$

□

Even if these extreme cases will be considered through synthetic experiments, real settings will comprise partially unrobust models. In such cases, the model achieving the highest PA unrobustness penalization will be selected.

**Proposition.** PA favors classifiers that exhibit desirable inductive biases and align with the common understanding of what constitutes a good model.

The proposition follows from the aforementioned interpretation and the way robustness contributions evolve with respect to the informativeness of the posterior. Common sense dictates that good models should assign high probabilities to the correct classes and low probabilities to the incorrect ones, regardless of their performance[2]. This is because desirable inductive biases should be constructed around sets of features that are predictive of the task at hand, and models should be able to extract those features from the data.

In this context, PA selects models in which matching pairs of observations are assigned high probabilities, while non-matching observations are assigned low probabilities. In the first case, the informativeness of posteriors at $\beta_0 = 1$ is already high and yields a high robustness contribution, thus avoiding $\beta$ from being pushed to very high values. If a mismatching observation $x''$ is assigned a low probability, the overall contribution from the pair $x', x''$ will not be as significantly decreased due to the small value of $\beta$. At the same time, these contributions will not drive $\beta$ to significantly low values, because their informativeness at $\beta_0 = 1$ will already be low.

### 3.6.3 Implementation

The goal of this project is to use PA 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 of the kernel optimization task has been developed with the purpose of this project. The implementation is integrated within the Torchmetrics[3] framework, facilitating its incorporation into any machine learning workflow. A brief code snippet demonstrating a simple PA evaluation is provided on the following page.

[2]Two models could achieve the same task performance but differ vastly in the informativeness of their posteriors.
[3]https://lightning.ai/docs/torchmetrics/stable/

The numerical implementation of the PA metric offers several notable features that enhance its functionality and flexibility. First, the metric initiates a single optimization process with a pair of datasets, but additional datasets can be incorporated for validation purposes, assuming that predictions will not vary very significantly between them. This implementation also supports the use of different models, which is particularly advantageous in cross-validation settings. If data samples across datasets are not corresponding, various pairing strategies can be used, including label matching, nearest-neighbor matching, or canonical correlation matching, if the feature extractor of the classifier is accessible.

Additionally, the implementation supports multi-device computation, allowing for parallelization of model evaluation through a distributed-data-processing (DDP) strategy on a CUDA-managed set of GPUs. This capability improves efficiency, especially when handling large datasets. The output of the model is computed only once per training epoch, so that predicted posteriors are stored and reused without having to continuously access the data, thus significantly reducing the time required for kernel optimization. Finally, the metric provides detailed logging and retrieval of information about the optimization process, facilitating better tuning of both the optimization algorithm and related 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[4].

```
from pametric import PosteriorAgreementBase, LogitsDataset

pa_metric = PosteriorAgreementBase(pa_epochs, beta_0)
results = pa_metric(
    LogitsDataset([logits0, logits1], y)
)

logPA, beta = results["logPA"], results["beta"]
```

[4]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 behavior 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 randomly generated, thus remaining agnostic of the model and the learning task.

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 $\boldsymbol{y} \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$ 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 $\hat{\boldsymbol{y}}', \hat{\boldsymbol{y}}''$ are considered.

- For a perfect classifier, predictions are $\hat{\boldsymbol{y}}' = \hat{\boldsymbol{y}}'' = \boldsymbol{y}$.
- For a constant classifier, predictions are $\hat{\boldsymbol{y}}' = \hat{\boldsymbol{y}}'' = \mathbf{0}$.
- For a random classifier, predictions $\hat{\boldsymbol{y}}', \hat{\boldsymbol{y}}''$ are generated by randomly permuting $\boldsymbol{y}$, 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. In particular,

$$F_{\hat{y}} = \frac{\Delta}{2}, \text{ and } F_{1-\hat{y}} = -\frac{\Delta}{2}, \text{ for } \hat{y} = \{\hat{y}', \hat{y}''\}.$$

In this way, $\boldsymbol{F}$ discriminates the predicted and non-predicted class by $\Delta \in \mathbb{R}$.

| Classifier | $\beta^*$ | **PA** | Accuracy |
|---|---|---|---|
| Perfect | 12.33 | **-0.0082** | 1.000 |
| Constant | 12.33 | **-0.0082** | 0.525 |
| Random | 0.000 | **-693.14** | 0.516 |

Table 4.1: Comparison of accuracy and PA values for the case $p = 1/2$.

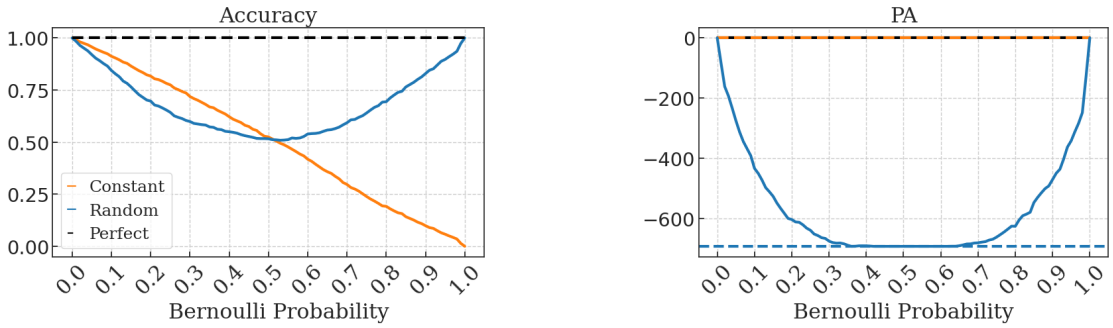



Figure 4.1: Evolution of PA and accuracy for constant, perfect and random classifiers across different values of $p \in [0, 1]$. Accuracy does not comply with the desired properties of a robustness metric and provides an inconsistent assessment that is exclusively driven by task performance. In contrast, PA continuously discriminates robust from unrobust classifiers for $p \in (0, 1)$ and reaches its minimum $-N \log 2$ (blue dashed line) for the random classifier when $p \in (0.3, 0.7)$.

Table 4.1 displays the values of PA and accuracy obtained for the case $p = 1/2$. It is clear that PA is independent of the task performance of the model, which in classification tasks is mostly reported through accuracy-based metrics, and instead discriminates effectively the random classifier from the rest. In particular, PA converges towards its minimum value $-N \log 2$ for the random classifier as $\beta \longrightarrow 0$, and converges towards zero for the other two as $\beta \longrightarrow \infty$. This behavior aligns with the intuitive assessment given in Example 3.2.1, since perfect and constant classifiers are robust by definition, while random classifiers are extremely unrobust.

Figure 4.1 extends these results across different values of $p \in [0, 1]$. It can be observed that the lower bound, represented by the blue dashed line, is achieved only after a certain mismatch threshold has been reached, of approximately 30% of the observations. This illustrates the trade-off navigated during $\beta$ optimization, in which matching observations are more heavily penalized the higher the value of $\beta$ is, whereas mismatching observations are more heavily penalized the lower the value of $\beta$ is. A balanced random classifier (i.e. balanced on the two classes) would converge to the lower bound for any possible original sample.

Finally, Figure 4.2 illustrates the $\beta$ optimization process for different values of prediction confidence gap $\Delta$. The confidence gap, expressed as a difference in the unnormalized log-odds $\boldsymbol{F}$, is very informative with regard to the quality of the model, as it is intuitively expected that the latent space represented by a high-confidence model encodes a better set of features to discriminate classes than one with a lower prediction confidence. In particular, the optimization concerns a random classifier with $p = 1/2$ and a sample of 100 observations, thus $\beta^* \longrightarrow 0$. Following the process already described in Section 3.6.2, we observe that the wider is the confidence gap (i.e. the informativeness of the posterior at $\beta_0 = 1$), the faster is the rate of convergence to the optimal value, considering that the number of mismatching observations is the same in all cases.

For that reason, results presented in this work will also include situations in which both samples are the same (i.e. $\boldsymbol{x}' = \boldsymbol{x}''$). In these cases, a pseudo-PA value will be obtained which will not correspond to the optimal with $\beta^* = 0$, but will depend on the rate with which it converges towards it. Given that convergence is faster for high confidence predictions, it will still be highly informative of the generalization capabilities of the model. More specifically, models displaying a lower proportion of high-confidence observations will be penalized. Given that low-confidence observations are associated with wrong predictions, the raking of models provided by a suboptimal PA for $\boldsymbol{x}' = \boldsymbol{x}''$ will be equivalent to that provided by standard accuracy. Further details on the empirical behavior of the PA metric can be found in Appendix B.1.1.

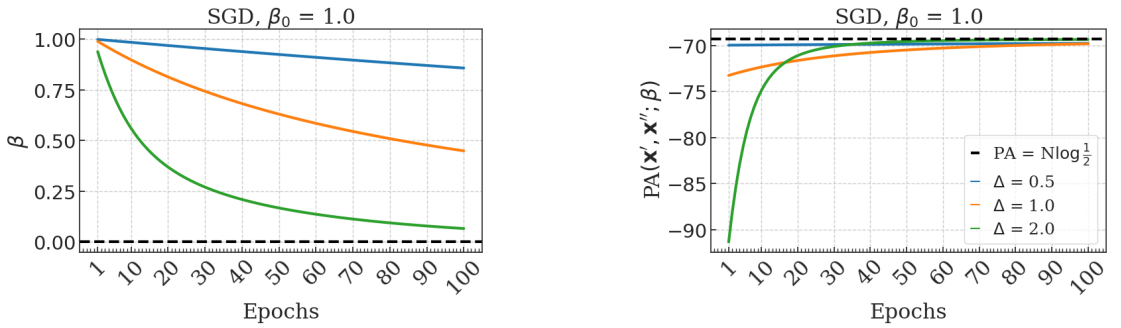



Figure 4.2: Evolution of the PA kernel optimization under different levels of prediction confidence for a random classifier with $p = 1/2$, computed over a sample of size $N = 100$. The rate of convergence is shown to depend on the informativeness of the posterior at $\beta_0 = 1$, and consequently on the confidence gap $\Delta$. An illustration of the original log-odds and its associated posterior distribution can be found in Appendix B.1.1.

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 it can be argued 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 synthetically designed, is relevant in any classification context because it provides a 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 perceptual standpoint and should therefore not perturb very significantly the predictions of robust models.

**Experiment 2.** The model under consideration is an image classifier for CIFAR10, a popular computer vision dataset composed of colored, $32 \times 32$ pixel images belonging to 10 different classes [35]. Sample $\boldsymbol{x}'$ contains 10.000 images and $\boldsymbol{x}''$ is generated by perturbing each image with white noise of increasing intensity. A pre-trained, undefended WideResNet-28-10 [72] architecture is 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 $\pm 3\sigma$.

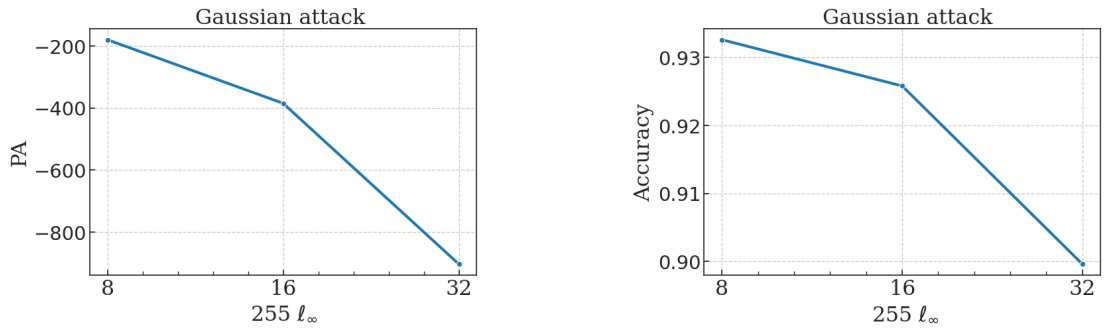



Figure 4.3: PA and accuracy displayed by the CIFAR10 classifier under increasing levels of white noise. Robustness and task performance are shown to be non-linearly related.

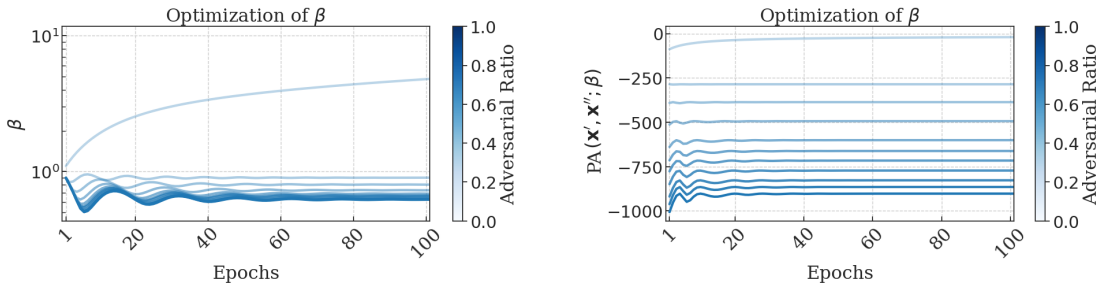



Figure 4.4: PA kernel optimization in the CIFAR10 gaussian noise setting for different ratio of perturbed observations. Perturbation magnitude is $\ell_\infty = 32 \ / \ 255$.

Figure 4.3 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.

Figure 4.4 expands these results by gradually adjusting the perturbation so that it only affects a specific ratio of observations. As expected, $\beta \longrightarrow \infty$ in the unperturbed case, and converges quickly to its optimal value in the rest of cases, even for a considerably large sample size. The decay in the PA value is less pronounced the higher fraction of perturbed observations 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 affected.

**Experiment 3.** The model under consideration is a sentiment classifier for IMDB, a popular NLP dataset containing 50,000 movie reviews labeled as either positive or negative [42]. A pre-trained DistilBERT-based architecture is adopted, with the appropriate tokenizer [53]. After a 100 epoch training using SGD with a learning rate of $10^{-3}$, the model with the highest validation accuracy is selected. Then, the original test dataset $\boldsymbol{x}'$ is incrementally perturbed to generate $\boldsymbol{x}''$, 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 as $2^L$.

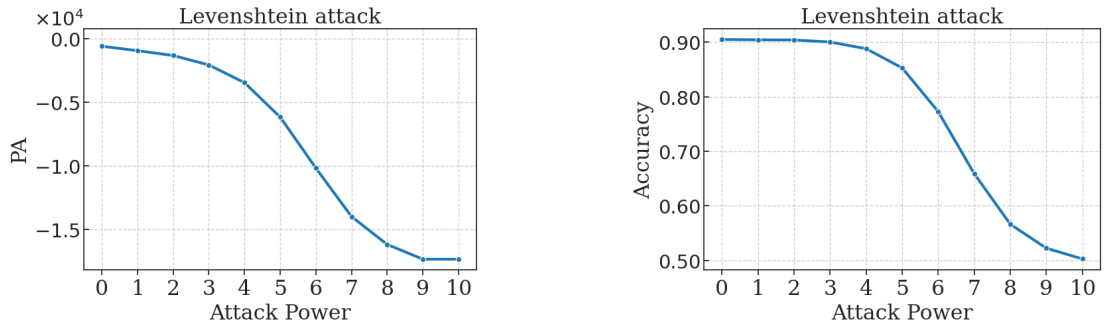



Figure 4.5: PA and accuracy for the IMDB sentiment classification under Levenshtein perturbations. The attack power is defined as $2^L$, being $L$ the Levenshtein distance between pairs of observations in $\boldsymbol{x}'$ and $\boldsymbol{x}''$.

Figure 4.5 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 designed to directly influence the sentiment prediction, which corresponds to an adversarial setting. For instance, Figure B.6 shows that accuracy can be manipulated to increase or decrease by perturbing original observations 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 displayed. The robustness against adversarial attacks in image classification tasks will be discussed in the next section.

## 4.2 Adversarial setting

The first real scenario in which covariate shift robustness will be measured is the adversarial setting. This setting serves as an archetypal use case for a 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 observations, 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 examples do not entail 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 attack failure rate (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 labels will be a reference for robustness, as it aligns with the aforementioned interpretation.

**Definition** (*Attack failure rate*)**.** Let $\hat{\boldsymbol{y}}', \hat{\boldsymbol{y}}'' \in \mathcal{Y}^N$ be the predicted class labels for $\boldsymbol{x}'$ and $\boldsymbol{x}''$, respectively, and let $\boldsymbol{y} \in \mathcal{Y}^N$ be the true labels. Considering the definition of accuracy provided in Section 3.2, the attack failure rate (AFR) can be expressed as

$$\begin{aligned} \mathrm{AFR}_{\mathrm{T}} &= \mathrm{Accuracy}(\hat{\boldsymbol{y}}'', \boldsymbol{y}), \\ \mathrm{AFR}_{\mathrm{P}} &= \mathrm{Accuracy}(\hat{\boldsymbol{y}}'', \hat{\boldsymbol{y}}'). \end{aligned}$$

**Definition** (*Adversarial ratio*)**.** 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 observations in the dataset, also known as adversarial ratio $\alpha \in [0, 1]$. The final adversarial dataset $\boldsymbol{x}''$ will be generated as

$$\boldsymbol{x}'' := \alpha \boldsymbol{x}'' + (1 - \alpha)\boldsymbol{x}',$$

where $\boldsymbol{x}'' = \boldsymbol{x}' + \boldsymbol{\Delta}$, 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 $\alpha$ and converge faster to the $\alpha = 1$ robustness value than any accuracy-based metric, in light of the behavior observed in Figure 4.4.

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 moderate attacks. This is because most adversarial perturbations 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 examples, 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.

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 observations with high confidence, as they contain the features encoded in the inductive bias of the classifier. Original observations that do not contain these features will be misclassified, and the lack of generalization to sampling randomness should be penalized for lowering the confidence in the predicted class. Regarding adversarial examples, 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 originally misclassified observations 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 subsets of observations.

**Proposition.** Let $\zeta_{\text{ERR}}$, $\zeta_{\text{MIS}}$ and $\zeta_{\text{ADV}}$ be the approximated robustness contributions of correctly classified original observations, misclassified original observations and misleading adversarial examples, respectively. Then, we can approximate PA as

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

with

$$\begin{aligned}
\zeta_{\text{ERR}} &= N \,\text{AFR}_{\text{T}}^{0}\, \text{AFR}_{\text{P}} \log\left(1 - 2\delta_{\text{ERR}}\right), \\
\zeta_{\text{MIS}} &= N(1 - \text{AFR}_{\text{T}}^{0})\, \text{AFR}_{\text{P}} \log\left(1 - 2\delta_{\text{MIS}}\right), \\
\zeta_{\text{ADV}} &= N \,\text{AFR}_{\text{T}}^{0}(1 - \text{AFR}_{\text{P}}) \log \delta_{\text{ADV}},
\end{aligned}$$

where $\text{AFR}_{\text{T}}^{0} \equiv \text{AFR}_{\text{T}}(\alpha = 0)$ is the accuracy of the model in the original data. 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). $\zeta_{\text{SAM}}$ aggregates the first two terms and will be interpreted as the sampling randomness contribution.

*Proof.* See Apendix B.1.2. □

| Defense | $\beta^*_{\text{PGD}}$ | $\Delta H_{\text{PGD}}$ | $\beta^*_{\text{FMN}}$ | $\Delta H_{\text{FMN}}$ |
|---|---|---|---|---|
| **Undefended** | 0.78 | 0.048 | 0.65 | 0.10 |
| **Engstrom et al.** | 15.63 | -1.204 | 2.59 | -0.71 |
| **Athalye et al.** | 35.48 | -3.049 | 19.84 | -2.13 |
| **Wong et al.** | 15.46 | -1.229 | 4.59 | -0.96 |
| **Addepalli et al.** | 15.89 | -2.023 | 6.08 | -1.71 |
| **Wang et al.** | 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 PGD $\ell_\infty = 8/255$ attacks with $\alpha = 1$. Entropy values are computed with the average posterior over correctly classified observations, which constitute the largest proportion of the dataset. Defended models converge to $\beta^* > 1$, whereas the undefended model displays $\beta^* < 1$. This is consistent with the interpretation of robustness in terms of the optimal informativeness of the posterior.

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 is used and differences by defect are more significantly penalized 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 (i.e. when $\alpha = 1$), which 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 $\alpha$ values. For that reason, the additional interpretability provided by this approximation will illustrate PA expression will help characterize the source of the robust and unrobust observed in the different models.

**Experiment 4.** The results provided in this section have been obtained using CIFAR10 [35], which is widely regarded as a standard benchmark for robustness evaluation. CIFAR10 is a balanced dataset containing 60.000 colored $32 \times 32$ pixel images belonging to 10 different classes. We will consider a pre-trained WideResNet-28-10 [72] as a baseline, undefended model and compare it to some state-of-the-art robust ResNet50 [26] models provided by the RobustBench [17] library under PGD [43] and FMN [49] 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.** [67], **Addepalli et al.** [1] and **Wang et al.** [65]. 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 $\mathbf{B}^{\ell_\infty}_{\infty}(x)$.

(a) Original (b) PGD, $\ell_\infty = 36/255$ (c) FMN

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

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 observations in the CIFAR10 dataset.

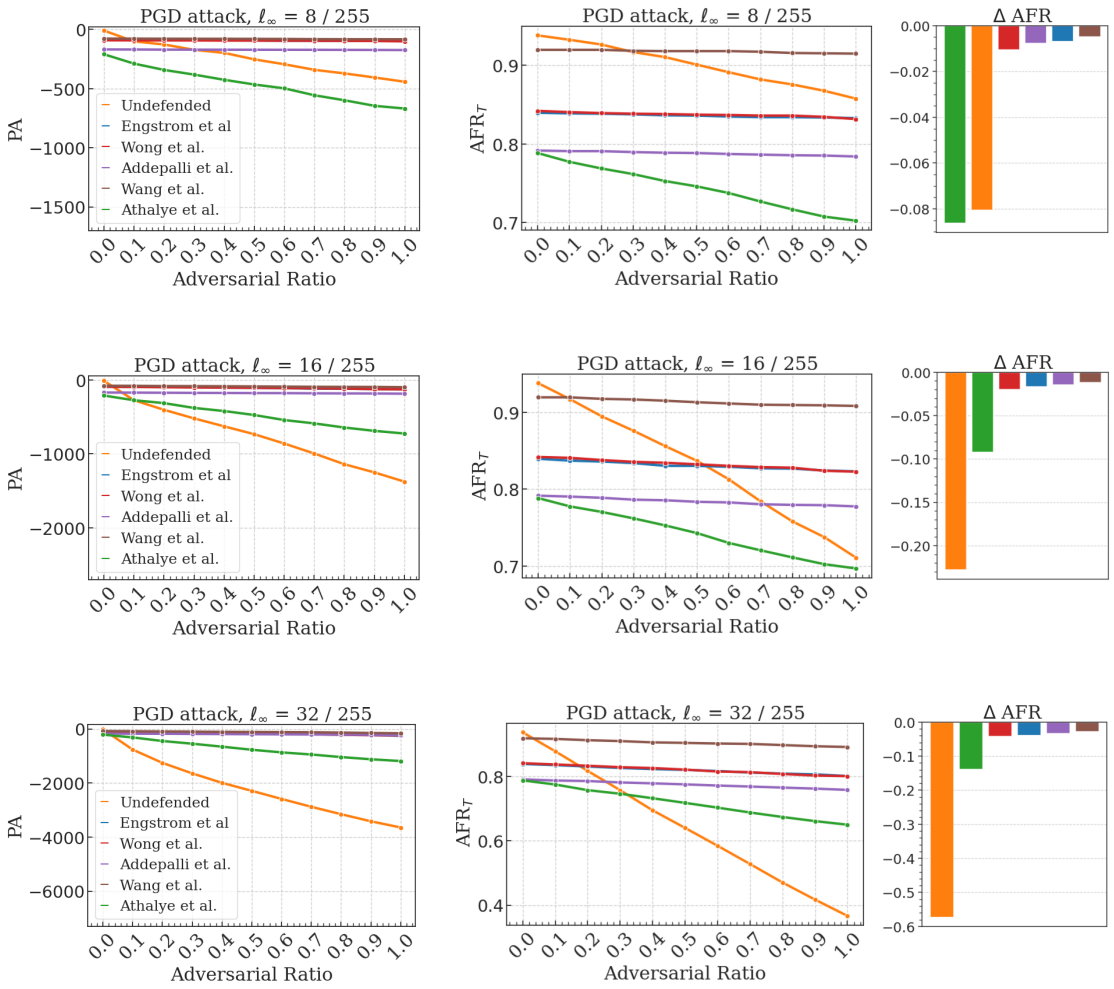



Figure 4.7: PA, $\text{AFR}_\text{T}$ and the AFR variation against increasing adversarial ratio $\alpha \in [0, 1]$ at different perturbation norm bounds $\ell_\infty$. A pre-trained, undefended WideResNet-28-10 and five RobustBench [17] defended models are subject to a 1000 step PGD attack. When $\alpha = 0$, PA $\longrightarrow 0$ as $\beta \longrightarrow \infty$ 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 $\text{AFR}_\text{T}$.

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 observations are misleading. Furthermore, it is also clear the **Athalye et al.** is significantly less robust to PGD attacks than its RobustBench counterparts, as its performance decreases way more significantly with increasing adversarial ratio.

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 behavior 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.

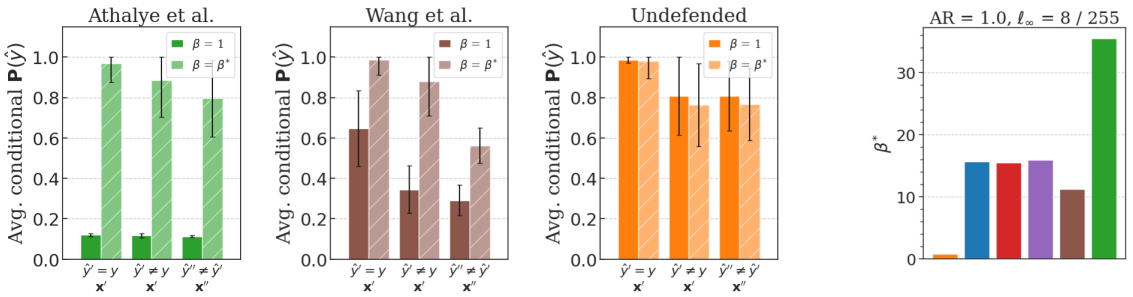



Figure 4.8: (**left**) Average posterior probability of the predicted class for correctly classified original observations, misclassified original observations and misleading adversarial observations. The probabilities assigned to correct ($\hat{y}' = y$) and incorrect ($\hat{y}' \neq y$, $\hat{y}'' \neq y$) predictions overlap in both **Undefended** and **Athalye et al.** models, whereas the opposite happens for the **Wang et al.** case. This highlights the suitability of its inductive bias with respect to both sampling randomness and adversarial perturbations. (**right**) Optimal $\beta^*$ value achieved by each model. **Undefended** and **Athalye et al.** models appear as outliers as they overestimate and underestimate, respectively, the information content of the features. Their poor robustness performance against PGD attacks reported in Figure 4.7 can be explained in these terms. These results have been obtained through a PGD attack with $\ell_\infty$=8/255, and they are consistent across different values of $\ell_\infty$.

We observe that the **Undefended** model provides overconfident predictions that maximize disagreement in misleading and misclassified observations, whereas **Athalye et al.** provides uncertain predictions that minimize disagreement in adversarial observations but have the opposite effect in correctly classified ones. The interpretation outlined in Chapter 2 contributes to the understanding of these results, as it is shown that $\beta^*$ effectively measures the quality of the informativeness estimation made by each model. On the one hand, 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.

Figure 4.8 **(left)** illustrates the previous reasoning by displaying the average posterior probability of the predicted class by each model (i.e. the maximum of the posterior), conditioned on the type of prediction assigned. This discrimination yields three groups of observations, namely original observations that are correctly classified by the model, original observations that are misclassified and perturbed examples that, having their associated unperturbed observation 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 behavior, in which original observations 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.

In the case of 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, $\text{AFR}_\text{P}$ 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 $\Delta$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 **Addepalli 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.

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

Table 4.3: Approximated PA contributions for a PGD attack with $\ell_\infty = 8/255$ and $\alpha = 1.0$. The number of originally misclassified and adversarially misleading observations is $N_{\text{MIS}} = \lfloor N(1 - \text{AFR}_{\text{T}}^0)\,\text{AFR}_{\text{P}} \rfloor$ and $N_{\text{ADV}} = \lfloor N\,\text{AFR}_{\text{T}}^0(1 - \text{AFR}_{\text{P}}) \rfloor$, respectively. The penalization argument $2\delta_{\text{ERR}}$ has not been included for being negligible in all cases.

Table 4.3 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 observations, and $\zeta_{\text{ERR}}$, $\zeta_{\text{MIS}}$ and $\zeta_{\text{ADV}}$ 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 observations and misleading adversarial observations, 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 observations 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 $\zeta_{\text{MIS}}$. The existing trade-off between standard and robust generalization arises when following this reasoning towards the minimization of $\zeta_{\text{MIS}}$, because reducing the number of misclassified observations 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 observations, 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, $\zeta_{\text{SAM}} = \zeta_{\text{ERR}} + \zeta_{\text{MIS}}$ can be understood as the lack of robustness to sampling randomness, and $\zeta_{\text{ADV}}$ as the lack of robustness to adversarial perturbations.

As expected, the standard generalization error term $\zeta_{\text{SAM}}$ 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 few misleading observations $N_{\text{ADV}}$. The discrimination of models based exclusively on $\zeta_{\text{SAM}}$ is very much aligned with that of $\text{AFR}_{\text{T}}$ in all cases except for the **Undefended** model, which is penalized more heavily for providing overconfident predictions with numerous 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 observations and outliers from the original dataset.

Further insights can be obtained by comparing these results with those of the **Athalye et al.** model, which has a similar accuracy on adversarial observations and a significantly worse accuracy on original observations. 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 $\zeta_{\text{SAM}}$ contribution than some of these models. Nevertheless, such uncertainty is also observed for original observations, which lowers accuracy on the original dataset and thus increases random sampling penalization $\zeta_{\text{SAM}}$. 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 $\text{AFR}_{\text{P}}$ 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 $\zeta_{\text{ADV}}$ is driven by the decrase 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.

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 observation 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 are obtained for FMN attacks, which are expected to be more effective than PGD attacks for being unbounded and lead to smaller $\beta^*$ values (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.

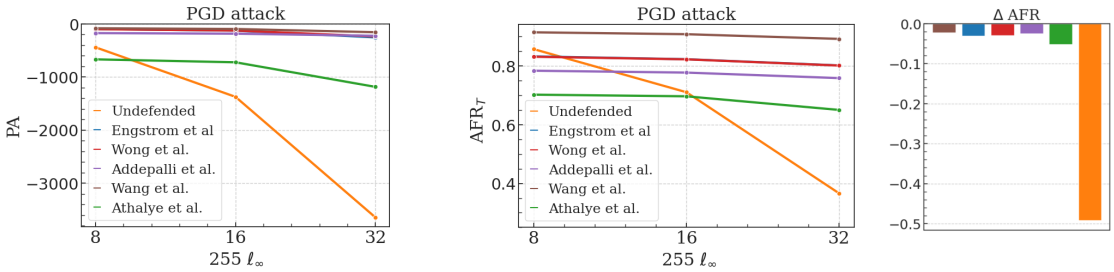



Figure 4.9: PA, $\text{AFR}_{\text{T}}$ and the AFR variation against increasing attack power for $\alpha = 1$. The undefended net and several RobustBench robust models are considered under a 1000 step PGD attack.

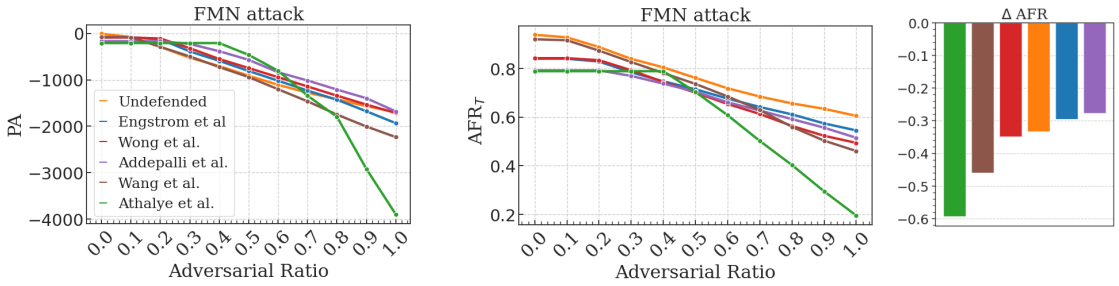



Figure 4.10: PA, $\text{AFR}_\text{T}$ and the AFR variation against increasing adversarial ratio. The undefended net and several RobustBench robust models are considered under a 1000 step FMN attack.

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. It must be noted that the adversarial ratio must be interpreted differently in this case, because the perturbations introduced are sorted by norm. In particular, an adversarial ratio $\alpha$ comprises perturbations up to the $\alpha$ quantile of the norm distribution. The consistency in the assessment provided by PA is lost for this reason, as prediction confidence is not constant across $\alpha$ values but instead decreases monotonically with it, which will trigger a different robustness response in each model.

It stems from these results that robust models have been defended with a compression strategy that succeeds at filtering out small perturbations, and for that reason maintain their performance at low adversarial ratio values [18]. In particular, **Athalye et al.** remains maximally robust until at least 40% of the observations are perturbed, at which point the defensive strategy is neutralized and a constant fraction of the additional perturbed observations 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 $\Delta$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 observations, which are the ones driving the metric after the $\alpha$ threshold is reached.

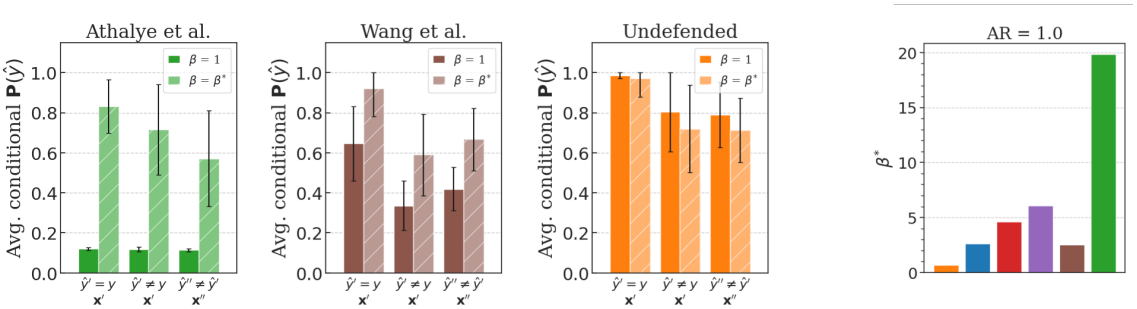



Figure 4.11: (**left**) Average posterior probability of the predicted class under FMN attack for correctly classified original observations, misclassified original observations, and misleading adversarial observations. The probabilities assigned to correct ($\hat{y}' = y$) and incorrect ($\hat{y}' \neq y$, $\hat{y}'' \neq y$) predictions are shown to overlap also for the **Wang et al.** model, which highlights the effectiveness of the FMN attack with respect to PGD. (**right**) Optimal $\beta^*$ value for each model.

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 behavior for $\beta = 1$, but optimal posteriors are less informative due to the increased number of misleading observations, 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 observations 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 behavior under both attacks.

Finally, Table 4.4 displays the approximated PA contributions for the FMN attack. In contrast with the PGD case, FMN is much more effective and $N_{\text{ADV}} > N_{\text{MIS}}$ in all cases, which makes the adversarial contribution $\zeta_{\text{ADV}}$ 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 $\zeta_{\text{SAM}}$ among robust models due to its superior accuracy. In the context of effective attacks, predictive certainty on original observations is highly rewarded because lower overall agreement makes sampling penalization terms $\log(1 - 2\delta_{\text{MIS}})$ more negative.

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 $\text{AFR}_{\text{P}}$, 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 proportion of perturbed observations, under which $\text{AFR}_{\text{P}}$ exhibits significant fluctuations that alter the ranking of the models at every step, as illustrated in Tables 4.5.

Furthermore, the analysis of the approximated robustness contributions highlights the fact 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 $\Delta$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 differently.

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

Table 4.4: Approximated PA contributions for a FMN attack with $\alpha = 1.0$. The number of originally misclassified and adversarially misleading observations is $N_{\text{MIS}} = \lfloor N(1 - \text{AFR}_{\text{T}}^0)\,\text{AFR}_{\text{P}} \rfloor$ and $N_{\text{ADV}} = \lfloor N\,\text{AFR}_{\text{T}}^0(1 - \text{AFR}_{\text{P}}) \rfloor$, respectively. The penalization argument $2\delta_{\text{ERR}}$ has not been included for being negligible in all cases except for the **Athalye et al.** model, which amounts to 0.36.

| PGD | $\alpha = 2/10$ PA | $\alpha = 2/10$ $AFR_P$ | $\alpha = 2/10$ $AFR_T$ | $\alpha = 4/10$ PA | $\alpha = 4/10$ $AFR_P$ | $\alpha = 4/10$ $AFR_T$ | $\alpha = 6/10$ PA | $\alpha = 6/10$ $AFR_P$ | $\alpha = 6/10$ $AFR_T$ |
|---|---|---|---|---|---|---|---|---|---|
| Addepalli et al. | **-172.5** | **0.995** | **0.785** | **-175.5** | 0.992 | **0.786** | **-177.6** | 0.989 | **0.783** |
| Wong et al. | -97.7 | 0.996 | 0.838 | -102.9 | 0.992 | 0.834 | -109.2 | **0.987** | 0.830 |
| Engstrom et al. | -94.2 | 0.996 | 0.836 | -104.6 | **0.990** | 0.830 | -110.3 | 0.988 | 0.829 |
| Wang et al. | -81.9 | 0.997 | 0.917 | -84.6 | 0.996 | 0.915 | -89.4 | 0.991 | 0.912 |
| **FMN** | PA | $AFR_P$ | $AFR_T$ | PA | $AFR_P$ | $AFR_T$ | PA | $AFR_P$ | $AFR_T$ |
| Addepalli et al. | -169.4 | 1.0 | **0.791** | -385.4 | 0.944 | **0.737** | –838.9 | 0.867 | 0.660 |
| Wong et al. | -111.2 | 0.991 | 0.834 | -553.1 | 0.901 | 0.743 | -944.8 | 0.810 | **0.653** |
| Engstrom et al. | -128.5 | 0.988 | 0.828 | -592.9 | 0.907 | 0.747 | -1020 | 0.836 | 0.675 |
| Wang et al. | **-291.6** | **0.952** | 0.873 | **-726.8** | **0.861** | 0.781 | **-1204** | **0.764** | 0.684 |

Table 4.5: Comparison of PA, $AFR_P$ and $AFR_T$ scores for a PGD attack with $\ell_\infty = 16 \,/\, 255$ and an FMN attack across different adversarial ratio values. Among robust models, the worst robustness score is emboldened for every case. PA displays higher consistency and discriminative power across varying $\alpha$ with respect to accuracy-based metrics. In the PGD case, PA is aligned with $AFR_T$ because sampling randomness is the principal source of unrobustness. In the FMN case, PA is aligned with $AFR_P$ because adversarial perturbations are the principal source of unrobustness.

As a conclusion to the analysis, several discrepancy measures between the model's representation of $\boldsymbol{x}'$ and $\boldsymbol{x}''$ have been considered, so that it can be determined whether they are able to discriminate robust models with a similar power as that of PA. On the one hand, posterior-based metrics such as Kullback-Leibler divergence and Wasserstein distance have been considered, each providing a notion of discrepancy between probability distributions. On the other hand, different distance measures in the latent space of feature representations were considered, including cosine similarity, centroid distance, and other cluster-based approaches like maximum mean discrepancy and Fréchet inception distance (FID). FID is particularly relevant, as it is a widely used metric in the generative adversarial network literature and thus fits intuitively well into our setting.

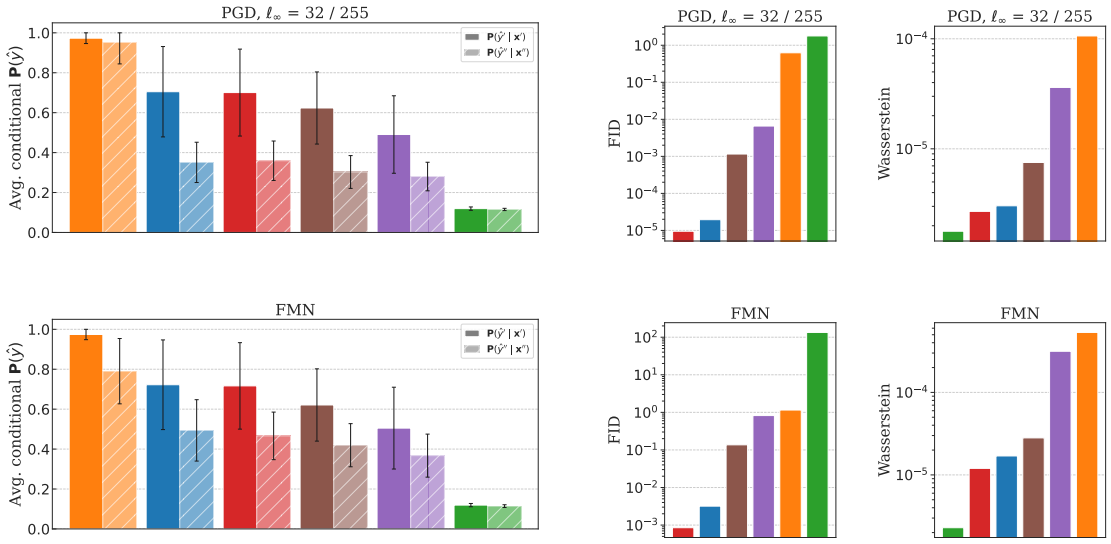



Figure 4.12: (**left**) Average posterior probability of the predicted class in the original and perturbed datasets with $\alpha = 1$. (**middle**) Average Fréchet Inception Distance (FID) between feature space representations of $\boldsymbol{x}'$ and $\boldsymbol{x}''$. (**right**) Average Wasserstein (W) distance between posteriors $\mathbf{P}^f(\cdot|\boldsymbol{x}')$ and $\mathbf{P}^f(\cdot|\boldsymbol{x}'')$ at $\beta_0 = 1$.

Figure 4.12 displays the FID and Wasserstein distance values for all models with $\alpha = 1$, for both PGD and FMN attacks, together with a representation of the average posterior for the original and perturbed datasets at $\beta_0 = 1$. Both metrics preserve the relative ordering of the models across different attacks, and their discriminative power does not appear to correlate with the baseline robustness and performance metrics considered. On the one hand, posterior distance is biased towards low-informative distributions, and for that reason yields the smallest value to the **Athalye et al.** model. In this model, the overlap between posteriors is the smallest in absolute terms but the largest in proportion, which constitutes the main source of generalization error. On the other hand, the FID between feature space representations yields a high value for models with significant posterior overlap, which serves as a first-order indicator of the quality of the inductive bias. Nevertheless, the remaining models are discriminated by several orders of magnitude, and no direct relationship with the measured robustness can be established. 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.

## 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. $\text{AFR}_\text{P}$), because significant disagreement in the remaining classes indicates vulnerability to distribution shifts present in source domains, and in turn to target domains as well.

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. The posterior agreement between two source datasets will be computed 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, **Arjovsky et al.**) and selective augmentation (LISA, **Yao et al.**), 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 the highest PA scores perform better on new domains.

Training datasets will be generated 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 modifications to the `diagvibsix` library[1] have been implemented with the purpose of this project so that datasets can be built with a specific configuration of factors for each observation, which will allow for a wide range of experiments in the domain shift and model selection settings.

**Definition** (*Shift ratio*)**.** In an analogous way to the adversarial case, datasets will be incrementally perturbed by including only a fraction $\alpha$ of the shifted observations, which in this context will be called shift ratio. The final shifted dataset $\boldsymbol{x}''$ will be generated as

$$\boldsymbol{x}'' := \alpha \boldsymbol{x}'' + (1 - \alpha)\boldsymbol{x}',$$

where $\boldsymbol{x}' \sim \boldsymbol{X}'$ and $\boldsymbol{x}'' \sim \boldsymbol{X}''$, as per Definition 3.2.

[1] https://github.com/viictorjimenezzz/diagvibsix/tree/librarization

**Experiment 5a.** 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 defined as

$$\mathbb{S} = \{X_0, X_1\},$$
$$\mathbb{T} = \{X_2, X_3, X_4, X_5\},$$

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 sampling experiment, namely $\tau_0^{\text{train}}$, $\tau_1^{\text{train}}$, $\tau^{\text{val}}$ and $\tau^{\text{test}}$, each involving disjoint subsets of MNIST. Following the notation introduced in Chapter 3, we can define:

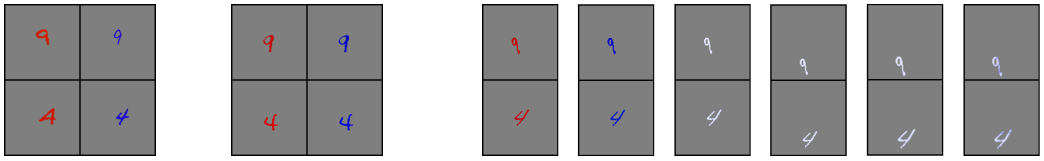
$$D^{\text{train}} = \{\boldsymbol{x}_0^{\text{train}}, \boldsymbol{x}_1^{\text{train}}\}, \text{ where } \boldsymbol{x}_j := \boldsymbol{x}_j \circ \tau_j^{\text{train}}, \ j = 0, 1$$
$$D^{\text{val}} = \{\boldsymbol{x}_0^{\text{val}}, \boldsymbol{x}_1^{\text{val}}\}, \text{ where } \boldsymbol{x}_j := \boldsymbol{x}_j \circ \tau^{\text{val}}, \ j = 0, 1$$
$$D_j^{\text{test}} = \{\boldsymbol{x}_j^{\text{test}}\}, \text{ where } \boldsymbol{x}_j^{\text{test}} := \boldsymbol{x}_j^{\text{test}} \circ \tau^{\text{test}}, \ j = 0, \ldots, 5$$

In this way, only training data is subject to both sampling randomness ($\tau_0^{\text{train}} \neq \tau_1^{\text{train}}$) and domain shift ($X_0 \nsim X_1$), emulating the conditions of real-world sampling experiments. In contrast, validation and test datasets entail each a single experiment instantiation, which means that distribution shift is the only accountable source of randomness between $\boldsymbol{x}'$ and $\boldsymbol{x}''$. 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.

The experiment definition comprises the complete characterization of source and target domains and the nature of 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 observation, the shifted dataset contains the same instances and in the same order as the original dataset, thus removing sampling randomness contributions from the robustness score. Table 4.6 stipulates the specific values defining each environment and Figure 4.13 illustrates them with an example.

| # Shift Factors | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Hue | red | **blue** | blue | blue | blue | blue |
| Lightness | dark | dark | **bright** | bright | bright | bright |
| Position | CC | CC | CC | **LC** | LC | LC |
| Scale | normal | normal | normal | normal | **large** | large |
| Texture | blank | blank | blank | blank | blank | **tiles** |
| *Shape* | *4,9* | *4,9* | *4,9* | *4,9* | *4,9* | *4,9* |

Table 4.6: Image factors associated to each of the environments considered in this experiment. CC and LC account for 'centered center' and 'centered low', respectively.



Train Validation Test (0 - 5)

Figure 4.13: Illustration of the training, validation and test datasets. Samples composing training datasets belong to different MNIST subsets, whereas samples composing validation and test datasets are corresponding (i.e. they are the same observation).

ERM, IRM [2] and LISA [69] algorithms were used to train a ResNet18 architecture for 100 epochs on dataset $D^{\text{train}}$ using Adam [33] optimizer with a learning rate of $10^{-3}$. Accuracy on validation dataset $D^{\text{val}}$ was monitored and weights achieving maximum performance were selected for evaluation. The LISA interpolation factor $p \in [0, 1]$ was also selected following this criterion and was ultimately set to 0.5, which indicates that both domain and factor interpolation between training samples contribute to improve task performance in this setting.

Figure 4.14 illustrates the discriminative capabilities of PA across test datasets. PA is shown to to decrease monotonically with increasing shift ratio and shift power in the cases of **Vanilla ERM** and **Arjovsky et al.**, with the latter being selected as a superior algorithm under these conditions. In contrast, **Yao et al.** does not exhibit the same consistent downward trend for $\alpha < 1$ and is notably penalized in the last dataset. This observation supports the robustness interpretation discussed in this work, suggesting that the model's inductive bias, rather than the data distribution, is responsible for this behavior. In particular, since LISA interpolates between observations from $\boldsymbol{x}_0^{\text{train}}$ and $\boldsymbol{x}_1^{\text{train}}$, the features encoded in the selected model may not align with those implicitly selected for the factor modification.

Table 4.7 expands on this analysis by providing the $\text{AFR}_\text{P}$ and $\text{AFR}_\text{T}$ scores associated with this experiment for $\alpha = 1$. PA assigns the highest score to **Yao et al.** in the first dataset, which is composed from a configuration of factors that has been seen by the model at training time, while consistently selecting **Arjovsky et al.** as the most robust for the remaining cases. In contrast, accuracy-based metrics provide an inconsistent assessment across different levels of shift power that does not clearly discriminate models by their response against source or target domain samples. In the absence of sampling randomness, a discontinuity in the feature representations that leads to this inconsistent behavior is not likely to occur, especially when PA displays a consistent downward trend in most cases.

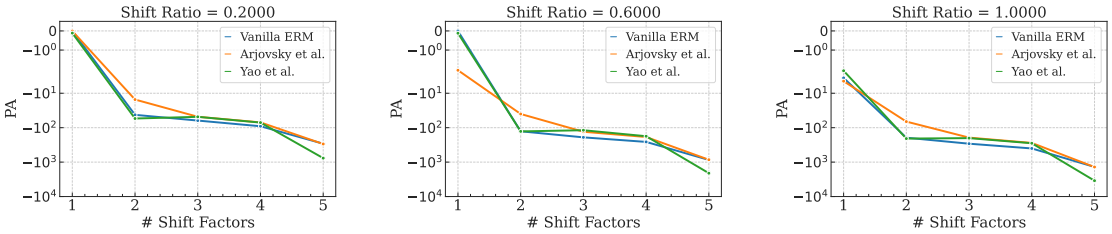



Figure 4.14: Evolution of PA under increasing levels of shift power and shift ratio $\alpha$ for **Experiment 5a**. PA is sensitive to the domain from which samples are drawn (i.e. source for dataset 1, target for the rest) and provides a consistent assessment across increasing levels of domain shift and shift ratio. In particular, **Yao et al.** achieves the highest score for the first dataset, whereas **Arjovsky et al.** algorithm is considered the most robust for the remaining cases.

| | 1 Shifted Factor | | | 2 Shifted Factors | | | 3 Shifted Factors | | | 4 Shifted Factors | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PA | $\text{AFR}_\text{P}$ | $\text{AFR}_\text{T}$ | PA | $\text{AFR}_\text{P}$ | $\text{AFR}_\text{T}$ | PA | $\text{AFR}_\text{P}$ | $\text{AFR}_\text{T}$ | PA | $\text{AFR}_\text{P}$ | $\text{AFR}_\text{T}$ |
| **Vanilla ERM** | -3.583 | 0.999 | 0.992 | -195.4 | 0.979 | 0.973 | 291.8 | 0.968 | 0.963 | -400.9 | 0.957 | 0.953 |
| **Arjovsky et al.** | -4.463 | 0.999 | 0.989 | **-66.54** | **0.994** | 0.987 | **-194.2** | 0.982 | 0.978 | **-277.1** | 0.974 | **0.972** |
| **Yao et al.** | **-2.219** | **0.999** | **0.993** | -207.2 | 0.980 | **0.975** | -200.9 | **0.983** | **0.979** | -282.9 | **0.976** | 0.970 |

Table 4.7: Comparison of PA, $\text{AFR}_\text{P}$ and $\text{AFR}_\text{T}$ scores for ERM, IRM and LISA learning algorithms with $\alpha = 1$ for **Experiment 5a**. The highest robustness score is emboldened for every case. PA provides a consistent assessment across target domains and selects **Arjovsky et al.** as the most robust model. In contrast, accuracy-based metrics are less discriminative and significantly inconsistent across datasets, thus yielding no clear verdict.

**Experiment 5b.** Alternative results were obtained by considering different sampling instantiations $\tau_0^{\text{val}} \neq \tau_1^{\text{val}}$ and $\tau_0^{\text{test}} \neq \tau^{\text{test}}$ to generate the first validation and test datasets. In particular,

$$D^{\text{val}} = \{\boldsymbol{x}_0^{\text{val}} \circ \tau_0^{\text{val}}, \boldsymbol{x}_1^{\text{val}} \circ \tau_1^{\text{val}}\} \text{ and } D_0^{\text{test}} = \{\boldsymbol{x}_0^{\text{test}} \circ \tau_0^{\text{test}}\}.$$

In this variation of **Experiment 5a**, samples $\boldsymbol{x}_0^{\text{test}}$ and $\boldsymbol{x}_j^{\text{test}}$, $j = \{1, \dots, 5\}$, are shifted due to both sampling randomness and image factor instantiations. This setting is not expected to change abruptly the performance score provided by accuracy-based metrics, but will definitely lower the PA robustness score, as the overlap between posteriors will decrease.

Figure 4.14 shows that PA succeeds at discriminating the different models by their predictive response under increasing shift ratio and shift power. In particular, **Vanilla ERM** can be identified to be non-robust by the fact that its score is maximum for the first test dataset, which belongs to the source domains, but decays rapidly to a minimum value after an additional factor is shifted. In contrast, **Arjovsky et al.** and **Yao et al.** models display a lower robustness score for the first test dataset, especially for small shift ratio values, but then decay at a lower rate than ERM for the following target datasets. In this case, **Yao et al.** is selected as the most robust model through both performance-based and robustness-based metrics. This change with respect to the previous experiment indicates that sample interpolation provides increased robustness against sampling randomness. In the previous setting, where samples $\boldsymbol{x}_0^{\text{test}}$ and $\boldsymbol{x}_j^{\text{test}}$ were composed of the same MNIST observations, this advantage was not exploited and the domain-invariance regularization performed by **Arjovsky et al.** was shown to be more effective.

These statements are further supported by the comparison of PA, $\text{AFR}_{\text{P}}$ and $\text{AFR}_{\text{T}}$ scores in Table 4.8. The robustness assessment provided by PA aligns with that of accuracy-based metrics in all target domains, which differs from the behavior observed in the absence of sampling randomness (see Table 4.7). Nevertheless, only PA assigns the highest score to **Vanilla ERM** on the first test dataset, which belongs to the source domain and thus represents a combination of factors that the model has seen during training. This is a clear indication that PA is able to capture the generalization capabilities of the model in a more substantial way than accuracy-based metrics, which are unable to discriminate different test datasets.

Figure 4.14 also shows an unexpected increase in PA after four shifted factors, which seems inconsistent with the alleged non-increasing behavior of PA, as per Section 3.2. Nevertheless, this phenomenon highlights again that robustness does not stem from the data generation process but rather from the latent representation of the model and the features selected for the construction of its inductive bias, as discussed in the introducting chapter. In this particular case, Table 4.9 proves that there is a clear discontinuity in the feature representation of the data when the texture factor is shifted (see Table 4.6), which results in a different discriminator function that ultimately leads to a different predictive outcome.

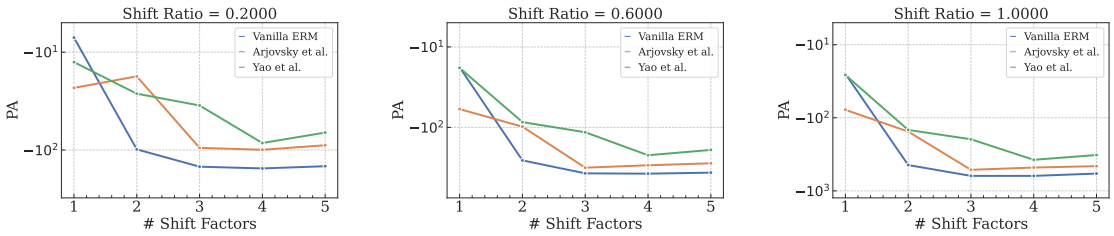



Figure 4.15: Evolution of PA under increasing levels of shift power and shift ratio $\alpha$ for **Experiment 5b**. Even in the presence of sampling randomness, PA is highly sensitive to source and domain environments and provides a consistent assessment of the generalization capabilities of the different models.

| | 1 Shifted Factor | | | 3 Shifted Factors | | | 5 Shifted Factors | | |
|---|---|---|---|---|---|---|---|---|---|
| | PA | $AFR_P$ | $AFR_T$ | PA | $AFR_P$ | $AFR_T$ | PA | $AFR_P$ | $AFR_T$ |
| **Vanilla ERM** | **-24.91** | 0.999 | 0.993 | -625.6 | 0.979 | 0.975 | -579.4 | 0.976 | 0.873 |
| **Arjovsky et al.** | -76.76 | 0.998 | 0.993 | -514.2 | 0.976 | 0.978 | -464.2 | 0.976 | 0.911 |
| **Yao et al.** | -26.21 | **0.999** | **0.994** | **-201.2** | **0.985** | **0.980** | **-324.4** | **0.988** | **0.945** |

Table 4.8: Comparison of PA, $AFR_P$ and $AFR_T$ scores for ERM, IRM and LISA learning algorithms with $\alpha = 1$ for **Experiment 5b**. The highest robustness score is emboldened for every case. PA is able to discriminate algorithms consistently and distinguish the first shifted factor dataset, which is drawn from source domains, from the rest.

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

Table 4.9: Average pairwise cosine similarity between feature space representations of original and augmented images (i.e. $\boldsymbol{x}_0^{\text{test}}$ vs. $\boldsymbol{x}_j^{\text{test}}$), for each of the shifted datasets in **Experiment 5b**. 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.

Finally, Figure 4.16 and Table 4.10 display the average posterior probability assigned to the predicted class for $\beta = 1$, and the subsequent optimal $\beta^*$ value derived from the maximization of posterior agreement, respectively. These results, which are reported for both **Experiment 5a** and **Experiment 5b**, give insight into the rationale behind the assessment provided by PA. In particular, given that accuracy was not discriminative across models and $\boldsymbol{x}_0^{\text{test}}$, $\boldsymbol{x}_j^{\text{test}}$ displayed a similar number of matching observations for all $j \in \{1, \dots, 5\}$, models providing the highest posterior probability to the correct class were selected as the most robust.

It follows also from the values of $\beta^*$ in Table 4.10 that the presence of sampling randomness between datasets under consideration does not significantly alter the robustness assessment provided in the **Vanilla ERM** and **Arjovsky et al.** cases, which indicates that the inductive bias of these models generalizes to different instantiations of MNIST. In contrast, the **Yao et al.** model displays a different generalization performance in between both experiments that does not stem from a difference in performance or $AFR_P$. This suggests that PA can provide a more informative assessment of the generalization capabilities of models and the suitability of the inductive bias of models under different conditions. In this particular case, the features learned by **Yao et al.** yield poorly informative posteriors in the absence of sampling randomness, which translates to a suboptimal robustness score in **Experiment 5a**.

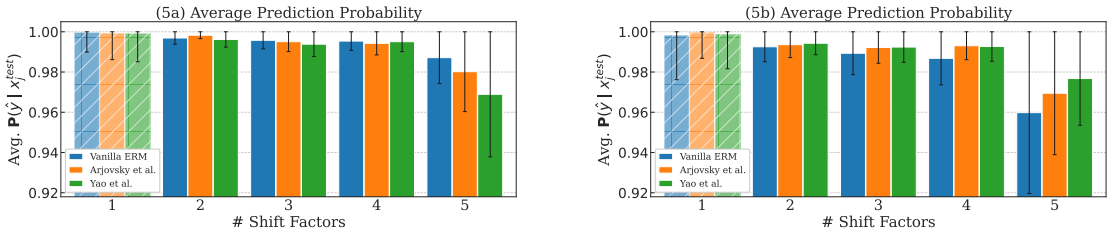



Figure 4.16: Average posterior probability of the predicted class for each of the test datasets, under the conditions established for **Experiment 5a** (left) and **Experiment 5b** (right). The first dataset is distinguished from the rest for belonging to the source domains.

| | Experiment 5a | Experiment 5b |
|---|---|---|
| **Vanilla ERM** | 2.303 | 3.022 |
| **Arjovsky et al.** | 4.927 | 3.138 |
| **Yao et al.** | 5.687 | 1.837 |

Table 4.10: Comparison of $\beta^*$ obtained after 1000 optimization epochs for both experiments conducted in the out-of-distribution setting. Given that performance in source domains is almost identical across models, the informativeness of the predictions is the main driver of the robustness assessment provided by PA.

Overall, these results highlight that posterior agreement possesses a superior discriminative power with respect to accuracy-based metrics. In particular, it provides a consistent assessment of the generalization capabilities of models under different sources of randomness, as it has shown to be sensitive to the presence of both sampling randomness and covariate shift and also to the type of domain (i.e. source or target) from which samples are drawn. This increased sensitivity is useful for the characterization and evaluation of inductive biases, which has been proven to determine the robust or unrobust behavior of models under different conditions.

The results obtained in this section further corroborate that all sources of randomness that affect the data generation process are accounted for in the PA score. In the adversarial setting, an approximation was performed so to discriminate the contribution of adversarial perturbations to the overall robustness measure. In this setting, a different procedure will be followed that will separately assess the suitability of the inductive bias against both distribution shift and sampling randomness.

**Experiment 5c.** ERM and IRM [2] algorithms were used to train a ResNet18 model for 50 epochs on dataset $D^{\text{train}}$ using Adam [33] optimizer with a learning rate of $10^{-2}$. A small subset of 128 observations from each validation sample $\boldsymbol{x}_0^{\text{val}}$, $\boldsymbol{x}_1^{\text{val}}$ generated for **Experiment 5a** is selected to generate a reduced validation dataset $D^{\text{sub}} = \{\boldsymbol{x}_0^{\text{sub}}, \boldsymbol{x}_1^{\text{sub}}\} \subset D^{\text{val}}$. Samples $\boldsymbol{x}_0^{\text{sub}}$, $\boldsymbol{x}_1^{\text{sub}}$ entail the same randomness instantiation $\tau^{\text{val}}$ and therefore are composed of the same MNIST observations and in the same order. Consequently, they are only subject to distribution shift, which in this case stems from the hue factor (i.e. blue vs red), as per Table 4.6.

At the end of every epoch, the principal component (see PCA [31]) of the feature space representation of $\boldsymbol{x}_0^{\text{sub}}$ and $\boldsymbol{x}_1^{\text{sub}}$ is computed separately, so that each MNIST observation in $D^{\text{sub}}$ can be associated with two values, one obtained under each sample. In a qualitative way, the suitability of the inductive bias represented in a particular epoch can be assessed by characterizing the principal component, which is the direction along which the data exhibits the highest variance.

Let $\Phi^c$ be the feature extractor of classifier $c$ after a training epoch. The principal direction of the feature space representations of samples $\boldsymbol{x}_0^{\text{sub}}$, $\boldsymbol{x}_1^{\text{sub}}$ will be denoted as $\boldsymbol{v}_0$, $\boldsymbol{v}_1$, respectively. The projections of each observation over these components were computed as

$$z_{0,n} = \langle \Phi^c(x_{0,n}^{\text{sub}}), \boldsymbol{v}_0 \rangle, \quad z_{1,n} = \langle \Phi^c(x_{1,n}^{\text{sub}}), \boldsymbol{v}_1 \rangle, \quad n = 1, \dots N_{\text{sub}}$$

On the one hand, the distribution of observations belonging to different classes across the principal direction is an indicator of the task-specific suitability of the inductive bias. In general, the inductive bias of the model should be constructed with the most predictive features for the task at hand, and therefore class membership should be the main driver of the variance in the feature space and thus be encoded in the principal component. The class-conditional variance of $\boldsymbol{z}_0$, $\boldsymbol{z}_1$ can be used as measure of the discriminative power of the latent features encoded, and consequently of the robustness of the model against sampling randomness.

On the other hand, the similarity between the projections of each observation along the principal direction is an indicator of cross-domain invariance and consequently of the robustness of the model against distribution shifts. If the inductive bias of the classifier were completely domain-agnostic, the feature representations of shifted observations would be the same, and so its projection along the principal direction. The mean squared error between the $\boldsymbol{z}_0$ and $\boldsymbol{z}_1$ can be used as a measure of the shift's influence in the inductive bias.

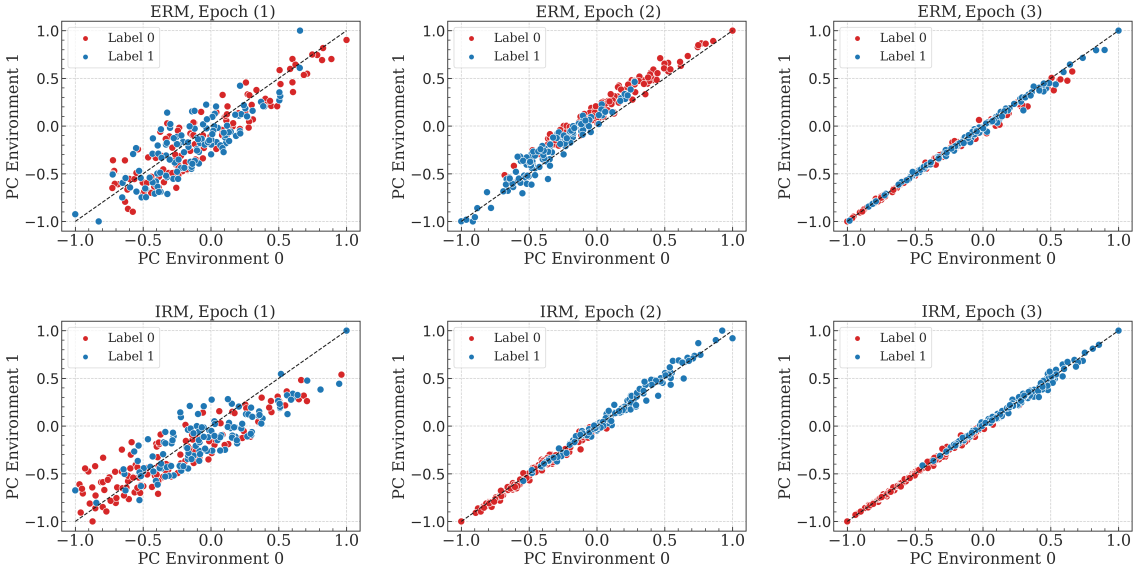



Figure 4.17: Normalized $\boldsymbol{z}_0$ vs $\boldsymbol{z}_1$ plot for ERM (top) and IRM (bottom) algorithms at three different training stages. Projections are colored by class membership, and the dashed line illustrates the $\boldsymbol{z}_0 = \boldsymbol{z}_1$ case. Under this configuration, ERM projections display either a high cross-domain error or a high class-conditional variance. In contrast, IRM is able to encode a representation that reduces both measures at the same time, thus indicating a more robust inductive bias.

Figure 4.17 illustrates the principal component projections of the feature space representations of samples $\boldsymbol{x}_0^{\text{sub}}$, $\boldsymbol{x}_1^{\text{sub}}$ for ERM and IRM algorithms at three different training stages. These results show that ERM is unable to encode a representation that is both discriminative and invariant to domain shifts, as it either displays a high cross-domain error or a high class-conditional variance. This indicates that its inductive bias is exclusively driven by domain-specific features or by class-specific features, and possibly the high learning rate avoids the model to converge to a more robust solution. In contrast, IRM is able to encode a representation that reduces both qualitative measures at the same time, which indicates that the inductive bias is able to capture the most predictive features for the task at hand without being significantly influenced by the shift in the hue factor.

These measures have been shown to qualitatively assess the suitability of the inductive bias for the aforementioned sources of randomness separately. For that reason, they will be monitored and used in the model selection setting to assist in the hyperparameter tuning process and to provide a more comprehensive interpretation of the source of robust or unrobust behavior reported by PA.

# 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 enhancing generalization performance under distribution shifts.

It is important to recognize that epoch-wise model selection represents a fundamentally different paradigm from the robustness assessment framework discussed in the previous chapter, which could be described as algorithm selection. In the previous scenario, models were selected based on standard validation accuracy and subsequently evaluated for their robustness to different sources of covariate shift. The assessment was thus conducted on a configuration of weights that had already been selected to encode a suitable inductive bias that demonstrated to perform well on samples drawn from the same distribution.

In that context, the robustness assessment provided by PA was implicitly constrained to a definition of generalization error that could be assumed to align with the error associated to the task performance, and predictions were expected to match the desired learning outcome. However, this assumption breaks down when extending the evaluation framework to epoch-wise selection, as consistency in predictive outcomes across different samples does not guarantee alignment with the performance over the task at hand. More specifically, there is no assurance that the inductive bias encoded by the most robust model actually represents a set of features that are relevant to the learning problem.

This reasoning becomes clearer when the problem is formulated in a data-agnostic way. After every epoch, the classifier encodes a different feature extractor function $\Phi$, resulting in distinct feature space vectors for each observation. Every epoch can therefore be considered as an independent sampling experiment in the feature space, and robustness between two realizations of these experiments is then evaluated using PA. Given that the underlying statistical model governing the sampling process is determined by the inductive bias of the model, which varies at every epoch, it is not possible for PA to mitigate or even detect overfitting to suboptimal biases associated with features or spurious correlations that are also present in the samples it relies upon.

Overall, the critical distinction between performance-based and robustness-based criteria has already been addressed, more notably in Section 4.1. It is clear that a classifier overfitting to specific features in the training data would lower its performance in validation data and simultaneously be considered increasingly robust. Besides, the ultimate measure of success in model selection will remain to be accuracy on test sets, particularly those from the target domain, as this is the primary objective of domain adaptation.

For these reasons, this chapter will explore two principal situations. First, the potential misalignment between robustness measurement and task performance will be monitored by considering a wide range of validation sample pairs for a single learning task. In most cases, $\boldsymbol{x}'$ will be drawn from a source environment and thus encode a set of features that the model has seen during training, while $\boldsymbol{x}''$ will be drawn from target environments. In particular, different levels of shift will be considered for $\boldsymbol{x}''$, as to simulate different degrees of access to target domains. The loss minimization process will ensure that the model's predictions for $\boldsymbol{x}'$ align with the desired learning outcomes and, in turn, robustness assessment will evaluate whether predictions for $\boldsymbol{x}''$ align with these expectations as well. The ability of the robustness criterion to select models that perform well in target domains will be in question.

Second, the behavior and suitability of robustness metrics as model selection criteria will be assessed under specific configurations of the training data that are crafted to encode suboptimal sets of features in the inductive bias. In particular, several degrees of co-occurence between image factors and output labels will be introduced, and the model selection capabilities of PA will be compared to that of standard accuracy under these conditions.

## 5.1 DiagVib-6 Benchmark

Building upon the exploratory results in the out-of-distribution setting, the robust model selection capabilities of PA will be assessed across a wide range of distribution shift conditions in a controlled experimental setup. In particular, different shift factors will be considered for both source environments and target environments. These findings will help identify the experimental conditions in which the discriminative power of PA is most effective.

Experiments have been conducted in a similar setting to that of **Experiment 5a-5b**, with a reduced dataset size to avoid repetition of MNIST observations. This ensures that each training observation uniquely represents a specific instance of the number drawing experiment, along with the corresponding domain shift perturbation. Given that factor modifications are not deterministic, this approach prevents the model's inductive bias from being influenced by an implicit data augmentation process.

**Experiment 6.** Model selection experiments in the domain generalization setting have been conducted within the DiagVib-6 [20] data generation pipeline, which was already described in Section 4.3. ERM, IRM [2] and LISA [69] algorithms have been used to train a ResNet18 for 100 epochs on source environments through SGD [51] and Adam [33] optimizers with learning rates $5 \times 10^{-5}$, $10^{-4}$, $5 \times 10^{-4}$ and $10^{-3}$. For each case, model validation and selection has been conducted on increasingly shifted datasets, encompassing both target and source domains, by storing the weights that displayed the best performance on different metrics. More specifically, PA, $\text{AFR}_\text{P}$ and accuracy on validation datasets have been computed every epoch and used as early stopping criteria.

The characterization of the inductive bias is outside of the scope of this work, but given the triviality 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 randomness instantiation, the nature of the shift and their relative frequency in the data. The optimization process will iteratively navigate the loss landscape and implicitly balance these features differently, leading to different predictive outcomes.

In this regard, two primary sources of inductive bias were examined by varying the nature of the shift defining the two source environments, 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 criteria.

For this experiment, two sampling instantiations $\tau_0^{\text{val}} \neq \tau_1^{\text{val}}$ are considered for the validation datasets, which means that $\boldsymbol{x}_0^{\text{val}}$ and $\boldsymbol{x}_1^{\text{val}}$ entail different instantiations of the shift factor and the sampling experiment. Table 5.1 and Table C.5 specify the configuration of factors considered for the training and validation datasets. In the former, source domains are shifted in the hue factor, while in the latter the position factor is considered.

**Dataset** (Training and validation `hue`)**.** Training datasets are always drawn from source domains. More specifically, they are composed of samples $\boldsymbol{x}_0^{\text{train}} \circ \tau_0^{\text{train}}$ and $\boldsymbol{x}_1^{\text{train}} \circ \tau_1^{\text{train}}$, with $\tau_0^{\text{train}} \neq \tau_1^{\text{train}}$. Validation datasets are composed of samples $\boldsymbol{x}_0^{\text{val}} \circ \tau_0^{\text{val}}$, $\boldsymbol{x}_1^{\text{val}} \circ \tau_1^{\text{val}}$, $\tau_0^{\text{val}} \neq \tau_1^{\text{val}}$, which are either drawn from source domains, target domains or a combination of both.

- When both $\boldsymbol{x}_0^{\text{val}}, \boldsymbol{x}_1^{\text{val}}$ are drawn from source domains, two configurations are considered, namely when they are drawn from the same distribution (SD) and when they are not, in which case are still defined in-distribution (ID) with respect to training samples.
- When $\boldsymbol{x}_0^{\text{val}}$ is drawn from the source and $\boldsymbol{x}_1^{\text{val}}$ is drawn from target domains, two more configurations are considered. In particular, depending on the magnitude of the shift entailed by $\boldsymbol{x}_1^{\text{val}}$, 1-factor mixed distribution (1F-MD) and 5-factor mixed distribution (5F-MD) samples can be considered. The former entails a shift in the factor that defines the experiment (i.e. hue), while the latter considers a scenario in which all possible factors are shifted.
- When both $\boldsymbol{x}_0^{\text{val}}, \boldsymbol{x}_1^{\text{val}}$ are drawn from target domains, a single out-of-distribution (OOD) configuration is considered, where both samples entail a different instantiation in the factor that defines the experiment.

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

Table 5.1: Image factors associated with each of the environments considered in this experiment. CC and UL account for 'centered center' and 'upper left', respectively.

**Dataset** (Test `hue`)**.** Source domains are shifted in the hue factor. Consequently, target domains entailed incremental shifts in the remaining factors.

| # Factors | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Hue | red | **green** | green | green | green | green |
| Lightness | dark | dark | **bright** | bright | bright | bright |
| Position | CC | CC | CC | CC | **UL** | UL |
| Scale | large | large | large | large | large | **small** |
| Texture | blank | blank | blank | **tiles** | 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 5.2: Image factors associated to each of the environments considered in this experiment. CC and UL account for 'centered center' and 'upper left', respectively.

For each validation dataset configuration, task performance (i.e. accuracy) will be reported on progressively shifted test datasets, as outlined in Table 5.2 for the hue factor and Table C.3 for the position factor experiments. Specifically, results will be associated with the optimization process that achieved the highest aggregated test accuracy across all the datasets. In other words, the optimization configuration that maximizes the cumulative performance across all test datasets will be reported, regardless of the model selection criterion. Under these conditions, only the best-performing models will be considered, allowing for a focused assessment of the suitability of the model selection criteria.

Tables 5.3-5.4 display the test performance of the selected models across different dataset configurations and model selection criteria. A key observation is that models trained under a shift in the hue factor (Table 5.3) exhibit significantly lower performance compared to those trained under a shift in the position factor (Table 5.4). This performance discrepancy is particularly noticeable in the first shifted test dataset, which contains the same configuration of factors in both experiments and shows a clear performance gap between the two.

Test datasets encode cumulative shifts, and each of them potentially favors specific features in the model's inductive bias. Consequently, models that demonstrate significant performance improvements in some datasets when selected using PA may exhibit the opposite behavior in others. Due to this variability, aggregated performance has been adopted as an experimental criterion to provide a more comprehensive evaluation, and the suitability of a selection metric will be determined based on its consistency across different datasets, not on a particular one.

Both experiments reveal a strong alignment in the model selection capabilities of validation accuracy and $\text{AFR}_\text{P}$. As described in previous chapters, accuracy provides an assessment based on task performance, while $\text{AFR}_\text{P}$ serves as our baseline robustness metric, thus accounting for changes in performance under distribution shifts. The strong alignment in their model selection capabilities is a significant observation, as it underscores that the discriminative power driving PA is fundamentally different from that of accuracy-based metrics.

| | **Acc. Test 0** | | | **Acc. Test 1** | | | **Acc. Test 2** | | | **Acc. Test 3** | | | **Acc. Test 4** | | | **Acc. Test 5** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ERM** | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA |
| SD | 99.4 | 99.4 | **99.5** | 42.1 | 42.1 | **55.0** | **70.4** | **70.4** | 69.7 | 68.0 | 68.0 | **68.3** | 52.6 | 52.6 | **52.7** | **41.7** | **41.7** | 30.2 |
| ID | **99.5** | **99.5** | 99.4 | 49.6 | 49.6 | **69.3** | **90.2** | **90.2** | 87.3 | 81.0 | 81.0 | **82.1** | **71.4** | **71.4** | 69.6 | **44.6** | **44.6** | 39.0 |
| 1F-MD | 99.3 | 99.3 | 99.3 | 47.4 | 47.4 | 47.4 | 78.6 | 78.6 | 78.6 | 70.7 | 70.7 | 70.7 | 63.7 | 63.7 | 63.7 | 36.7 | 36.7 | 36.7 |
| 5F-MD | 99.5 | 99.5 | 99.5 | 49.6 | 49.6 | 49.6 | 90.2 | 90.2 | 90.2 | 81.0 | 81.0 | 81.0 | 71.4 | 71.4 | 71.4 | 44.6 | 44.6 | 44.6 |
| OOD | 99.3 | 99.3 | 99.3 | 60.9 | 60.9 | 60.9 | 92.6 | 92.6 | 92.6 | 83.9 | 83.9 | 83.9 | 64.9 | 64.9 | 64.9 | 32.3 | 32.3 | 32.3 |
| **IRM** | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA |
| SD | 99.3 | 99.3 | 99.3 | 71.5 | 71.5 | **83.3** | 70.6 | 70.6 | **91.9** | 65.3 | 65.3 | **85.9** | **75.0** | **75.0** | 66.5 | 28.8 | 28.8 | **46.7** |
| ID | 99.4 | 99.4 | 99.4 | 44.3 | 44.3 | 44.3 | 88.1 | 88.1 | 88.1 | 76.1 | 76.1 | 76.1 | 59.4 | 59.4 | 59.4 | 45.2 | 45.2 | 45.2 |
| 1F-MD | 99.4 | 99.4 | 99.4 | 44.3 | 44.3 | 44.3 | 88.1 | 88.1 | 88.1 | 76.1 | 76.1 | 76.1 | 59.4 | 59.4 | 59.4 | 45.2 | 45.2 | 45.2 |
| 5F-MD | **99.4** | **99.4** | 99.3 | 31.2 | 31.2 | **83.3** | 88.7 | 88.7 | **91.9** | 73.8 | 73.8 | **85.9** | 65.5 | 65.5 | **66.5** | **50.2** | **50.2** | 46.7 |
| OOD | 99.5 | 99.5 | 99.5 | 62.4 | 62.4 | 62.4 | 90.1 | 90.1 | 90.1 | 84.1 | 84.1 | 84.1 | 52.2 | 52.2 | 52.2 | 42.6 | 42.6 | 42.6 |
| **LISA** | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA |
| SD | 99.3 | 99.3 | **99.4** | **88.4** | **88.4** | 83.1 | 53.9 | 53.9 | **78.5** | 57.9 | 57.9 | **80.5** | 63.1 | 63.1 | **77.0** | **38.8** | **38.8** | 35.4 |
| ID | 99.5 | 99.5 | 99.5 | 59.2 | 59.2 | **88.8** | **62.5** | **62.5** | 60.4 | 71.7 | 71.7 | **76.8** | 70.9 | 70.9 | **71.5** | 35.2 | 35.2 | **36.4** |
| 1F-MD | 99.5 | 99.5 | 99.5 | 88.8 | 88.8 | 88.8 | 54.4 | 54.4 | 54.4 | 76.8 | 76.8 | 76.8 | 71.5 | 71.5 | 71.5 | 36.4 | 36.4 | 36.4 |
| 5F-MD | 99.2 | 99.2 | **99.4** | **86.6** | **86.6** | 85.1 | 71.1 | 71.1 | **74.8** | 77.3 | 77.3 | **83.7** | 73.7 | 73.7 | **81.9** | **49.4** | **49.4** | 48.6 |
| OOD | **99.3** | 99.2 | 99.2 | 91.8 | **95.8** | 84.4 | 59.0 | 63.0 | **83.0** | 77.6 | 79.0 | **88.0** | 73.8 | 73.6 | **84.8** | 34.4 | 40.4 | **47.3** |

Table 5.3: Test performance under increasing levels of shift for models selected through different configurations of validation datasets. PA, $\text{AFR}_\text{P}$ and accuracy are used as early stopping criteria for model selection in the hue factor experiment (see Tables 5.1-5.2).

| | Acc. Test 0 | | | Acc. Test 1 | | | Acc. Test 2 | | | Acc. Test 3 | | | Acc. Test 4 | | | Acc. Test 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ERM** | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA |
| SD | 99.5 | 99.5 | 99.5 | **99.5** | **99.5** | 99.4 | **85.7** | **85.7** | 84.7 | 42.0 | 42.0 | **66.3** | 45.0 | 45.0 | **51.4** | 38.5 | 38.5 | **51.8** |
| ID | 99.3 | 99.3 | **99.4** | **99.5** | **99.5** | 99.4 | 84.9 | 84.9 | **86.6** | 66.3 | 66.3 | **68.9** | **52.0** | **52.0** | 50.2 | 46.3 | 46.3 | **48.1** |
| 1F-MD | 99.3 | 99.3 | **99.4** | **99.5** | **99.5** | 99.4 | 84.9 | 84.9 | **86.6** | 66.3 | 66.3 | **68.9** | **52.0** | **52.0** | 50.2 | 46.3 | 46.3 | **48.1** |
| 5F-MD | **99.6** | **99.6** | 99.1 | **99.4** | **99.4** | 99.1 | 84.3 | 84.3 | **90.1** | **65.2** | 65.2 | 58.6 | 66.2 | 66.2 | **73.3** | 57.4 | 57.4 | **57.6** |
| OOD | **99.5** | **99.5** | 99.4 | **99.5** | **99.5** | 99.4 | **87.0** | **87.0** | 86.6 | **75.1** | **75.1** | 68.9 | **56.9** | **56.9** | 50.2 | **57.5** | **57.5** | 48.1 |
| **IRM** | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA |
| SD | 99.2 | 99.2 | **99.5** | 99.4 | 99.4 | **99.5** | 83.3 | 83.3 | **89.9** | 43.5 | 43.5 | **44.3** | 53.4 | 53.4 | **59.0** | 44.6 | 44.6 | **48.2** |
| ID | 99.3 | 99.3 | 99.3 | 99.5 | 99.5 | 99.5 | 77.7 | 77.7 | 77.7 | 57.4 | 57.4 | 57.4 | 55.9 | 55.9 | 55.9 | 50.4 | 50.4 | 50.4 |
| 1F-MD | 99.3 | 99.3 | 99.3 | 99.5 | 99.5 | 99.5 | **84.6** | **84.6** | 77.7 | **69.8** | **69.8** | 57.4 | **64.4** | **64.4** | 55.9 | **57.5** | **57.5** | 50.4 |
| 5F-MD | 98.8 | 98.8 | 98.8 | 98.8 | 98.8 | 98.8 | 84.6 | 84.6 | 84.6 | 81.8 | 81.8 | 81.8 | 69.5 | 69.5 | 69.5 | 63.1 | 63.1 | 63.1 |
| OOD | 99.3 | 99.3 | 99.3 | 99.5 | 99.5 | 99.5 | 84.6 | 84.6 | 84.6 | 69.8 | 69.8 | 69.8 | 64.4 | 64.4 | 64.4 | 57.5 | 57.5 | 57.5 |
| **LISA** | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA |
| SD | 99.3 | 99.3 | 99.3 | 99.4 | 99.4 | 99.4 | 85.3 | 85.3 | 85.3 | 56.5 | 56.5 | 56.5 | 53.9 | 53.9 | 53.9 | 44.2 | 44.2 | 44.2 |
| ID | **99.5** | **99.5** | 99.3 | 99.3 | 99.3 | **99.4** | 62.3 | 62.3 | **85.3** | 50.6 | 50.6 | **56.5** | 52.9 | 52.9 | **53.9** | 42.9 | 42.9 | **44.2** |
| 1F-MD | 99.3 | 99.3 | 99.3 | 99.4 | 99.4 | 99.4 | 85.3 | 85.3 | 85.3 | 56.5 | 56.5 | 56.5 | 53.9 | 53.9 | 53.9 | 44.2 | 44.2 | 44.2 |
| 5F-MD | **99.3** | **99.3** | 99.0 | **99.2** | **99.2** | 99.1 | **80.0** | **80.0** | 77.5 | **88.6** | **88.6** | 77.2 | 68.4 | 68.4 | **73.7** | 64.5 | 64.5 | 64.5 |
| OOD | **99.5** | **99.5** | 99.3 | 99.3 | 99.3 | **99.4** | 62.3 | 62.3 | **85.3** | 50.6 | 50.6 | **56.5** | 52.9 | 52.9 | **53.9** | 42.9 | 42.9 | **44.2** |

Table 5.4: Test performance under increasing levels of shift for models selected through different configurations of validation datasets. PA, $\text{AFR}_\text{P}$ and accuracy are used as early stopping criteria for model selection in the position factor experiment (see Tables C.5-C.3).

Overall, models selected through PA demonstrate equal or superior performance under increasing levels of domain shift. In particular, ID and SD dataset configurations, which address the domain adaptation problem for being limited to source-related factor configurations, show substantial improvements compared to those chosen by accuracy-based metrics. When PA-based selection underperforms, it does so by a small margin, whereas in cases where it outperforms, the margin of improvement is much larger.

Alternative results were obtained using the same factor configuration but with a single instantiation of the sampling experiment $\tau^{\text{val}}$, in a way that samples $\boldsymbol{x}_0^{\text{val}}$ and $\boldsymbol{x}_1^{\text{val}}$ are composed of the same MNIST observations. Tables B.1-B.2 present these results, from which no conclusive evidence can be drawn. More specifically, accuracy and $\text{AFR}_\text{P}$ are not aligned in their selection of models, and neither of the metrics provides a consistent superior model selection capabilities.

These experiments yield some important and a priori counterintuitive conclusions. First, they demonstrate that PA is more effective for model selection in the presence of sampling randomness; that is, when posterior agreement is maximized between different observations of the same class. One might intuitively expect that filtering out sampling randomness would lead to selecting a model encoding the most robust set of features with respect to the distribution shift, and that these features would be proven to generalize under target domain shifts, regardless of the particular sampling instance. However, these results suggest that different instances of the same class improve PA model selection performance, even when source environments are accounted for in validation and overfitting to irrelevant features is mitigated. Besides, PA is shown to be more effective in in-distribution model selection settings (i.e. ID and SD), and does not always outperform accuracy-based metrics in mixed-distribution or out-of-distribution scenarios, at least in the ERM and IRM cases.

## 5.2 In-distribution DiagVib-6

Thus far, experiments in the domain generalization setting have been conducted under synthetic conditions, where the effect of the shift was uniform across the sample and the presence of predictive and non-predictive factors was balanced. These experiments have demonstrated that PA successfully discriminates robust from non-robust learners and also provides increased early-stopping performance compared with current baseline metrics, particularly in domain adaptation conditions.

However, real-world datasets are subject to heteroscedastic sources of 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 shifts. 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 feature representations. These representations may generalize well to sampling variability within source environments but fail to adapt to distribution shifts in target environments, which poses an additional challenge to the domain adaptation problem.

**Experiment 7.** Let shape, hue, lightness, position, scale and texture be MNIST image factors that can be manipulated throught 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-occurrent 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-occurrent with other instances of $F^L$ and $F^P$, respectively, thus breaking the exclusivity condition. In this work, three particular settings will be considered:

- Zero generalization opportunities (ZGO): Each instance of $F^P$ is exclusively co-occurrent 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-occurrent with all instances of $F^L$, so that all combinations of factors are present in source domains.

The setting for ID model selection (i.e. domain adaptation) 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.

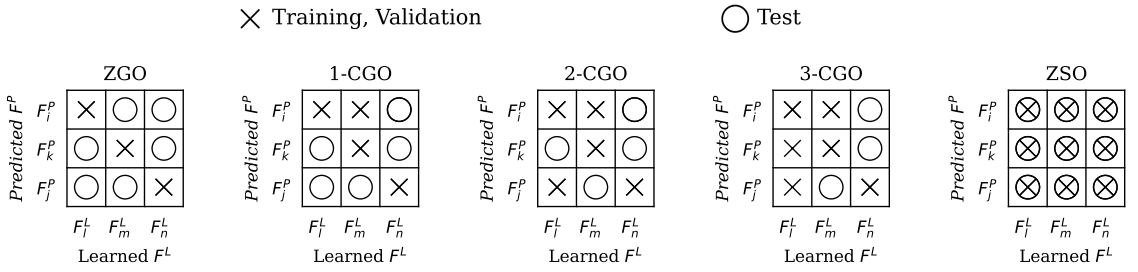



Figure 5.1: Representation of the co-occurrence pattern in between learning factors $F^L$ and predicted factors $F^P$ for the ZGO, CGO and ZSO settings that will be considered in this experiment.

In this case, only ERM and IRM [2] algorithms will be considered, since data augmentation algorithms such as LISA [69] would bypass the synthetically-induced co-occurrence patterns in the datasets, thus preventing a proper analysis of the PA model selection capabilities under these specific conditions. These algorithms will be used to train a ResNet18 model for 100 epochs using Adam optimizer with learning rates $10^{-2}$ and $10^{-3}$.

In order to encompass the primary drivers of inductive bias from a perceptual perspective, these experiments will be conducted for both hue and position as learning factors $F^L$ and shape as the predicted factor $F^P$. The specific configuration of factor values and co-occurrence patterns that defines each dataset can be found in Table C.2.2. As has been the case throughout this project, test performance will be reported in increasingly shifted test datasets, as specified in Table C.6. In this case, however, selected models will be those achieving the highest accuracy in the first test dataset only, which is the one assessing the performance in the complementary co-occurring learning factors, and therefore determining whether the selected model has been able to overcome the limitations posed by the suboptimal configuration of the data. In total, training, validation and test samples are composed of 30.000, 15.000 and 3.000 MNIST observations, respectively.

Tables 5.5-5.6 present the results of this experiment by reporting the test performance obtained through accuracy-based model selection (Acc) and the improvement in performance obtained through PA-based model selection (ΔAcc). These results show that PA is able to effectively select models that generalize better to unseen combinations of factors, particularly in the first environment, for both hue and position factor experiments.

This setting highlights the potential of robustness-driven model selection, as Tables B.3-B.4 show that similarly positive results are be obtained with $\text{AFR}_\text{P}$ as early stopping criterion. Figure 5.2 and Table 5.7 expand these insights by illustrating the rationale behind the robustness assessment provided by PA. Under heavy spurious overfitting, mismatch between predictions in $\boldsymbol{x}_0^{\text{test}}$ and $\boldsymbol{x}_1^{\text{test}}$ is very significant, and in most cases models are considered to be non-robust. Nevertheless, even when the performance difference between ERM and IRM models is minimal, $\beta^*$ starts converging to non-zero values at a lower co-occurrence threshold in the IRM case, if compared to ERM. This underscores the superior robustness-fostering capabilities of IRM.

| | **Test 1** | | **Test 2** | | **Test 3** | | **Test 4** | | **Test 5** | |
|---|---|---|---|---|---|---|---|---|---|---|
| **ERM** | Acc. | ΔAcc. | Acc. | ΔAcc. | Acc. | ΔAcc. | Acc. | ΔAcc. | Acc. | ΔAcc. |
| ZGO | 53.2 | ±0.01 | 54.6 | ±0.01 | 55.7 | ±0.01 | 66.7 | ±0.01 | 66.6 | ±0.01 |
| 1-CGO | 62.9 | **+9.5** | 64.7 | **+10.2** | 60.8 | **+0.3** | 62.9 | **+2.2** | 64.2 | **+0.5** |
| 2-CGO | 69.1 | **+9.4** | 71.2 | **+7.8** | 71.9 | **+0.3** | 76.2 | **-2.4** | 77.0 | **-2.8** |
| 3-CGO | 73.1 | **+16.6** | 85.6 | **+3.6** | 70.1 | **+9.7** | 71.4 | **+6.4** | 72.1 | **+6.7** |
| ZSO | 99.6 | ±0.01 | 92.8 | **-0.1** | 89.9 | **+0.2** | 85.9 | ±0.01 | 85.9 | ±0.01 |
| **IRM** | Acc. | ΔAcc. | Acc. | ΔAcc. | Acc. | ΔAcc. | Acc. | ΔAcc. | Acc. | ΔAcc. |
| ZGO | 50.1 | **+5.9** | 50.5 | **+4.9** | 52.8 | **+9.5** | 64.4 | **+1.1** | 64.9 | **+1.2** |
| 1-CGO | 63.0 | **+7.0** | 65.9 | **+7.6** | 59.4 | **+2.2** | 60.1 | **+2.2** | 59.0 | **+1.8** |
| 2-CGO | 69.0 | **+10.6** | 69.7 | **+10.0** | 67.5 | **+4.7** | 64.5 | **+13.0** | 65.1 | **+12.6** |
| 3-CGO | 79.5 | **+11.6** | 83.0 | **+9.8** | 73.6 | **+10.9** | 70.7 | **+11.0** | 72.2 | **+11.3** |
| ZSO | 99.4 | **+0.1** | 93.4 | **+1.3** | 89.2 | **+0.2** | 87.0 | **+1.6** | 87.0 | **+1.6** |

Table 5.5: Test performance under increasing levels of shift for models selected through different configurations of factor co-occurrence for the hue learning factor experiment. Specifically, the performance of models selected through validation accuracy (Acc) and the difference between accuracy-based and PA-based selection (ΔAcc) is reported. PA is able to select models that perform better than the one selected through accuracy in the most cases.

| ERM | Test 1 Acc. | Test 1 ΔAcc. | Test 2 Acc. | Test 2 ΔAcc. | Test 3 Acc. | Test 3 ΔAcc. | Test 4 Acc. | Test 4 ΔAcc. | Test 5 Acc. | Test 5 ΔAcc. |
|---|---|---|---|---|---|---|---|---|---|---|
| ZGO | 50.2 | ±0.01 | 50.0 | ±0.01 | 52.4 | ±0.01 | 52.0 | ±0.01 | 50.6 | ±0.01 |
| 1-CGO | 44.1 | **+4.2** | 41.4 | **+0.5** | 43.6 | **+1.5** | 45.0 | **+1.0** | 44.8 | **+1.2** |
| 2-CGO | 64.4 | **+7.0** | 53.4 | **+1.7** | 55.1 | **+2.7** | 57.7 | **+1.3** | 56.7 | **+1.3** |
| 3-CGO | 73.2 | **+17.9** | 55.3 | **+8.3** | 52.8 | **+8.5** | 51.8 | **+7.0** | 51.8 | **+7.2** |
| ZSO | 99.1 | ±0.01 | 92.1 | **-0.2** | 88.6 | **+0.2** | 87.0 | **+0.1** | 85.7 | **+0.2** |
| **IRM** | Acc. | ΔAcc. | Acc. | ΔAcc. | Acc. | ΔAcc. | Acc. | ΔAcc. | Acc. | ΔAcc. |
| ZGO | 50.5 | ±0.01 | 50.0 | ±0.01 | 53.1 | ±0.01 | 52.5 | ±0.01 | 51.6 | ±0.01 |
| 1-CGO | 48.2 | **+0.8** | 43.6 | **+0.6** | 43.2 | **-0.4** | 44.0 | **-1.7** | 44.1 | **-2.1** |
| 2-CGO | 65.0 | **+7.1** | 53.1 | **+5.6** | 52.8 | **+1.3** | 59.2 | **-3.7** | 60.1 | **-4.7** |
| 3-CGO | 80.4 | **+13.8** | 63.2 | **+7.7** | 57.5 | **+6.0** | 51.4 | **+7.4** | 52.6 | **+5.2** |
| ZSO | 99.6 | **-0.3** | 93.8 | **-1.0** | 88.2 | **-2.5** | 89.5 | **-8.1** | 89.6 | **-8.4** |

Table 5.6: Test performance under increasing levels of shift for models selected through different configurations of factor co-occurrence for the position learning factor experiment. Specifically, the performance of models selected through validation accuracy (Acc) and the difference between accuracy-based and PA-based selection (ΔAcc) is reported. PA is able to select models that perform better than the one selected through accuracy in the most cases.

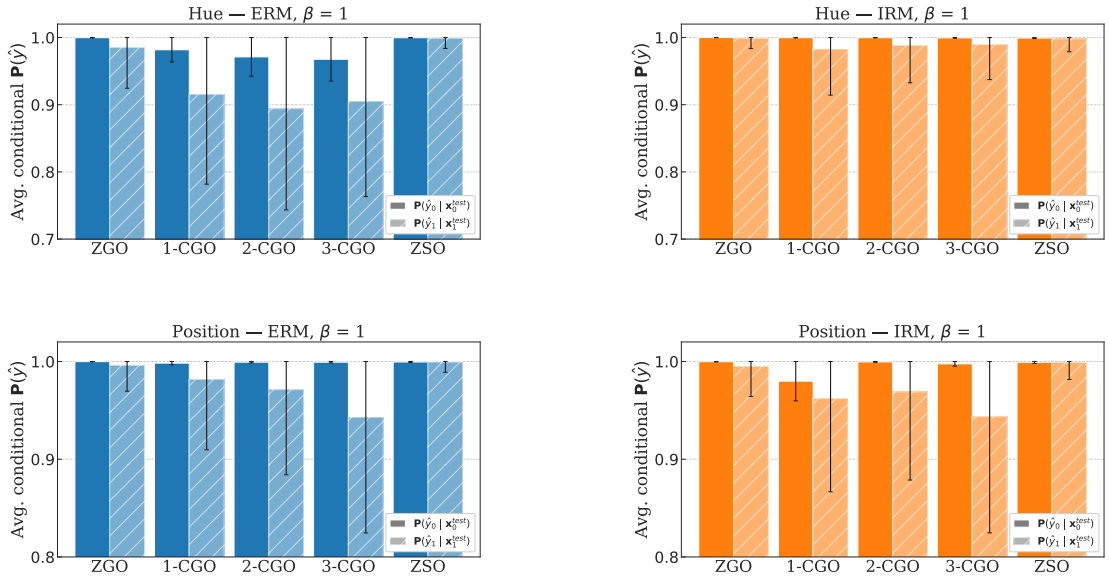



Figure 5.2: Average posterior distribution at the predicted class for the different ZGO, CGO and ZSO settings for the hue and position learning factor experiments. The posterior distribution is computed for the first test dataset, which is the one assessing the performance in the complementary co-occurring learning factors.

| Hue | **ZGO** | **1-CGO** | **2-CGO** | **3-CGO** | **ZSO** |
|---|---|---|---|---|---|
| **ERM** | 0.000 | 0.000 | 0.000 | 0.000 | 85.98 |
| **IRM** | 0.000 | 0.000 | 2.110 | 17.34 | 116.3 |
| position | **ZGO** | **1-CGO** | **2-CGO** | **3-CGO** | **ZSO** |
| **ERM** | 0.000 | 0.000 | 0.000 | 0.000 | 1720 |
| **IRM** | 0.000 | 0.000 | 3.984 | 34.107 | 2254 |

Table 5.7: A comparison of the $\beta^*$ values obtained in **Experiment 7** reveals that ERM models are entirely non-robust up until the ZSO setting, whereas IRM models demonstrate generalization capabilities (i.e. $\beta^* > 1$) starting from the 1-CGO setting.

## 5.3 WILDS Benchmark

In light of the results obtained in this chapter, the model selection capabilities of PA will be finally assessed on standard benchmark datasets in out-of-distribution settings. In particular, several WILDS [34] datasets will be considered, as they offer a comprehensive set of domain generalization tasks that is representative of real-world scenarios and are widely used in the literature to report performance in robust learning.

Every dataset under consideration entails a specific configuration of learning opportunities that will shift the inductive bias towards suboptimal representations. The domain generalization performance of the models will be assessed on target domains and the selection capabilities of the different metrics will be compared. In particular, only domain adaptation settings will be considered, as PA has been proven to consistently outperform accuracy-based metrics in these conditions.

Every experiment is associated to a specific WILDS dataset, where the data configuration, network architecture and optimization settings were drawn from either the original publication or from the configurations specified for IRM [2] and LISA [69] algorithms in their respective works. To ensure consistency, the setup and construction of the required environments strictly adhered to the guidelines outlined in these publications.

With regard to PA model selection, it must be noted that samples associated to each domain in the validation set were initially not sorted or paired in any way. For the purpose of these experiments, and given the results obtained in previous sections, observations from validation samples $\boldsymbol{x}_0^{\text{val}}, \boldsymbol{x}_1^{\text{val}}$ were paired by performing a nearest-neighbour search in the feature space after the first training epoch, conditioned on class membership. In that way, posterior agreement would be computed in a similar setting to the one considered in the synthetic experiments, under both sampling randomness (even if not entirely random due to the pairing) and domain shift. This pairing method is clearly suboptimal, as it is biased by the specific representation of features delivered after the first training epoch. However, it was selected for efficiency purposes, and it was conducted using the L2 Euclidean distance implementation of the FAISS [21] library.

**Experiment 8.** The **waterbirds** dataset was considered. The classification task involves the prediction of two classes of birds, namely `waterbird` and `landbird`. The dataset is balanced in the classes of birds, but images encode a spurious correlation that arises from their different habitats. In particular, the configuration considered in this experiment is such that the background of the images is unbalanced in the train dataset, which contains 3554 images with water background and only 1241 images with land background. Both validation and test datasets are balanced, and the shortcut opportunity encoded by the background subpopulation shift will be exploited. More specifically, source domains will be defined from the background type, and robustness and performance in between these will be considered as model selection criteria.

**Experiment 9.** The **celebA** dataset was considered. The classification task involves the prediction of hair color from images of American celebrities, namely `blonde` and `not blonde`. The subpopulation shift arises from the spurious correlation existing between the gender of the celebrity and the color of their hair. In particular, Table 5.8 contains the relative frequency of the gender-hair color combinations in the training dataset. Source domains will be defined from the hair color (i.e. the label), and robustness and performance in between these will be considered as model selection criteria.

| | `blonde` | `not blonde` |
|---|---|---|
| Male | 1741 | 89931 |
| Female | 28234 | 82685 |

Table 5.8: Relative frequency of the gender-hair color combinations in the `celebA` [34] dataset. The `blonde` class is underrepresented, especially in male pictures.

**Experiment 10.** The **camelyon17** dataset was considered. The classification task involves the identification of tumor tissue in lymph node patches sampled from different hospitals, which amounts to the classes `tumor` and `no tumor`. The out-of-distribution setting originates from the differences in the samples taken from different hospitals, as illustrated in Figure 1.5. In particular, the composition of the training, validation and test datasets is specified in Table 5.9. Each of the three first hospitals defines a source environment, and the generalization capabilities over observations in the fourth hospital will be examined.

| Hospital | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Training | 53425 | 116959 | 132052 | - |
| Validation | 6011 | 12879 | 14670 | - |
| Test | - | - | - | 85054 |

Table 5.9: Hospital of origin of the lymph node patches that compose training, validation and test datasets in the `camelyon17` [34] dataset.

| | **Average Acc.** | | | **Worst-case Acc.** | | |
|---|---|---|---|---|---|---|
| | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA |
| **ERM** | **78.52** | 73.72 | **78.52** | **67.11** | 58.90 | **67.11** |
| **IRM** | **90.16** | 89.70 | **90.16** | **89.21** | 88.67 | **89.21** |

Table 5.10: Average and worst-case test accuracy for the `waterbirds` [34] dataset. PA outperforms $\text{AFR}_\text{P}$ due to the vulnerability of the latter to the specific sampling instantiation. More specifically, while PA has a monotonically increasing evolution during training, $\text{AFR}_\text{P}$ is only monotonic in mean, and the variation around the mean at every epoch hinders its selection consistency.

| | **Average Acc.** | | | **Worst-case Acc.** | | |
|---|---|---|---|---|---|---|
| | Acc. | $\text{AFR}_\text{P}$ | PA | Acc. | $\text{AFR}_\text{P}$ | PA |
| **ERM** | 97.80 | **98.27** | **98.27** | 97.76 | **97.77** | **97.77** |
| **IRM** | 98.12 | **98.41** | **98.41** | **98.06** | 98.05 | 98.05 |

Table 5.11: Average and worst-case test accuracy for the `celebA` [34] dataset. Robustness-based selection metrics outperform accuracy in most cases.

| | Accuracy | | |
|---|---|---|---|
| | Acc. | $AFR_P$ | PA |
| **ERM** | 86.73 | 86.73 | 86.73 |
| **IRM** | 67.98 | 67.98 | 67.98 |
| **LISA** | 81.1 | **81.8** | **81.8** |

Table 5.12: Test accuracy for the `camelyon17` [34] dataset. No significant improvement is observed, with the exception of the LISA model.

Tables 5.10-5.12 display the average performance and worst-case performance (when applicable) obtained on test datasets for the different model selection criteria. Overall, PA-based model selection shows the best results, as it is consistently able to select models that generalize better to test data, while accuracy and $AFR_P$ are shown to underperform in some settings. However promising, these results should be taken with caution and further experimentation should be pursued to fully determine the generalizability of this approach to other datasets and other possible configurations of source domains.

# Chapter 6

# Conclusions

The main goal of this project was the assessment of the suitability of the Posterior Agreement framework as a robust model selection criterion within the context of deep learning models for image classification tasks. The work entailed deriving and implementing an operative version of PA for discrete hypothesis classes, exploring its properties, and comparing its discriminative power against that of baseline accuracy-based metrics. These objectives guided the research process, shaping the experiments and analyses that ultimately lead to these conclusions.

The first step towards evaluating the suitability of PA as a robustness metric involved its characterization under the three sources of randomness that are relevant in image classification tasks, namely sampling randomness (i.e. in-distribution setting), adversarial perturbations and distribution shifts, which were described and formalized in Chapter 3.

**Experiment 1** concerned the empirical exploration of the properties of PA by artificially generating predictions from random, perfect, and constant classifiers. Results showed that PA complies with the desired properties of a robustness metric (see Proposition 3.2) and successfully discriminates the random classifier, which is maximally unrobust, from the perfect and constant classifiers, which are robust by definition. The PA score aligned with this intuition and provided a consistent assessment that was independent of the task performance (see Figure 4.1).

These results motivated the exploration of more realistic in-distribution scenarios, specifically to evaluate the sensitivity of the metric under the presence of noise in the data. **Experiment 2** and **Experiment 3** involved two different models for two different learning tasks in which samples were perturbed with random noise. In both cases, PA was shown to correlate non-linearly with the performance of the model and to display higher sensitivity to small perturbations, even the output prediction of the classifiers was not affected (Figures 4.3-4.5).

In light of the properties displayed by PA, the extension of the robustness measurement to the adversarial setting was considered in **Experiment 4**. This setting involved the assessment of the discriminative power of PA under adversarial perturbations generated through PGD and FMN attacks in the CIFAR10 dataset (see Figure 4.6). PA was shown to be highly sensitive to perturbations and provided consistent discriminative power under increasing levels of attack power and attack ratio, distinguishing clearly between robust and non-robust defenses (see Figure 4.7, 4.9).

The robustness assessment in the adversarial setting was further analyzed by breaking down PA contributions for correctly classified original observations, misclassified original observations and misleading adversarial observations. The comparison of $\beta^*$ between defenses and the analysis of the average posterior probability in these cases was shown to expand the understanding of the behavior of models under covariate shift, and to provide additional discriminative power based on the informativeness of the posterior (see Figures 4.8, 4.11).

An approximation of the PA contributions in this setting effectively distinguished the primary sources of randomness influencing the PA score in each experiment, thereby providing a clearer understanding of the sources of robust and unrobust behavior measured (see Tables 4.3-4.4). The results indicated that sampling randomness was the main contributor to the PA score for poorly effective attacks like PGD, therefore aligning its assessment with $\mathrm{AFR}_{\mathrm{T}}$. In contrast, generalization error to adversarial perturbations was shown to dominate the PA score for highly effective attacks like FMN, thus aligning instead with $\mathrm{AFR}_{\mathrm{P}}$.

Robustness assessment in the out-of-distribution setting was considered in **Experiment 5a-5b**, where the DiagVib-6 data generation pipeline was leveraged to generate synthetic datasets in which shifts were defined as modification of certain image factors for a MNIST digit prediction task. Datasets were subject to both sampling randomness and domain shift, and also to domain shift only, and robustness was evaluated through PA on performance-selected models. PA was shown to possess a superior discriminative power with respect to accuracy-based metrics (see Figures 4.14-4.15). In particular, PA was able to discriminate models under different levels of shift power and shift ratio, and to be sensitive to the presence of both sampling randomness and covariate shift. PA was able to discriminate the most robust model in the source domain, and then provide a consistent assessment on the most robust model on target domains (see Tables 4.7-4.8).

The successful results on robustness assessment motivated the exploration of PA as an early-stoppping criterion for robust model selection. From a data-agnostic perspective, this approach presented two main limitations. First, the vulnerability of robustness-based assessments to overfitting to features that are not relevant for the learning task. Second, the possibility of overfitting to unsuitable biases encoding spurious correlations that are only manifested in source domains, and therefore not be suitable for domain adaptation tasks. Two main experiments were designed to evaluate these conditions by leveraging the DiagVib-6 synthetic data generation pipeline.

In **Experiment 6**, the model selection capabilities of PA were assessed in a set of validation datasets having increasing access to target domains (see Tables 5.1-5.2). Results showed that PA consistently outperformed accuracy-based metrics in settings in which validation datasets were subject to both sampling randomness and domain shift instantiations (see Tables 5.3-5.4). This behavior was found to be consistent across image factors and learning algorithms.

In **Experiment 7**, the inductive bias of the model was manipulated to overfit to specific co-occurrences of image factors and the model selection capabilities of PA were assessed in this context (see Figure 5.1). More specifically, a zero-generalization dataset was initially considered and then iteratively ampliated to contain more generalizable features. In all of these settings, robustness-based model selection, and in particular PA-based selection, was shown to improve performance on shifted test sets that contained a complementary co-occurrence data configuration.

Finally, the domain adaptation capabilities of PA were assessed on three WILDS benchmark datasets, each encompassing a specific subpopulation shift or out-of-distribution setting to which the model's generalization capabilities were assessed (see **Experiment 8-10**). Even if these results have a merely exploratory purpose, PA was shown to select the best performing models in all scenarios, sometimes aligning with performance-based criteria and others with robustness-based criteria instead (see Tables 5.10-5.12).

All in all, both algorithm selection and model selection by means of PA was conducted on a wide range of settings, encompassing all possible sources of generalization error in image classification tasks. PA was shown to be sensitive to different sources of randomness and to possess a superior discriminative power and consistency with respect to baseline accuracy-based metrics. Furthermore, PA demonstrated superior model selection capabilities in the same-distribution and in-distribution settings, in both synthetic experiments and benchmark datasets.

## Future work

The results presented in this work provide a solid foundation for future research in PA-based robustness assessment and model selection in discrete, finite hypothesis class problems. In particular, some areas of interest for future work include PA-driven training, PA-based cross-validation and the analysis of the phenomenon of double descent through the lens of posterior agreement.

Regarding PA-driven training, results suggest that PA favors classifiers with desirable inductive biases, aligning with the common understanding of what constitutes a good model. This opens the possibility of using PA as a guiding criterion for optimizing deep learning models, potentially replacing performance-based metrics when tuning parameters such as learning rates or regularization weights.

PA-based cross-validation could also be an interesting area of research. This work provides evidence in favor of the model selection capabilities of PA in same-distribution settings. In that sense, the maximization of PA could be considered the objective of a cross-validation pipeline for parameter tuning, model selection or hyperparameter optimization, in which instances of the same model with different subsets of training data are expected converge to the same distribution over the hypothesis space.

Finally, the exploration of the phenomenon of double descent is perhaps the most intriguing yet challenging avenue for future work. In deep learning models, the hypothesis class could be defined as a (re-)parametrization of the space of weights of the model. In such case, the nature of the inductive bias driving models further ahead of the interpolation threshold could be analyzed from the perspective of the posterior agreement. In particular, by assessing the values of PA obtained when comparing models with similar performance at both sides of the interpolation threshold.

# 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 $\{1, \ldots, K\}$. 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 $\{1, \ldots, K\}^N$. Therefore, we identify every function $c \in \mathcal{C}$ with the tuple $(c(1), \ldots, c(N))$. Conversely, we identify every tuple $(c_1, \ldots, c_N) \in \{1, \ldots, K\}^N$, 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:

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

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

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

Using Lemma 3.6.1 we can rewrite the denominator as:

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

Therefore, the posterior distribution can be written as:

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

□

## 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 $X''$.

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

*Proof.* Trivial, commutative property. □

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

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

*Proof.* See Lemma 2.3.1. □

**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:

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

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

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

We will require three important results from optimization theory:

**T1** The minimum of $G(\beta) = -\mathrm{PA}(X', X'')$ 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:

$$G(\beta) = -k(\boldsymbol{x}', \boldsymbol{x}'') = \sum_{n=1}^{N} -\log \left[ \sum_{j=1}^{K} e_j' e_j'' \right] + \sum_{n=1}^{N} \log \left[ \sum_{k=1}^{K} e_k' \sum_{p=1}^{K} e_p'' \right]$$

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:

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

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

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 $\boldsymbol{x}', \boldsymbol{x}''$. 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

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

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 $e_j = 0\ \forall j = \{1, ...K\}$, which is not possible in a classification setting since $\beta \in \mathbb{R}^+$. We end up with the following expression:

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

Now we proceed analogously with the second term:

$$G_2^n(\beta) = G_2(\beta) = \log\left[\sum_{j=1}^{K} e'_j \sum_{k=1}^{K} e''_k\right] = \log\left[\sum_{k=1}^{K}\sum_{j=1}^{K} e'_j e''_k\right]$$

$$\frac{d}{d\beta}G_2(\beta) = \frac{\sum_{k=1}^{K}\sum_{j=1}^{K}(F'_j + F''_k)e'_j e''_k}{\sum_{k=1}^{K}\sum_{j=1}^{K} e'_j e''_k}$$

$$\frac{d^2}{d^2\beta}G_2(\beta) = \frac{\sum_{k=1}^{K}\sum_{j=1}^{K}(F'_j + F''_k)^2 e'_j e''_k}{\sum_{k=1}^{K}\sum_{j=1}^{K} e'_j e''_k} - \frac{\left(\sum_{k=1}^{K}\sum_{j=1}^{K}(F'_j + F''_k) e'_j e''_k\right)^2}{\left(\sum_{k=1}^{K}\sum_{j=1}^{K} e'_j e''_k\right)^2} > 0 \iff$$

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

$$\iff \sum_{k=1}^{K}\sum_{q=1}^{K}\sum_{j=1}^{K}\sum_{i=1}^{K}(F'_j + F''_k)^2 e'_j e''_k e'_i e''_q - (F'_j + F''_k) e'_j e''_k (F'_i + F''_q) e'_i e''_q > 0 \iff$$

$$\iff \sum_{k=1}^{K}\sum_{q=1}^{K}\sum_{j=1}^{K}\sum_{i=1}^{K}(F'_j + F''_k) e'_j e''_k e'_i e''_q [(F'_j + F''_k) - (F'_i + F''_q)] > 0$$

We can define as well:

$$\frac{d^2}{d^2\beta}G_2(\beta) = \sum_{k=1}^{K}\sum_{q=1}^{K}\sum_{j=1}^{K}\sum_{i=1}^{K} S_{(jk),(iq)} = \sum_{k=1}^{K}\sum_{q=1}^{K}\sum_{j=1}^{K}\sum_{i=1}^{K}(F'_j + F''_k) E_{(jk),(iq)} \Delta_{(jk),(iq)}$$

$$E_{(jk),(iq)} = e'_j e''_k e'_i e''_q = E_{(ik),(jq)} = E_{(jq),(ik)} = E_{(iq),(jk)}$$

$$\Delta_{(jk),(iq)} = (F'_j - F'_i) + (F''_k - F''_q) = -\Delta_{(iq),(jk)}$$

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

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

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

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

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:

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

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

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

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

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

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

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

□

**Properties.** $\mathrm{PA}\,(\boldsymbol{x}', \boldsymbol{x}''; \beta)$ 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:

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

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

$$\mathbf{P}^c_i(j \mid x'_i) = \mathbf{P}^c_i(j \mid x''_i) = \delta_j(y_i) \Longrightarrow \lim_{\beta\to\beta^*=\infty}\mathrm{PA}\,(\boldsymbol{x}', \boldsymbol{x}''; \beta) = 0$$

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

$$\mathbf{P}^c_i(j \mid x'_i) = \mathbf{P}^c_i(j \mid x''_i) = \delta_j(0) \Longrightarrow \lim_{\beta\to\beta^*=\infty}\mathrm{PA}\,(\boldsymbol{x}', \boldsymbol{x}''; \beta) = 0$$

# Appendix B

# Supplementary Results

## B.1 Robustness assessment

### B.1.1 In-distribution setting

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.

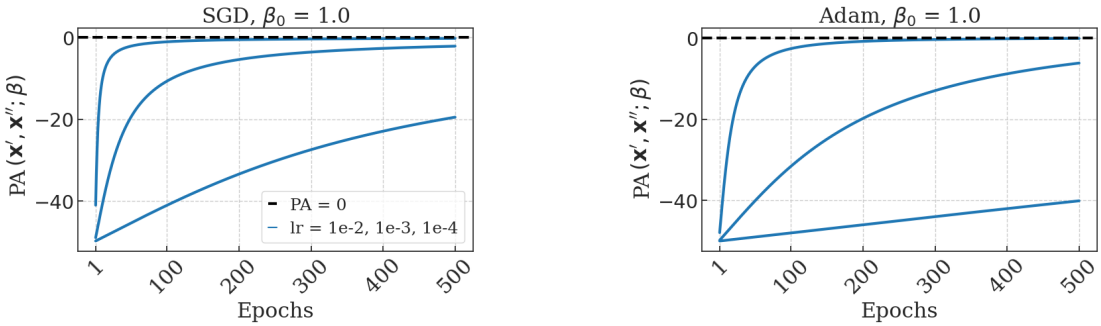



Figure B.1: Evolution of the $\beta$ optimization for a robust sample.

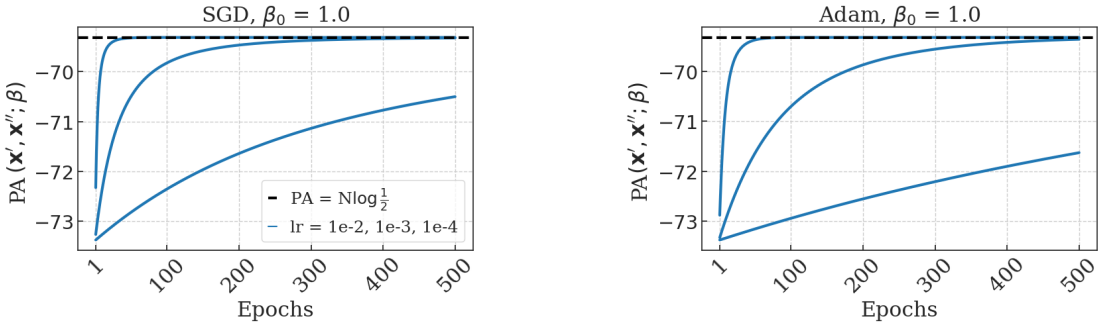



Figure B.2: Evolution of the $\beta$ optimization for a non-robust sample.

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.

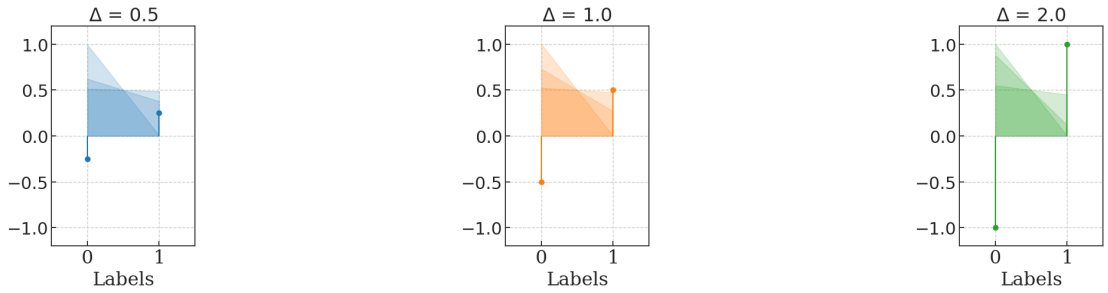



Figure B.3: Logit distributions associated with the behaviour observed in Figure 4.2.

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.

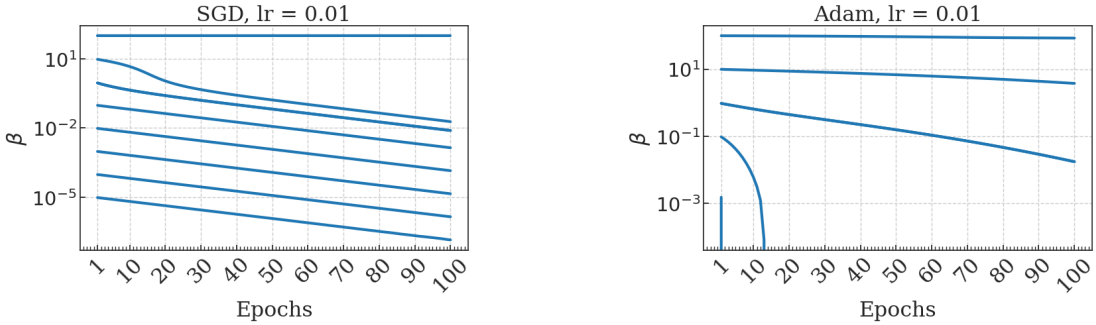



Figure B.4: Evolution of $\beta$ optimization for different initial values for a non-robust classifier.

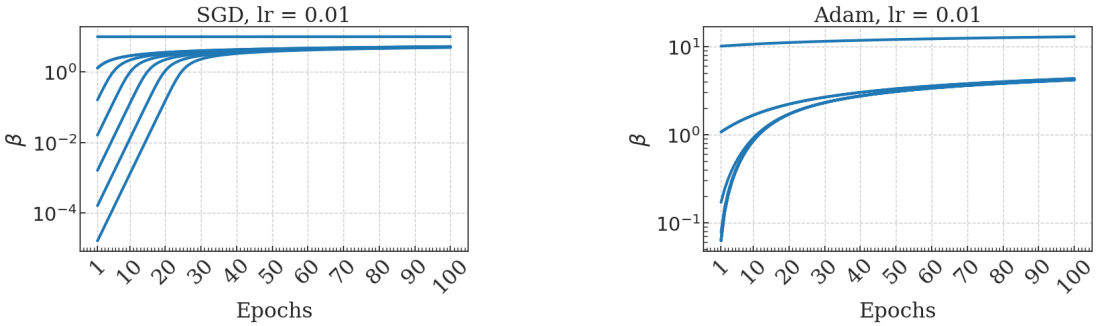



Figure B.5: Evolution of $\beta$ optimization for different initial values for a robust classifier.

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.

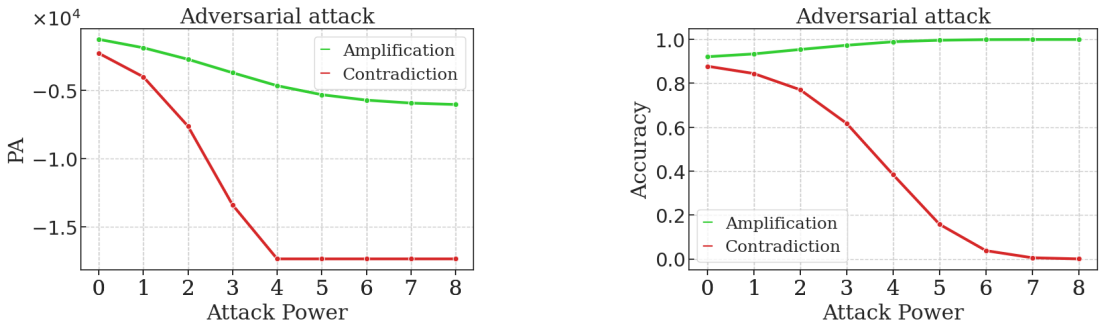



Figure B.6: PA and accuracy for the IMDB sentiment classification task under simple adversarial attacks. Observations are perturbed by replacing some words with positive or negative adjectives that either encourage (amplification) or discourage (contradiction) the true sentiment of the review. The attack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

### B.1.2 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 $\text{AFR}_\text{P}$, 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.

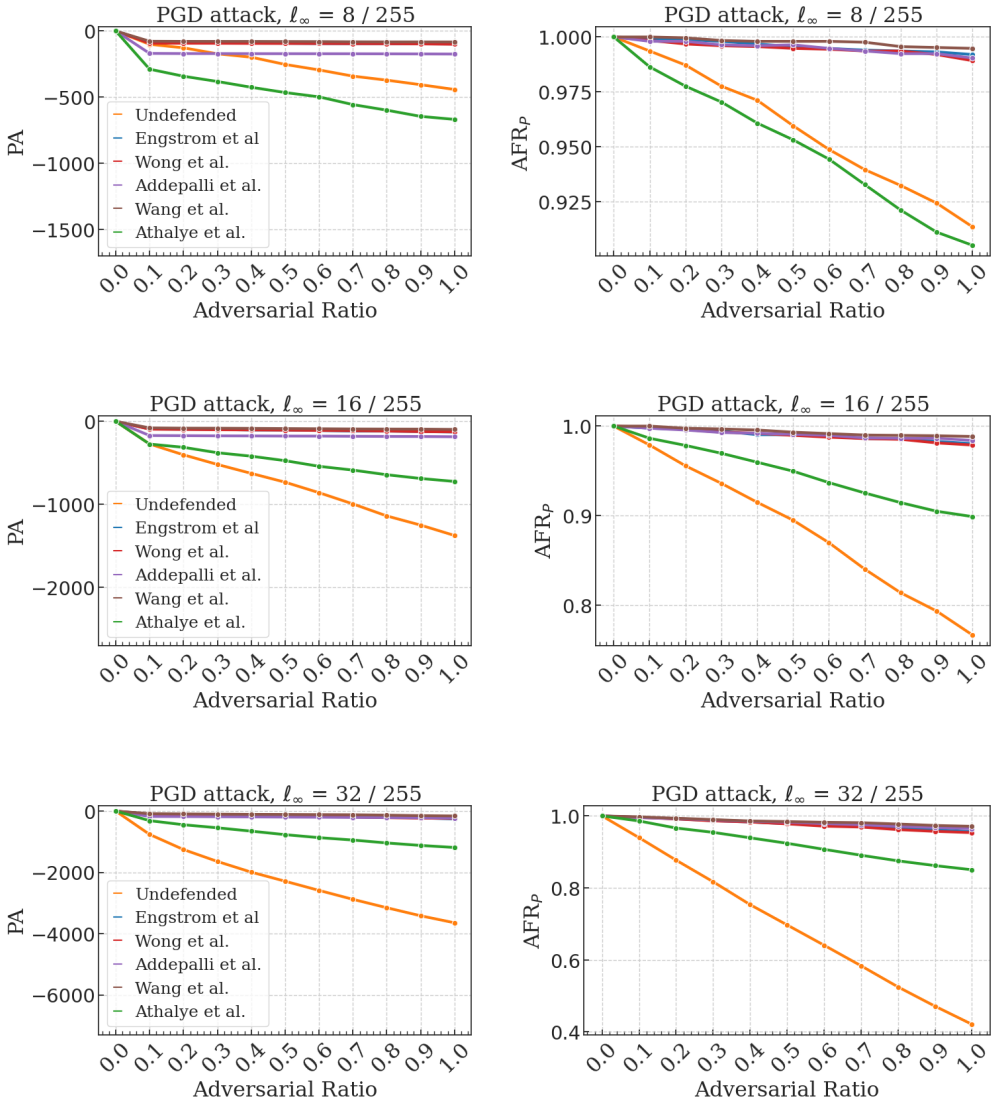



Figure B.7: PA and $\text{AFR}_\text{P}$ variation under increasing adversarial ratio at different perturbation norm bounds. The undefended net and several RobustBench robust models are considered against a 1000 step PGD attack.

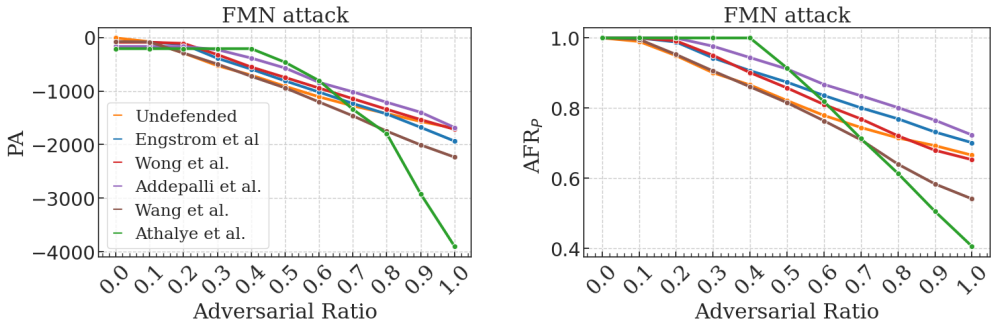



Figure B.8: PA and $\text{AFR}_\text{P}$ variation under increasing adversarial ratio. T he undefended net and several RobustBench robust models are considered against a 1000 step FMN attack.

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

$$\text{PA} \approx N\,\text{AFR}_\text{T}\,\text{AFR}_\text{P}\log\left(1-2\delta_\text{ERR}\right)+N(1-\text{AFR}_\text{T})\,\text{AFR}_\text{P}\log\left(1-2\delta_\text{MIS}\right)+N\,\text{AFR}_\text{T}(1-\text{AFR}_\text{P})\log\delta_\text{ADV},$$

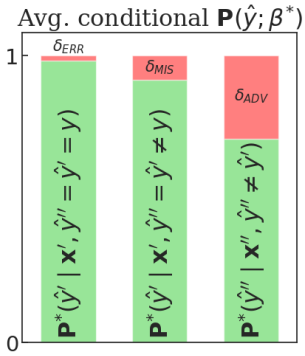



Figure B.9: Illustrative representation of the terms and posterior values constrained considered for the PA approximation.

*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:

$$\hat{y}_\text{first} = \arg\max_{\mathcal{Y}} \mathbf{P}(y \mid \boldsymbol{x}) = \hat{y}$$
$$\hat{y}_\text{next} = \arg\max_{\mathcal{Y}\setminus\{\hat{y}\}} \mathbf{P}(y \mid \boldsymbol{x})$$

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

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

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.
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 $\mathfrak{f} = \text{AFR}_{\text{T}}^0\,\text{AFR}_{\text{P}}$. Then

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

which yields

$$\zeta^i_{\text{ERR}} \approx \log\left\{(1-\delta_{\text{ERR}})^2 + \delta^2_{\text{ERR}}\right\} \approx \log\left(1 - 2\delta_{\text{ERR}}\right),$$

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 $\mathfrak{f} \approx (1 - \text{AFR}_{\text{T}}^0)\,\text{AFR}_{\text{P}}$. Then

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

which yields

$$\zeta^i_{\text{MIS}} \approx \log\left\{(1-\delta_{\text{MIS}})^2 + \delta^2_{\text{MIS}}\right\} \approx \log\left(1 - 2\delta_{\text{MIS}}\right),$$

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 $\mathfrak{f} = \text{AFR}_{\text{T}}^0(1 - \text{AFR}_{\text{P}})$. Then

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

which yields

$$\zeta^i_{\text{ADV}} \approx \log\left\{\left(1-\delta_{\text{ERR}}\right)\delta_{\text{ADV}} + \delta_{\text{ERR}}\left(1-\delta_{\text{ADV}}\right)\right\} \approx \log\delta_{\text{ADV}},$$

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. □

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.

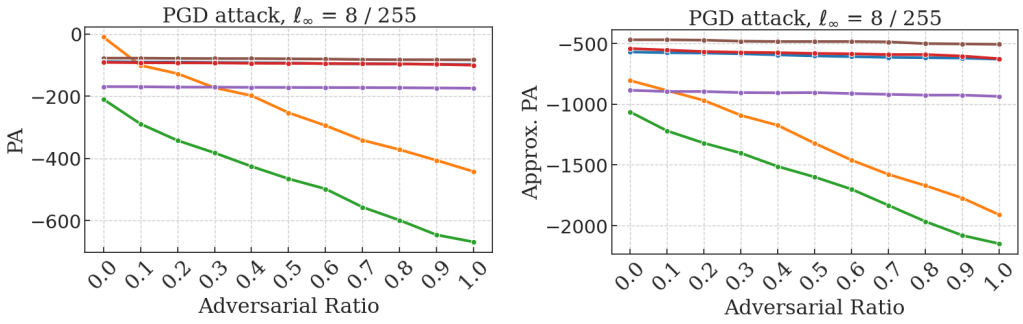



Figure B.10: True and approximated PA values under increasing adversarial ratio for a PGD attack with $\ell_\infty$=8/255.

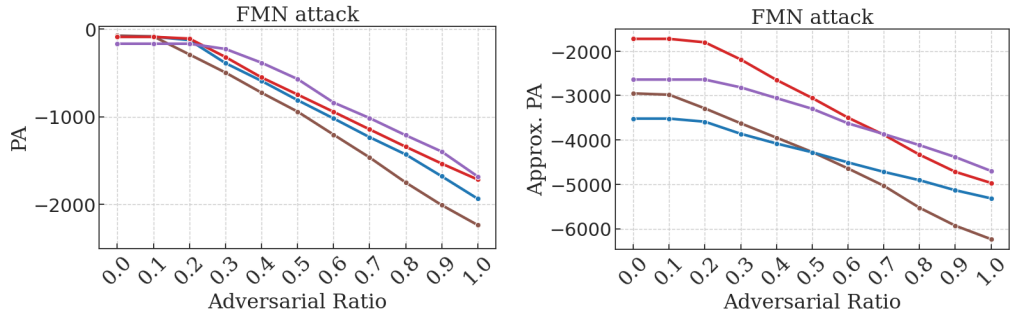



Figure B.11: True and approximated PA values under increasing adversarial ratio for a FMN attack.

Results obtained for 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.

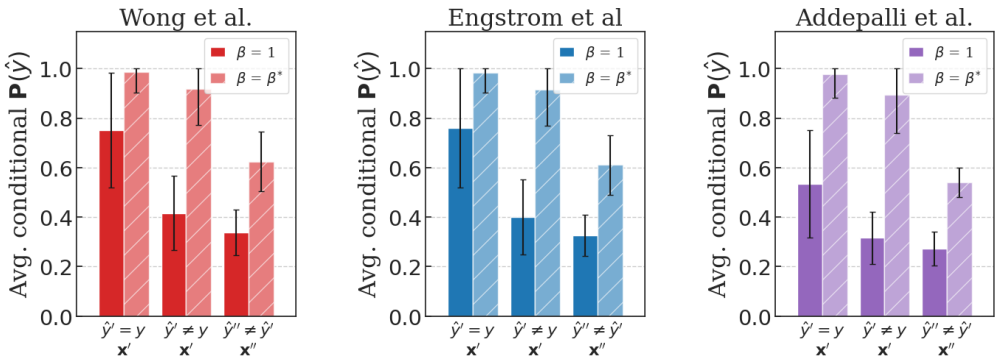



Figure B.12: Average posterior probability of the predicted class for correctly classified original samples, misclassified original samples, and misleading adversarial samples, respectively. Results have been obtained through a PGD attack with $\ell_\infty$=8/255 and sorted by increasing $\beta^*$.

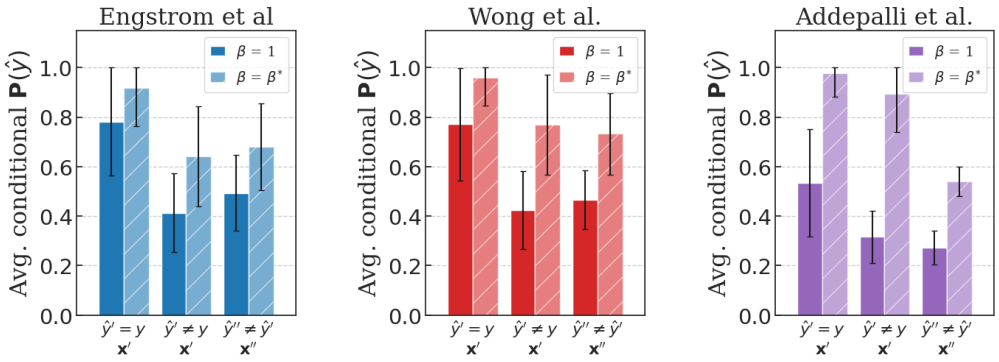



Figure B.13: Average posterior probability of the predicted class for correctly classified original samples, misclassified original samples, and misleading adversarial samples, respectively. Results have been obtained through a FMN attack and sorted by increasing $\beta^*$.

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.

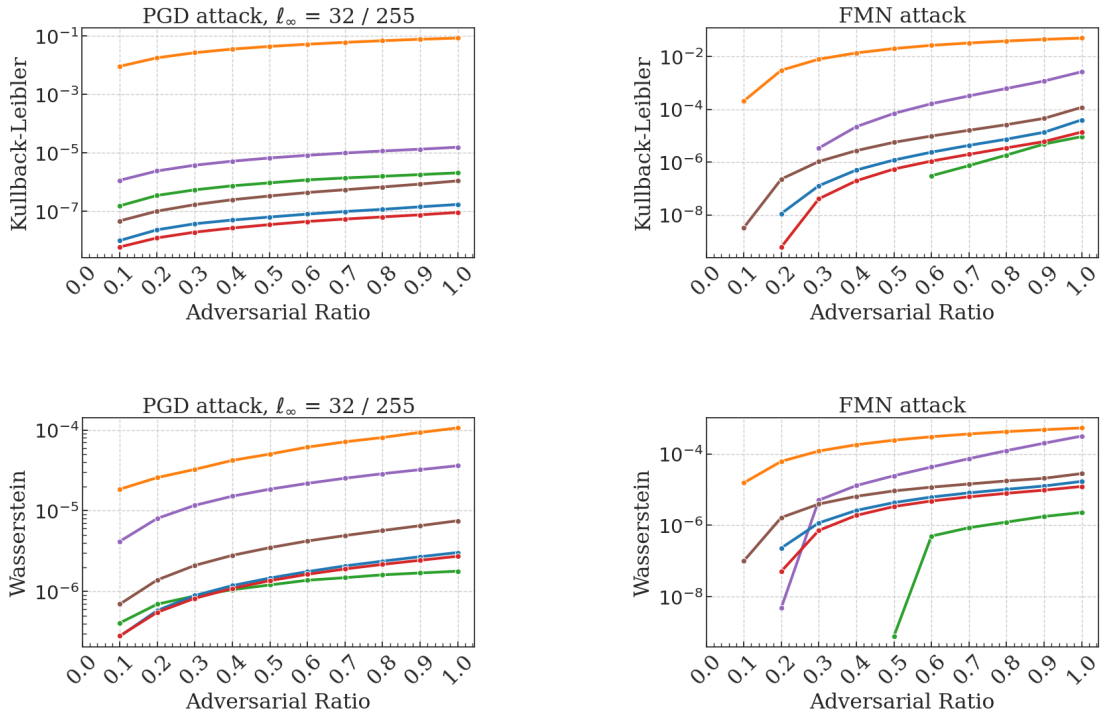



Figure B.14: Comparison of FMN and PGD attacks using probability-based distances.

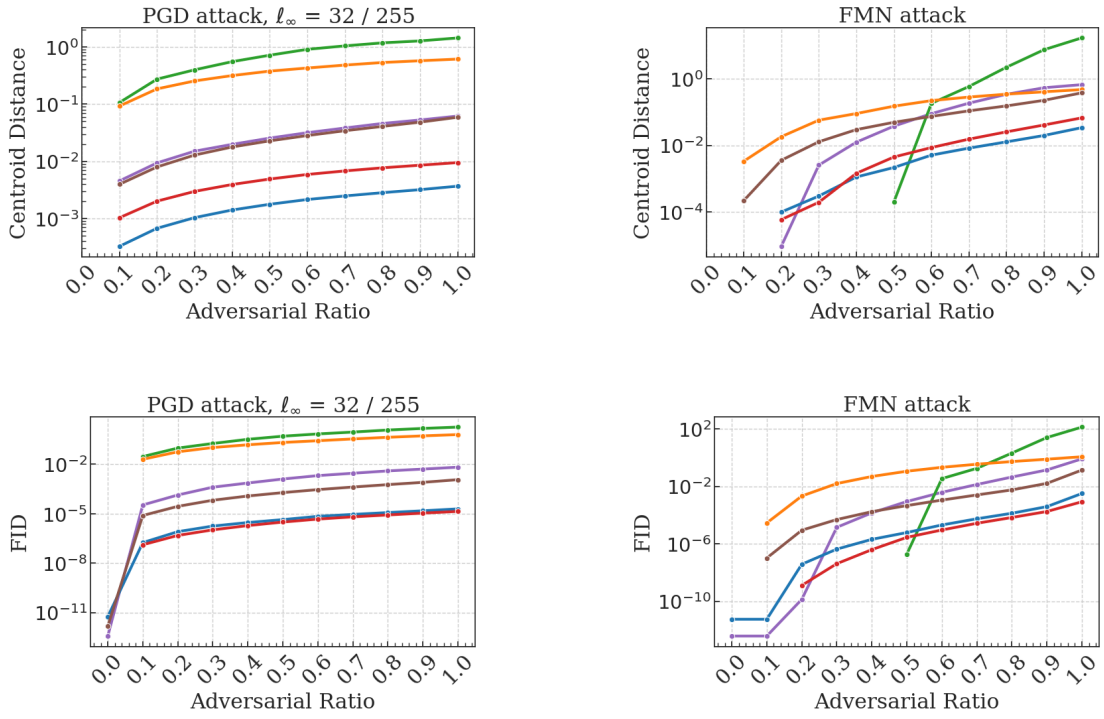



Figure B.15: Comparison of FMN and PGD attacks using feature-space-based distances.

## B.2 Model Selection

### B.2.1 DiagVib-6 Benchmark

| | Acc. Test 0 | | | Acc. Test 1 | | | Acc. Test 2 | | | Acc. Test 3 | | | Acc. Test 4 | | | Acc. Test 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ERM** | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA |
| SD | 99.3 | 99.4 | **99.5** | 54.8 | **71.5** | 64.0 | **81.0** | 54.3 | 79.9 | 74.5 | 48.6 | **76.1** | 53.4 | 34.9 | **57.9** | **35.8** | 32.7 | 35.2 |
| ID | 99.4 | **99.5** | **99.5** | **72.8** | 69.9 | 70.1 | 85.5 | 85.8 | **88.5** | **79.5** | 76.0 | 79.1 | 55.6 | **66.6** | 60.7 | 38.3 | **38.9** | 34.6 |
| 1F-MD | **99.5** | 99.4 | 99.4 | 48.2 | 45.1 | **51.0** | 74.6 | 77.6 | **79.8** | 65.7 | 68.9 | **75.3** | 59.7 | 58.2 | **63.3** | 42.8 | **44.8** | 38.3 |
| 5F-MD | 99.5 | 99.5 | 99.5 | 49.6 | 49.6 | 49.6 | 90.2 | 90.2 | 90.2 | 81.0 | 81.0 | 81.0 | 71.4 | 71.4 | 71.4 | 44.6 | 44.6 | 44.6 |
| OOD | 99.3 | **99.4** | **99.4** | 60.9 | **69.3** | **69.3** | **92.6** | 87.3 | 87.3 | **83.9** | 82.1 | 82.1 | **64.9** | 59.6 | 59.6 | 32.3 | **39.0** | **39.0** |
| **IRM** | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA |
| SD | 99.4 | 99.2 | **99.6** | **61.4** | 52.2 | 45.1 | 85.7 | 69.5 | **91.6** | 78.3 | 58.6 | **78.5** | 55.7 | **58.9** | 56.5 | 43.2 | 27.2 | **45.5** |
| ID | **99.6** | 99.4 | 99.4 | 44.8 | 31.2 | **51.2** | **91.7** | 88.7 | 79.0 | **78.6** | 73.8 | 67.6 | 56.5 | **65.5** | 57.2 | 45.4 | **50.2** | 47.8 |
| 1F-MD | 99.5 | 99.5 | 99.5 | 62.4 | 62.4 | 62.4 | 90.1 | 90.1 | 90.1 | 84.1 | 84.1 | 84.1 | 52.2 | 52.2 | **52.2** | 42.6 | 42.6 | 42.6 |
| 5F-MD | **99.4** | **99.4** | 99.3 | 31.2 | 31.2 | **83.3** | 88.7 | 88.7 | **91.9** | 73.8 | 73.8 | **85.9** | 65.5 | 65.5 | **66.5** | **50.2** | **50.2** | 46.7 |
| OOD | **99.5** | **99.5** | 99.2 | **62.4** | **62.4** | 48.8 | 90.1 | 90.1 | **92.3** | **84.1** | **84.1** | 80.6 | 52.2 | 52.2 | **60.0** | 42.6 | 42.6 | **51.1** |
| **LISA** | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA |
| SD | **99.4** | 97.9 | 99.3 | 87.7 | 78.5 | **88.0** | **63.8** | 45.8 | 62.6 | **79.3** | 39.8 | 74.9 | **74.8** | 40.1 | 71.1 | **42.3** | 26.2 | 38.7 |
| ID | **99.5** | 99.3 | 99.4 | 83.0 | **84.7** | 81.0 | 53.3 | 67.0 | **86.1** | 67.9 | **70.6** | 69.1 | 56.5 | **66.3** | 61.8 | 33.9 | 33.6 | **42.1** |
| 1F-MD | **99.5** | 99.4 | 99.3 | 88.8 | 71.0 | **91.7** | **54.4** | 50.4 | 42.3 | **76.8** | 66.3 | 61.9 | **71.5** | 62.4 | 56.2 | **36.4** | 32.8 | 28.9 |
| 5F-MD | 99.3 | 99.3 | 99.3 | 76.1 | 76.1 | 76.1 | 70.2 | 70.2 | 70.2 | 71.4 | 71.4 | 71.4 | 70.6 | 70.6 | 70.6 | 48.6 | 48.6 | 48.6 |
| OOD | **99.4** | 99.3 | 99.3 | **81.0** | 70.7 | 70.7 | **86.1** | 80.2 | 80.2 | 69.1 | **75.9** | **75.9** | 61.8 | **72.4** | **72.4** | **42.1** | 36.8 | 36.8 |

Table B.1: Test performance under increasing levels of shift for models selected through different configurations of validation datasets. PA, $AFR_P$ and accuracy are used as early stopping criteria for model selection in the hue factor experiment (see Tables 5.1-5.2).

| | Acc. Test 0 | | | Acc. Test 1 | | | Acc. Test 2 | | | Acc. Test 3 | | | Acc. Test 4 | | | Acc. Test 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ERM** | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA |
| SD | 99.3 | 99.2 | **99.4** | 99.4 | 99.3 | **99.5** | 86.6 | 82.1 | **87.4** | 68.5 | 59.7 | **74.5** | 50.8 | 47.6 | **56.4** | 47.4 | 46.4 | **57.6** |
| ID | 99.3 | **99.4** | **99.4** | 99.4 | 99.5 | **99.6** | **86.6** | 86.5 | 85.3 | 68.5 | **75.3** | 74.2 | 50.8 | **57.1** | 54.8 | 47.4 | **57.1** | 55.6 |
| 1F-MD | 99.3 | **99.4** | 99.3 | **99.5** | 99.4 | 99.4 | 84.9 | 83.6 | **89.0** | 66.3 | 66.3 | **68.0** | 52.0 | 47.0 | **53.6** | 46.3 | 41.4 | **53.6** |
| 5F-MD | 99.1 | 99.1 | 99.1 | 99.1 | 99.1 | 99.1 | 90.1 | 90.1 | 90.1 | 58.6 | 58.6 | 58.6 | 73.3 | 73.3 | 73.3 | 57.4 | 57.4 | 57.4 |
| OOD | 99.3 | 99.3 | 99.3 | 99.4 | 99.4 | 99.4 | **89.0** | 89.0 | 84.2 | **68.0** | **68.0** | 60.4 | **53.6** | **53.6** | 46.8 | **53.6** | **53.6** | 45.0 |
| **IRM** | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA |
| SD | **99.4** | 98.8 | 99.3 | **99.4** | 98.7 | 99.2 | **85.0** | 65.3 | 83.2 | 55.0 | **58.4** | 51.2 | 59.0 | **68.9** | 52.1 | 51.4 | **58.8** | 45.2 |
| ID | **99.4** | 99.3 | 99.2 | 99.4 | 99.3 | **99.5** | **86.4** | 81.1 | 83.0 | 50.4 | 46.2 | **62.4** | 48.2 | 54.2 | 70.4 | 39.8 | 45.5 | **59.6** |
| 1F-MD | **99.4** | 99.3 | 99.3 | **99.5** | 99.4 | 99.3 | 85.3 | 87.1 | **89.5** | 30.9 | **56.8** | 44.0 | 49.6 | 49.7 | **68.8** | 38.2 | 44.3 | **54.8** |
| 5F-MD | 99.2 | 99.2 | 99.2 | 99.5 | 99.5 | 99.5 | 83.0 | 83.0 | 83.0 | 62.4 | 62.4 | 62.4 | 70.4 | 70.4 | 70.4 | 59.6 | 59.6 | 59.6 |
| OOD | 99.4 | 99.3 | **99.5** | 99.4 | 99.4 | 99.4 | **85.0** | 82.7 | 84.4 | **55.0** | 39.9 | 51.6 | **59.0** | 44.4 | 58.7 | **51.4** | 37.2 | 49.9 |
| **LISA** | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA |
| SD | **99.4** | 98.6 | **99.4** | 99.2 | 98.6 | **99.3** | **90.4** | 55.8 | 78.8 | **74.9** | 37.2 | 61.9 | **71.9** | 28.5 | 60.2 | 54.4 | 29.4 | **55.5** |
| ID | **99.4** | 99.2 | 99.3 | **99.3** | 99.0 | 99.1 | **85.5** | 62.3 | 78.4 | 49.7 | 74.5 | **89.9** | 62.8 | 46.6 | **63.5** | 54.2 | 48.6 | **59.1** |
| 1F-MD | **99.4** | **99.4** | 99.3 | **99.3** | **99.3** | 99.1 | **85.5** | **85.5** | 78.4 | 49.7 | 49.7 | **89.9** | 62.8 | 62.8 | **63.5** | 54.2 | 54.2 | **59.1** |
| 5F-MD | 98.8 | 98.8 | 98.8 | 99.0 | 99.0 | 99.0 | 70.4 | 70.4 | 70.4 | 66.3 | 66.3 | 66.3 | 63.9 | 63.9 | 63.9 | 58.5 | 58.5 | 58.5 |
| OOD | **99.4** | 99.3 | **99.4** | **99.6** | 99.2 | **99.6** | 78.0 | **78.4** | 78.0 | **80.3** | 77.4 | **80.3** | 60.5 | **68.2** | 60.5 | 55.9 | **56.7** | 55.9 |

Table B.2: Test performance under increasing levels of shift for models selected through different configurations of validation datasets. PA, $AFR_P$ and accuracy are used as early stopping criteria for model selection in the position factor experiment (see Tables C.5-C.3).

### B.2.2 In-distribution DiagVib-6

| | Acc. Test 1 | | | Acc. Test 2 | | | Acc. Test 3 | | | Acc. Test 4 | | | Acc. Test 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ERM** | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA |
| ZGO | 53.2 | 53.2 | 53.2 | 54.6 | 54.6 | 54.6 | 55.7 | 55.7 | 55.7 | 66.7 | 66.7 | 66.7 | 66.6 | 66.6 | 66.6 |
| 1-CGO | 62.9 | **72.4** | **72.4** | 64.7 | **74.8** | **74.8** | 60.8 | **61.1** | **61.1** | 62.9 | **65.1** | **65.1** | 64.2 | **64.7** | **64.7** |
| 2-CGO | 69.1 | **78.6** | **78.6** | 71.2 | **78.9** | **78.9** | 71.9 | **72.2** | **72.2** | **76.2** | 73.8 | 73.8 | **77.0** | 74.2 | 74.2 |
| 3-CGO | 73.1 | **89.8** | **89.8** | 85.6 | **89.3** | **89.3** | 70.1 | **79.8** | **79.8** | 71.4 | **77.7** | **77.7** | 72.1 | **78.8** | **78.8** |
| ZSO | 99.6 | 99.6 | 99.6 | **92.8** | 92.7 | 92.7 | 89.9 | **90.0** | **90.0** | 85.9 | **86.0** | **86.0** | 85.9 | **86.0** | **86.0** |
| **IRM** | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA |
| ZGO | 50.1 | **56.0** | **56.0** | 50.5 | **55.4** | **55.4** | 52.8 | **62.3** | **62.3** | 64.4 | **65.5** | **65.5** | 64.9 | **66.0** | **66.0** |
| 1-CGO | 63.0 | 64.0 | **70.1** | 65.9 | 64.3 | **73.4** | 59.4 | 59.5 | **61.6** | 60.1 | 56.0 | **62.4** | 59.0 | 56.9 | **60.7** |
| 2-CGO | 69.0 | **82.4** | 79.5 | 69.7 | **83.2** | 79.6 | 67.5 | **78.5** | 72.2 | 64.5 | 77.0 | **77.5** | 65.1 | **79.0** | 77.7 |
| 3-CGO | 79.5 | **91.1** | **91.1** | 83.0 | **92.8** | **92.8** | 73.6 | **84.5** | **84.5** | 70.7 | **81.7** | **81.7** | 72.2 | **83.5** | **83.5** |
| ZSO | 99.4 | **99.5** | **99.5** | 93.4 | **95.1** | 94.7 | 89.2 | 79.8 | **89.4** | 87.0 | 80.4 | **88.6** | 87.0 | 80.4 | **88.6** |

Table B.3: Test performance under increasing levels of shift for models selected through different configurations of factor co-occurrence. PA, $AFR_P$ and accuracy are used as early stopping criteria for model selection in the hue factor experiment (see Tables C.6 and C.2.2-C.2.2).

| | Acc. Test 1 | | | Acc. Test 2 | | | Acc. Test 3 | | | Acc. Test 4 | | | Acc. Test 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ERM** | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA |
| ZGO | 50.2 | 50.2 | **50.2** | 50.0 | **50.0** | **50.0** | 52.4 | 52.4 | 52.4 | 52.0 | 52.0 | 52.0 | 50.6 | 50.6 | 50.6 |
| 1-CGO | 44.1 | 48.2 | **48.2** | 41.4 | **42.0** | **42.0** | 43.6 | **45.0** | **45.0** | 45.0 | **46.0** | **46.0** | 44.8 | **46.0** | **46.0** |
| 2-CGO | 64.4 | **71.8** | 71.4 | 53.4 | 52.5 | **55.1** | 55.1 | 55.9 | **57.8** | 57.7 | **60.5** | 59.0 | 56.7 | **59.0** | 58.0 |
| 3-CGO | 73.2 | **91.1** | **91.1** | 55.3 | **63.6** | **63.6** | 52.8 | **61.3** | **61.3** | 51.8 | **58.8** | **58.8** | 51.8 | **59.1** | **59.1** |
| ZSO | **99.1** | 98.4 | **99.1** | **92.1** | 91.6 | 91.9 | 88.6 | 86.2 | **88.8** | 87.0 | 84.3 | **87.1** | 85.7 | 84.6 | **85.9** |
| **IRM** | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA | Acc. | $AFR_P$ | PA |
| ZGO | 50.5 | 50.5 | 50.5 | 50.0 | 50.0 | 50.0 | 53.1 | 53.1 | 53.1 | 52.5 | 52.5 | 52.5 | 51.6 | 51.6 | 51.6 |
| 1-CGO | 48.2 | **49.0** | **49.0** | 43.6 | **44.2** | **44.2** | **43.2** | 42.8 | 42.8 | **44.0** | 42.4 | 42.2 | **44.1** | 42.1 | 42.0 |
| 2-CGO | 65.0 | **72.2** | 72.0 | 53.1 | 57.6 | **58.6** | 52.8 | 54.0 | **54.1** | **59.2** | 58.6 | 55.6 | **60.1** | 58.7 | 55.4 |
| 3-CGO | 80.4 | **94.1** | **94.1** | 63.2 | **70.9** | **70.9** | 57.5 | **63.5** | **63.5** | 51.4 | **58.9** | **58.9** | 52.6 | **57.8** | **57.8** |
| ZSO | **99.6** | 99.0 | 99.3 | **93.8** | 93.3 | 92.8 | 88.2 | **90.0** | 85.8 | **89.5** | 83.0 | 81.3 | **89.6** | 86.2 | 81.2 |

Table B.4: Test performance under increasing levels of shift for models selected through different configurations of factor co-occurrence. PA, $AFR_P$ and accuracy are used as early stopping criteria for model selection in the position factor experiment (see Tables C.6 and C.2.2-C.2.2).

# 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 $\boldsymbol{x}'$ and $\boldsymbol{x}''$ are exclusively subject to sampling randomness, in the sense that each observation in $\boldsymbol{x}'$ is paired with an observation in $\boldsymbol{x}''$ 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 $\boldsymbol{x}'$ and $\boldsymbol{x}''$, 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:

$$\mathcal{S} = \{X_0, X_1\},$$
$$\mathcal{T} = \{X_1, X_2, X_3, X_4, X_5\},$$

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:

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

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.

$$D^{\text{train}} = \{\boldsymbol{x}_0^{\text{train}}, \boldsymbol{x}_1^{\text{train}}\}, \text{ where } \boldsymbol{x}_j := \boldsymbol{x}_j \circ \tau_j^{\text{train}}, \ j = 0, 1$$
$$D^{\text{val}} = \{\boldsymbol{x}_0^{\text{val}}, \boldsymbol{x}_1^{\text{val}}\}, \text{ where } \boldsymbol{x}_j := \boldsymbol{x}_j \circ \tau_j^{\text{val}}, \ j = 0, 1$$

  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.

$$D^{\text{train}} = \{\boldsymbol{x}_0^{\text{train}}, \boldsymbol{x}_1^{\text{train}}\}, \text{ where } \boldsymbol{x}_j := \boldsymbol{x}_j \circ \tau_j^{\text{train}}, \ j = 0, 1$$
$$D^{\text{val}} = \{\boldsymbol{x}_0^{\text{val}}, \boldsymbol{x}_1^{\text{val}}\}, \text{ where } \boldsymbol{x}_j := \boldsymbol{x}_j \circ \tau^{\text{val}}, \ j = 0, 1$$

  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.

| # Shift Factors | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Hue | red | **blue** | blue | blue | blue | blue |
| Lightness | dark | dark | **bright** | bright | bright | bright |
| Position | CC | CC | CC | **LC** | LC | LC |
| Scale | normal | normal | normal | normal | **large** | large |
| Texture | blank | blank | blank | blank | blank | **tiles** |
| *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 DiagVib-6 Benchmark

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.

| # Shift Factors | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Hue | red | **green** | green | green | green | green |
| Lightness | dark | dark | **bright** | bright | bright | bright |
| Position | CC | CC | CC | CC | **UL** | UL |
| Scale | large | large | large | large | large | **small** |
| Texture | blank | blank | blank | **tiles** | 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.

| # Shift Factors | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Hue | green | green | green | **red** | green | green |
| Lightness | dark | dark | dark | dark | **bright** | bright |
| Position | UL | **CC** | CC | CC | CC | CC |
| Scale | large | large | **small** | small | small | small |
| Texture | blank | blank | blank | blank | blank | **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.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 | *Shape* |
|---|---|---|---|---|---|---|---|
| Training | 0 | red | dark | CC | large | blank | *1,4,7,9* |
| | 1 | **blue** | dark | CC | large | blank | *1,4,7,9* |
| Validation | 0 | red | dark | CC | large | blank | *1,4,7,9* |
| SD | 1 | red | dark | CC | large | blank | *1,4,7,9* |
| ID | 1 | **blue** | dark | CC | large | blank | *1,4,7,9* |
| 1F-MD | 1 | **magenta** | dark | CC | large | blank | *1,4,7,9* |
| 5F-MD | 1 | **green** | **bright** | **UL** | **small** | **tiles** | *1,4,7,9* |
| Validation OOD | 0 | **yellow** | dark | CC | large | blank | *1,4,7,9* |
| | 1 | **magenta** | dark | CC | large | blank | *1,4,7,9* |

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.

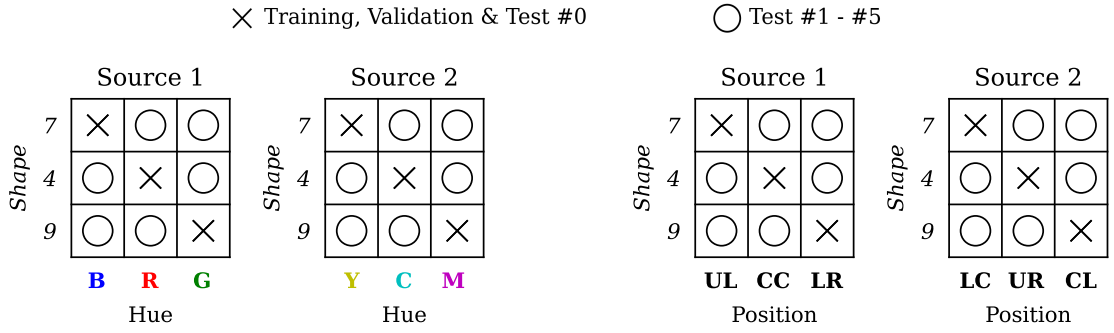
| | Env. | Hue | Lightness | Position | Scale | Texture | *Shape* |
|---|---|---|---|---|---|---|---|
| Training | 0 | green | dark | UL | large | blank | *1,4,7,9* |
| | 0 | green | dark | LR | large | blank | *1,4,7,9* |
| Validation | 0 | green | dark | UL | large | blank | *1,4,7,9* |
| SD | 1 |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

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.

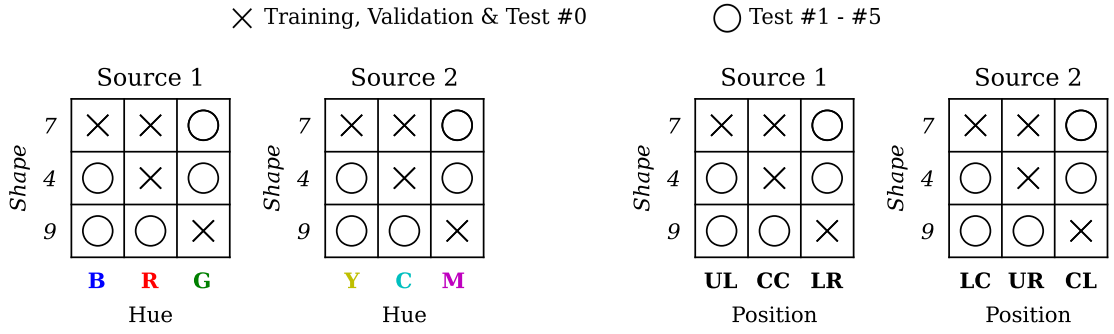
### C.2.2 In-distribution DiagVib-6

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.

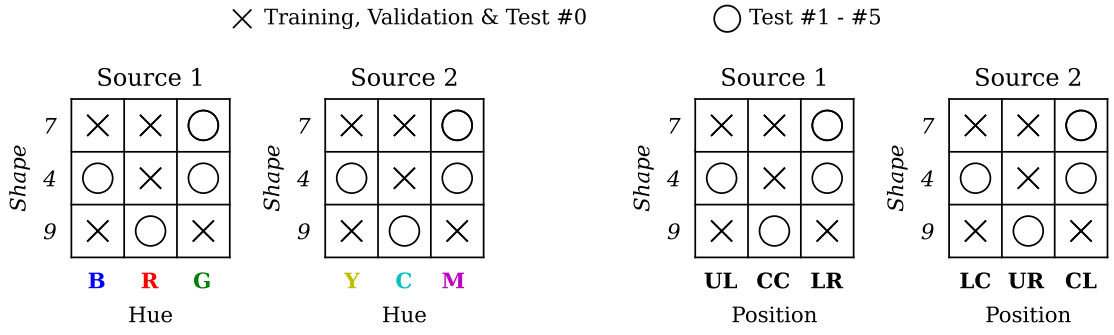
**Dataset** (ID setting under ZGO)**.** Let the predicted factor $F^P$ be `shape` (i.e. the image digits), and let the learned factor $F^L$ be `hue` and `position`, each for a different dataset. The zero shortcut opportunity datasets are obtained as follows:



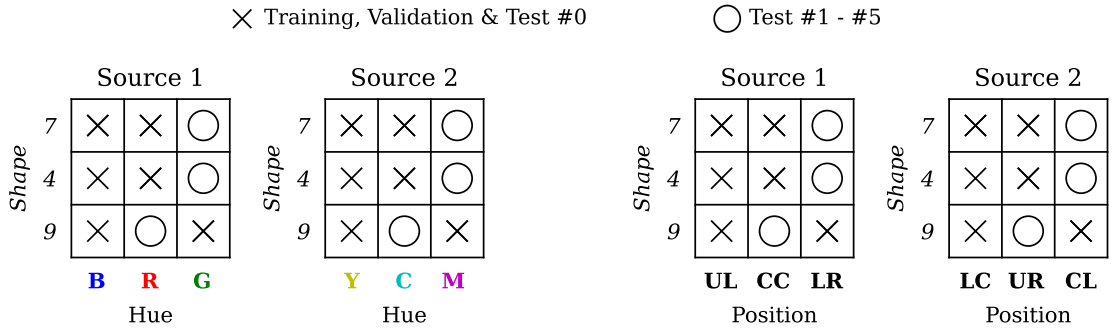

**Dataset** (ID setting under 1-CGO)**.** Let the predicted factor $F^P$ be `shape` (i.e. the image digits), and let the learned factor $F^L$ be `hue` and `position`, each for a different dataset. The single compositional generalization opportunity datasets are obtained as follows:



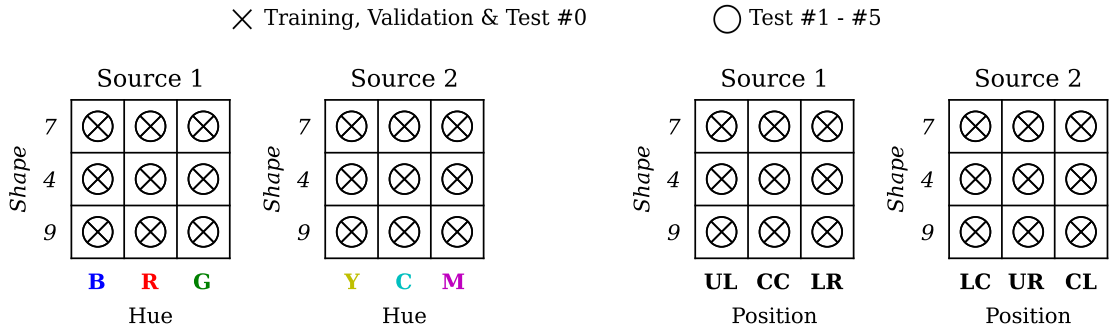

**Dataset** (ID setting under 2-CGO)**.** Let the predicted factor $F^P$ be `shape` (i.e. the image digits), and let the learned factor $F^L$ be `hue` and `position`, each for a different dataset. The double compositional generalization opportunity datasets are obtained as follows:



**Dataset** (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:




**Dataset** (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 | **CR** | CR | CR | CR |
| Scale | large | large | large | **small** | small | small |
| Brightness | bright | bright | bright | bright | **dark** | dark |
| Texture | tiles | tiles | tiles | tiles | tiles | **wood** |

| `position` | #0 | #1 | #2 | #3 | #4 | #5 |
|---|---|---|---|---|---|---|
| Scale | large | large | **small** | small | small | small |
| Hue | blue | blue | blue | **red** | red | red |
| Brightness | bright | bright | bright | bright | **dark** | dark |
| Texture | tiles | tiles | tiles | tiles | tiles | **wood** |

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 Mummadi, 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] Facebook AI Research. Faiss: A library for efficient similarity search and clustering of dense vectors. `https://github.com/facebookresearch/faiss`, 2017. Accessed: 2024-09-15.

[22] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples.

[23] Peter Grünwald and Teemu Roos. Minimum Description Length Revisited. 11(01):1930001.

[24] 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.

[25] Allan Gut. *An Intermediate Course on Probability*. Springer, second edition.

[26] 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.

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models.

[28] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial Examples Are Not Bugs, They Are Features.

[29] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, first edition.

[30] Ortiz Jimenez. The inductive bias of deep learning: Connecting weights and functions.

[31] Ian T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, New York, 2nd edition, 2002.

[32] 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.

[33] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization.

[34] 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.

[35] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images.

[36] 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.

[37] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to Generalize: Meta-Learning for Domain Generalization. 32(1).

[38] 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.

[39] Jian Liang, Ran He, and Tieniu Tan. A Comprehensive Survey on Test-Time Adaptation under Distribution Shifts.

[40] Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards Out-Of-Distribution Generalization: A Survey.

[41] Cristopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer.

[42] 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.

[43] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks.

[44] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning.

[45] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain Generalization via Invariant Feature Representation.

[46] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Second edition.

[47] Kevin P. Murphy. *Probabilistic Machine Learning. An Introduction.* The MIT Press.

[48] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-Adversarial Domain Adaptation.

[49] Maura Pintor, Fabio Roli, Wieland Brendel, and Battista Biggio. Fast Minimum-norm Adversarial Attacks through Adaptive Norm Constraints.

[50] Joaquin Quiñonero-Candela, editor. *Dataset Shift in Machine Learning*. Neural Information Processing Series. MIT Press.

[51] Sebastian Ruder. An overview of gradient descent optimization algorithms.

[52] David E Rumelhart, Geoffrey E Hintont, and Ronald J Williams. Learning representations by back-propagating errors.

[53] 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.

[54] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Mądry. Adversarially Robust Generalization Requires More Data.

[55] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein Distance Guided Representation Learning for Domain Adaptation.

[56] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition.

[57] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks.

[58] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE.

[59] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness May Be at Odds with Accuracy.

[60] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep Learning for Computer Vision: A Brief Review. 2018:1–13.

[61] Chengyu Wang, Minghui Qiu, Jun Huang, and Xiaofeng He. Meta Fine-Tuning Neural Language Models for Multi-Domain Text Mining.

[62] 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.

[63] 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.

[64] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better Diffusion Models Further Improve Adversarial Training.

[65] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training, 2023.

[66] 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.

[67] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020.

[68] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating Adversarial Examples with Adversarial Networks.

[69] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving Out-of-Distribution Robustness via Selective Augmentation.

[70] Samuel Yu, Peter Wu, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. PACS: A Dataset for Physical Audiovisual CommonSense Reasoning.

[71] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features.

[72] 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.

[73] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically Principled Trade-off between Robustness and Accuracy.

[74] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond Empirical Risk Minimization.

[75] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain Generalization: A Survey. pages 1–20.

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Institute for Machine Learning
Prof. Dr. Joachim M. Buhmann

**Title of work:**

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

**Thesis type and date:**

Master Thesis, September 2024

**Supervision:**

Prof. Dr. Joachim M. Buhmann, Dr. João Carvalho, Dr. Alessandro Torcinovich

**Student:**

Name: Victor Jimenez Rodriguez
E-mail: vjimenez@student.ethz.ch
Legi-Nr.: 900-0030-00L

**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`

Zurich, 16. 9. 2024: _____