ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

DINFK
Institute for Machine Learning

Victor Jimenez Rodriguez

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. Joao Borges de Sa Carvalho, Dr. Alessandro Torcinovich
Prof. Dr. Joachim M. Buhmann

September 2024

IDSC-XX-YY-ZZ

Preface

The research presented in this thesis was conducted within the Information and Science Engineering Group at the Institute for Machine Learning (ETH Zurich), during the period October 2023 - September 2024, under the supervision of Prof. Dr. Joachim M. Buhmann. The thesis was co-supervised at Universitat Politecnica de Catalunya by Prof. Dr. Alexandre Parera i Lluna.

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 this hypothesis and we explore the use of PA as a model selection criterion for deep learning models in supervised classification tasks.

Starting from the theoretical principles leading to PA, we derive a computationally-efficient approximation to its value in discrete hypothesis set problems, and we show that it is a valid alternative to cross-validation in the presence of distribution shifts. Additionally, we follow some threads from the original hypothesis and we extend the use of PA to some other specific settings in which a domain shift can be defined, such as subpopulation (unbalanced) settings, mislabelled datasets and data augmentation strategies.

REWRITE AT THE END

Notation

Symbols

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

Indicies

a	Ambient
air	Air

Acronyms and Abbreviations

NEDC	New European Driving Cycle
ETH	Eidgenössische Technische Hochschule

Chapter 1

Introduction

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

1.1 The robustness challenge

PARAGRAPH MISSING: Context on machine learning, deep learning and classification tasks.

Robustness can be defined as the ability of a machine learning model to maintain its predictive power on unseen observations that present some kind of transformation or variation [39]. Overall, three sources of variability are relevant in the context of image classification, namely sampling randomness, adversarial attacks, and out-of-distribution generalization (see Figure 1.1) [9].

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 [45]. 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 [47].

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. Lack of representativity of certain aspects of the data and the presence of spurious correlations can lead to models that generalize well to sampling randomness within the same dataset 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 [39].

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 [24]. 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 of standard performance in robust models [47, 58]. 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 naturally invariant to [23].

(a) Original (b) Sampling (c) Adversarial (d) Domain

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

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

1.1.1 Adversarial setting

As it was already mentioned, certain perturbations to 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 [45].

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 their non-linear nature. 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 significally change the outcome [18].

Following this intuition, several attacks have been proposed to evaluate the robustness of models to adversarial samples. One common principle to induce model failure is finding vulnerable directions in the feature space and adjusting the perturbation to have the desired misleading effect. Adversarial examples generated by these attacks can be used to train robust models via regularization, promoting generalization to those features present in the worst-case examples and thus selecting models insensitivized to them.

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 [42]. These limitations express a fundamental trade-off that arises from an intrinsic difference between robust and non-robust features [47, 58].

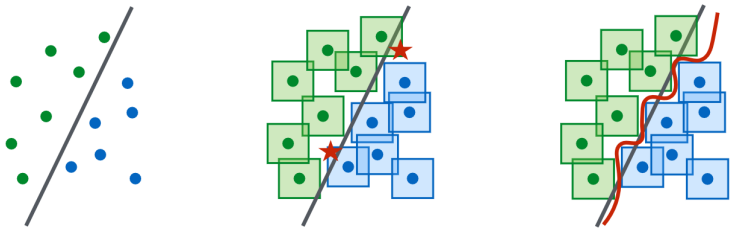
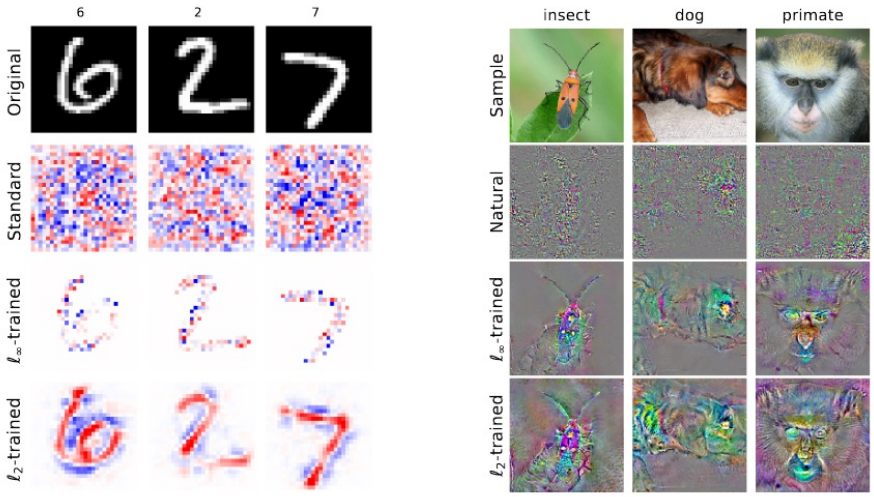


Figure 1.2: A conceptual illustration of standard vs. adversarial decision boundaries. (**left**) A set of linearly-separable points. (**middle**) Decision boundary learned via standard training. (**right**) Decision boundary learned via adversarial training. Both methods achieve zero training error, but only the robust model is able to generalize to ℓ_∞ perturbations. Source: [33]

Features selected by 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 (see Figure 1.3) [23].



(a) MNIST (b) Restricted ImageNet

Figure 1.3: Scaled loss gradient with respect to input images. Input pixels yielding the most predictive power are aligned with perceptually relavant features for the case of adversarial models, while appearing completely random in the case of standard models. Source: [47]

Furthermore, adversarial perturbations of robust models have been shown to display salient characteristics; that is, their features are perceived to belong to the class they are misclassified to, as illutrated in Figure 1.4.

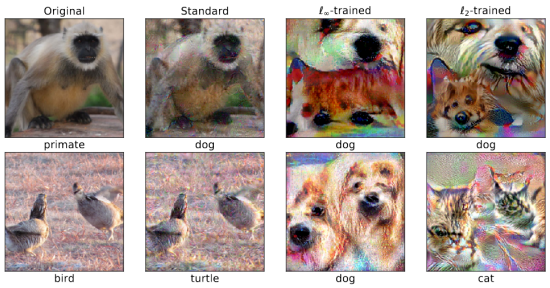


Figure 1.4: Adversarial examples for standard and PGD-trained models. Perturbed images produced for robust models effectively capture salient data characteristics and appear similar to examples of a different class. Source: [47]

Overall, these and other findings suggest that robustness in the adversarial setting is a fundamental property of the 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 providing 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 [35, 39]. 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 [60, 50, 32].

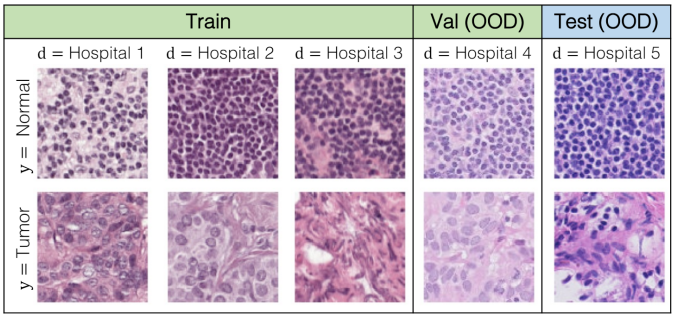


Figure 1.5: The `camelyon17` (WILDS) dataset comprises tissue patches from different hospitals. The goal is to accurately predict the presence of tumor tissue in patches taken from hospitals that are not in the training set. Source: [27]

A distribution shift can arise for various reasons, namely the unfeasibility of collecting diverse enough data, the lack of representativity of certain features, the changing or time-dependent nature of the data and also the implicit bias induced in the data collection process. This last point 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.

Several approaches can be taken to address this issue, depending on the nature of the shift and the access to its causal structure (see Figure 3 and Table 2 in [50]). 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 [50, 32].

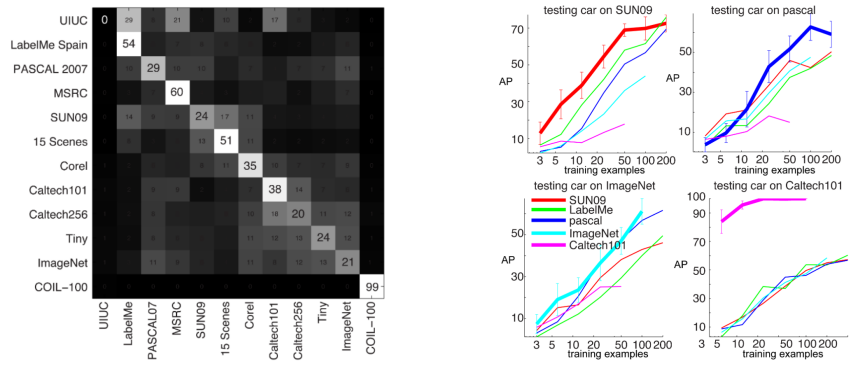


Figure 1.6: **(left)** Confusion matrix associated with a dataset identification task. There is a clearly pronounced diagonal, which indicates that each dataset posesses unique traits that make it distinguishable from the rest. **(right)** Cross-dataset generalization for "car" detection as function of training data. The vertical gap between two curves represents the decrease in performance resulting from training on a different dataset, and horizontal shift corresponds to the increase in amount of data needed to reach the same level of performance. Source: [46]

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 other 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 known as domain adaptation.

In this work we will focus instead on domain generalization, which refers to the case in which sampling from target domains is not feasible and feature invariance can be only enforced from the source [4]. 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. kernel latent space [35], adversarial [37] or model-based [1]) and the similarity metric will determine the the particularities of the method [43, 31].

[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 feature space.

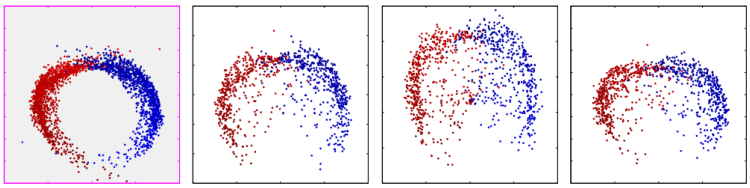


Figure 1.7: Projections of a binary synthetic dataset in the two principal DICA dimensions. The shaded box depicts the projection of training data, whereas the unshaded boxes show projections of unseen test datasets. Source: [35]

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

(a) Original (b) Mixup [59] (c) CutMix [57]

Figure 1.8: Mixup and Cutmix strategies can be used to interpolate between different labels and/or domains by generating intermediate observations. Source: [57]

Unlike in the adversarial setting, there is no common way of measuring the shift in distribution betweeen 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 by 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 [60]. 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 [27].

1.2 Related work

In the adversarial front, early work [45] unveiled the nature of the susceptibility of deep learning models to adversarial examples and FGSM [18] was introduced as an intuitive approach for model regularization. Since then, several gradient-based methods have been proven to enhance adversarial robustness, such as PGD [33], C&W [11], FMN [38] and many others (see [30] 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 [3].

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 [30]. Regarding adversarial training (i.e. defenses), robustness can be achieved by a variety of methods, such as ensemble learning, defensive distillation, generative adversarial networks [54, 34], diffusion models [52, 22] and adaptive-boundary methods [14]. In this project, the Robustbench attack library [15] will be used for adversarial robustness evaluation in the `CIFAR10` dataset [28].

In the domain generalization front, the existing rich taxonomy of methods can be classified into three main groups, namely data manipulation, representation learning and alternative learning strategies [50, 60, 32]. Data manipulation strategies refer to augmentation and generation, as for example randomization or adversarial augmentation [55, 59, 57]. Representation learning strategies are primarily divided into domain-invariant methods (e.g. IRM [1] or kernel-based [35, 2]) and feature disentanglement methods, which encompass causality-inspired approaches and general multi-component analysis. Other learning strategies include meta-learning [29, 49], ensemble learning or self-supervised learning.

Regarding robustness characterization, a wide range of metrics have been conceived (see [20] for reference), but accuracy-based criteria are still the most common. Alternatively, some theoretically-grounded approaches have been proposed, such as CLEVER [53], ACTS [51] or PA [9], which is the one we will explore 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 (and its multiple variations, such as DiagVib-6 [17]), PACS [56], VLCS [25] or WILDS [27].

1.3 Objectives

WRITE AT THE END

The main objective of this thesis is thus to assess the suitability of this framework in the context of deep learning model robustness in image classification tasks. For that, an operative version of posterior agreement will be derived, and an efficient implementation of its computation will be used as a metric to evaluate and select models based on the robustness of their response to different sources and levels of variability. The results of this work will be compared with the current state-of-the-art in robustness evaluation, namely robustbench [**?**] and WILDS [27] benckmarks in the adversarial and out-of-distribution settings, respectively, and an overall analysis of the use of the metric as an early-stopping criterion will be provided.

- Lead to the derivative work that will be presented in the thesis.=> benchmarking
- Lead to derivative but next level (more useful, current interest...) work => model selection
- Lead to non-derivative (i.e. probably unsuccessful) work => model selection beyond robustness
- Outline the structure of the thesis in terms of hypothesis.

Chapter 2

Theoretical background

2.1 The learning framework

Statistical learning theory encompasses the mathematical framework used to study generalization in machine learning [36]. 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 measure of probability in $\mathcal{X}$, and let $\underline{\mathrm{X}} = (X_1, ..., X_N) \overset{\text{iid}}{\sim} X$ be a N-sized (simple) random sample of X [12]. A supervised dataset D is a realization $\boldsymbol{x} \sim \underline{\mathrm{X}}$ paired with its output values under the target function mapping.

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

$\mathcal{D}$ will represent the class of supervised datasets generated from $\underline{\mathrm{X}}$.

The quality of the approximation can be measured with the expected risk $\mathcal{R}(f)$:

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

where $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ denotes a loss function. The Glivenko-Cantelli theorem allows us to estimate the expected risk with its empirical (plug-in) analogous when N is large enough [21].

Definition (Empirical risk)**.** Let D and $\mathcal{L}$ be the dataset and loss function of our problem, respectively. The empirical risk of $f \in \mathcal{F}$ computed on D is defined as

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

Training, therefore, amounts to minimizing the empirical risk over the function class $\mathcal{F}$:

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

The best learning algorithm will be that achieving the lowest generalization error under sampling randomness, which is estimated as the empirical risk on a different dataset $D' \in \mathcal{D}$.

It can be shown that generalization error is ultimately linked to the complexity of the function class. The definition of complexity depends on the nature of the problem, but intuitively measures the cardinality of the subset of $\mathcal{F}$ that the algorithm is able to represent. A complex or high-capacity algorithm will be able to represent a larger subset of $\mathcal{F}$ and achieve a low empirical error, but will be also prone to overfitting to the specific learning realization thus yielding a higher generalization error [36].

As a general principle, the inductive bias of the algorithm (i.e. the set of constraints imposed on $\mathcal{F}$ during learning) should be aligned with that of our target function [24]. Given that more expressive classes are always preferred by optimization algorithms, the ERM objective function is tweaqued to include a regularization term penalizing complexity.

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

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

where $\lambda \in \mathbb{R}$ controls 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 [44, 36, 48].

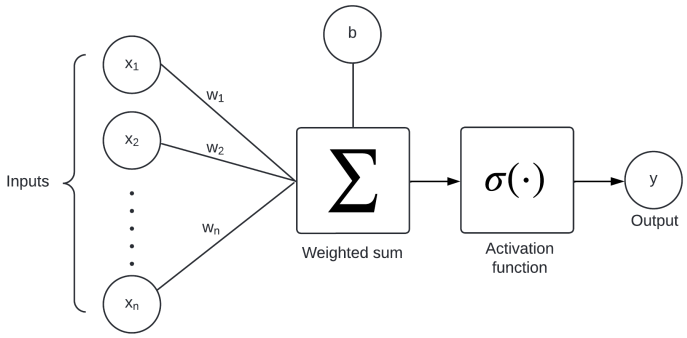


Figure 2.1: The output of a node is computed by applying a non-linear activation function σ to the weighted sum of its inputs $\boldsymbol{x}$ plus a bias term b.

Let $\boldsymbol{x}_k \in \mathbb{R}^{d_k}$ be the input to the layer $k \leq L$, 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 σ_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:

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

where $\gamma \in \Gamma \subset \mathbb{R}^S$ represents the set of parameters of the network. 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 the backpropagation algorithm [41]. In practice, more efficient variations of gradient descent are used, such as stochastic gradient descent [40] or the Adam optimizer [26].

A neural network architecture NN can be expressed as a parametrization of the function space $\mathcal{F}$:

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

where Γ is the parameter space associated with this particular architecture. The functional landscape $\mathcal{F}_\Gamma$ consists of all mappings $f(\gamma) : \mathcal{X} \longmapsto \mathcal{Y}$ that that can be realized by some parameter configuration $\gamma \in \Gamma$.

The universal approximation theorem states that an arbitrarily wide architecture is 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 linked to better generalization capabilities, as they intuitively represent a robust region in the parametrization space and should be preferred over sharp minima [24].

In this work we will explore a different approach to the generalization problem, based on a definition of the generalization error that relies on the implicit randomness of the data sampling process.

2.3 Posterior agreement

As mentioned at the start of the chapter, the input of learning algorithms are datasets containing samples 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 [7]. 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; that is, when datasets are drawn from the same random vector but entail different instantiations of the noise associated with the sampling process.

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 noise. 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 [13, 10, 8].

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 under this framework.

Definition (Data distribution)**.** The (simple) random sample $\underline{\mathrm{X}} \overset{\text{iid}}{\sim} X$ has a probability distribution described by the density function $f_{\underline{\mathrm{X}}}$:

$$f_{\underline{\mathrm{X}}} = \prod_{n=1}^{N} f_X(x).$$

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

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

Definition (Sample)**.** Let X be a random variable associated with a measure of probability in $\mathcal{X}$. Let $\tau \in \mathbb{T}$ be a source of randomness allowed by such measurement. The set of transformations $\mathbb{T}$ is composed of the possible experimental conditions for the data sampling process from X. The

dependency of measurement realizations $\boldsymbol{x} \sim \underline{\mathrm{X}}$ on experimental conditions will be captured by index τ, and we will implicitly consider sample

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

to be a realization of experiment $\underline{\mathrm{X}} \overset{\text{iid}}{\sim} X$ under conditions τ. Given the stochastic nature of τ, we will also refer to it as noise instantiation.

These concepts will allow us to quantify the nature of the randomness that we are enforcing models to be invariant to. A dataset drawn from $\underline{\mathrm{X}}$ entails a noise instantiation τ that is not observed, which implies that we don't have access to the true information content of the data that derives from its true distribution $f_{\underline{\mathrm{X}}}$.

2.3.1 Posterior distribution

Definition (Hypothesis class)**.** Let $\mathcal{D}$ be the class of datasets generated from N-sized realizations of $\underline{\mathrm{X}}$. 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 Θ is the output space of hypothesis representing all possible outcomes of a function f learned on a dataset sampled from $\underline{\mathrm{X}}$ [7].

Intuitively, this framework interprets complexity from the perspective of the possible set of 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 Θ. Nevertheless, the underlying transformation is not homeomorphic in general, and more suitable generalization regularization constraints can be defined in Θ, 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 $\boldsymbol{x} \sim \underline{\mathrm{X}}$ 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 Θ can thus be postulated following this intuition.

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

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

Since we assume a mapping f and a data distribution $\mathbf{P}_X$, we can describe the randomness of the hypothesis outcome conditioned on the distribution of the data.

Definition (Posterior)**.** Let $\mathfrak{P}^f$ be a probability distribution family under consideration. A probability distribution over the hypothesis class can be defined as a conditional distribution given an realization $\boldsymbol{x} \sim \underline{\mathrm{X}}$. We will refer to this distribution as the posterior over Θ 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 Θ within the framework of probability theory. For instance, we can obtain the (prior) probability of a hypothesis to be selected by f as

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

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 define a robustness-based generalization error, we will proceed in an analogous way as we did in the previuous section. We will consider datasets $D', D'' \in \mathcal{D}$ each arising from a different sampling realization $\boldsymbol{x}', \boldsymbol{x}'' \sim \underline{\mathrm{X}}$, respectively. Both realizations are independent and they differ in the implicit noise entailed by their measurement:

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

Two posterior selection principles are derived from the robustness-informativeness trade-off:

P1 Posteriors should be expressive enough to cover the realizable subset of the hypothesis space.
P2 Equally likely inputs drawn from the same experiment should yield similar sets of hypothesis.

Definition (Description length)**.** Let $\mathcal{F}_\Gamma(\cdot)$ be the function class containing all functions represented by a parametrization Γ. 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 [19]. 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 [7], 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, we will normalize loss values by dividing by the description length of the prior:

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

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

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

It amounts to the expected loss over the normalized posteriors on the validation data $\boldsymbol{x}''$ weighted over the posterior distribution on the training data $\boldsymbol{x}'$. 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.

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

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

□

2.3.3 Maximum posterior agreement

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

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

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

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

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

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

Theorem 2.3.1. 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.

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

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

$$\mathcal{H}_{\boldsymbol{\Pi}^f}(\theta) \leq \log|\Theta|, \quad \forall\boldsymbol{\Pi}^f.$$

□

Chapter 3

Experimental setup

This chapter delineates the covariate shift setting within the supervised classification framework and introduces an operative formulation of posterior agreement. This formulation represents the cornerstone of this work as it allows for robustness-based model selection in discrete hypothesis classes.

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 $K \in \mathbb{N}$ be the cardinality of $\mathcal{Y}$. A K-class classifier can be defined as the composition of three functions:

- A feature extractor, mapping the input space to a d-dimensional feature space.

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

- A discriminant function, assigning a score to each of the K classes given a feature vector.

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

- A decision rule, yielding the class label from a vector of scores. We will set it to be the maximum a posteriori (MAP) rule.

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

A classifier is 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 are parametrized NN architectures in $\Gamma \subseteq \mathbb{R}^S$, such that:

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

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

The concepts defined in the previous chapter allow us to formalize 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 $\Gamma \subseteq \mathbb{R}^S$. Let RRM_D be the regularized risk minimization problem for c on D. Let $\mathcal{L}$ be the cross-entropy loss function for the classifier c.

$$\mathcal{L}(x, y) = -\log F_y(\Phi(x);\gamma)$$

The K-class classification problem is the RRM_D with loss function $\mathcal{L}$ parametrized in Γ.

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

No further characterization of the regularization factor will be provided in this chapter, as specific learning models and methods will be introduced together with the results.

3.2 Robustness in covariate shift settings

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

Definition (Covariate shift)**.** Let $\boldsymbol{x}'$ and $\boldsymbol{x}''$ be two N-sized samples of $\underline{\mathrm{X}} \overset{\text{iid}}{\sim} X$. A covariate shift exists between $\boldsymbol{x}'$ and $\boldsymbol{x}''$ if their (empirical) distributions are significantly different[1] for N large enough:

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

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 [39].

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 noise 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 $f_{X'} \neq f_{X''}$. The randomness entailed by their respective measurement process is also different in general (i.e. $\mathbb{T}' \neq \mathbb{T}''$). In such case

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

lead to a covariate shift known as out-of-distribution (OOD), given that the fundamental source of distribution shift is the difference in the probability measure over the support induced by each experiment. [39]

In the OOD case, $\boldsymbol{x}'$ 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 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 thus a covariate shift will be induced leading 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. [39]

Definition (Adversarial shift)**.** Let $\boldsymbol{x}' \sim \underline{\mathrm{X}}$ 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 adversarial examples, sampling randomness is not the source of distribution shift, as both $\boldsymbol{x}'$ and $\boldsymbol{x}''$ arise from the same realization of the experiment.

In this work we must consider a wider concept of sampling randomness that does not only comprise the implicit noise instantiation of each realization $\boldsymbol{x} \sim \underline{\mathrm{X}}$ 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 defined in the theoretical introduction.

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' and D'' be datasets generated from realizations $\boldsymbol{x}'$ and $\boldsymbol{x}''$, respectively. A robustness metric is a function $\Omega : \mathcal{D}'' \times \mathcal{F} \longmapsto \mathbb{R}$ that quantifies the generalization capability of a learned $\hat{f}_{D'} \in \mathcal{F}$ to observations in D''.

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

$$\mathrm{ACC}_{D'}(D'') = \frac{1}{N} \sum_{n=1}^{N} \delta_{y''_n} \left(\hat{c}_{D'}(x''_n) \right).$$

As it was argued before, we will interpret the concept of generalization from the perspective of the possible learning outcomes of a specific 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 resolution limit does not depend on the model but on the 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, regardless of the complexity of the model. The more complex the model is, the higher will be the resolution of its associated hypothesis space, but the more prone will be to overfit to the noise and thus yield unstable hypothesis. A regularization or model selection procedure derived from the robustness metric should then find the sweet spot between resolution and stability.

Properties (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 shift.

P2 (Independent discriminability) The metric should differentiate models only 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 [9]. Let us consider the following example.

Example 3.2.1. Let $\mathcal{D}$ be a class of balanced binary supervised datasets; that is, containing exactly the same number of observations of each label. We will examine the following three classifiers evaluating observations in $D \in \mathcal{D}$:

C1 A random classifier, returning a random prediction to each observation in the dataset. Overall performance tends to 50% accuracy as dataset size increases.

C2 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.

C3 A perfect classifier, returning the correct prediction to each observation in the dataset. Overall performance is 100% accuracy.

In terms of performance, **C1** and **C2** are equivalent when the dataset is large enough, and **C3** would be selected as the best. Nevertheless, a robustness metric compliant with **P2** would evaluate **C1** to be non-robust, while **C2** and **C3** would be considered equivalent and achieve maximum robustness, since their set of hypothesis remains the same for every dataset in $\mathcal{D}$.

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

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

3.3 Adversarial setting

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

Definition (Perturbation)**.** Let $\boldsymbol{x}'$ be a realization of $\underline{\mathrm{X}} \overset{\text{iid}}{\sim} X$ 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 ℓ_p-norm ball of radius ϵ centered at x. A perturbation Δ 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 ϵ_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 ϵ_p.

As it was previously outlined, the existence of adversarial examples in NNs was initially associated with their heavily non-linear nature and, as a consequence, to a lack of smoothness over the hypothesis space [45]. 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 [18].

Example 3.3.1. Let $w \in \mathbb{R}^d$ be the weight vector of a NN 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 the same intuition, we can define the most adversarial direction of perturbation as the one 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 [18].

An effective regularizer for adversarial training can be built by including the FGSM term on the objective that makes the model robust to ϵ_p-constrained perturbations [18]. A multi-step version can be immediately derived that sistematically 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 ϵ_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 Π is the projection operator. This is known as projected gradient descent attack [33].

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 oder 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 [33].

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:

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

This is known as the fast minimum-norm (FMN) attack [38].

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 OOD sense, and only a subset of them are available. We can formalize the problem as follows.

Definition (Domain generalization)**.** Let $\mathcal{S} = \{X_1^S, \cdots, X_S^S\}$ and $\mathcal{T} = \{X_1^T, \cdots, X_T^T\}$ 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 we will metonymically refer to them as domains. Set $\mathcal{S}$ encompasses source domains, and $\mathcal{T}$ target domains (see Section 1.1.2) [32, 50].

According to Definition 3.2, datasets sampled from each domain entail a OOD shift that will lead to non-zero generalization error. The domain generalization problem consists of selecting the model with the lowest generalization error between source target domains without having access to the 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 generating the data, and we expect our robustness assessment to be sensitive to the intensity of these variations.

In this work, we will also explore a more epistemologically grounded approach to robustness assessment that will be agnostic to the nature of the model 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 [20]), we will use the geometrical properties of the feature space and the resulting probability distribution over the output space to quantify the shift. Even though this approach might seem biased towards the specific parameters of the classifier network, we will redefine the model selection problem in a way that these definitions make sense.

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 [60].

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), we must eventually assess whether our metric is able to differentiate between robust and non-robust models. For that reason, we will consider ERM (see Definition 2.1) as our baseline vanilla model and compare its generalization performance to covariate shift with two models representing two different robustness enhancement strategies.

As a first approach, NN architectures will be trained by means of IRM, a regularization method driven by feature alignment [1]. 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 [32].

Definition (IRM)**.** Let $\mathcal{R}^d$ be the risk of a classifier c over domain $d \in \mathcal{S}$. The IRM problem minimizes risk over all domains while enforcing the feature extractor to yield domain-invariant representations [1]:

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

A surrogate version of the problem simplifies its implementation:

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

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 [1] 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 [59] and CutMix [57].

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 [55].

3.6 Robustness assessment with posterior agreement

As it was argued in Section 3.2, accuracy and by extension the custom law of quantifying robustness by reporting model performance in benchmark datasets does not offer any theoretical mechanism for the true characterization of robustness and ultimately depends on the nature of the shift that the model is being made robust to.

In this section we will derive a practical version of posterior agreement (PA) that can be used in supervised classification tasks to assess the generalization performance to different kinds of shifts. Even before reaching the final expression, a fundamental distinction between PA and accuracy can be made, namely the fact that posterior agreement is computed with the output of the discriminant function, which encodes the confidence in the classification, whereas accuracy only considers the output prediction label. Confidence information increases the discriminative power of the metric, for instance when comparing models with similar predictive capabilities, but also involves a probability distribution over the output space that allows for a broader theoretical interpretation.

3.6.1 Posterior in classification tasks

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

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

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

Theorem 3.6.1 (Classification posterior)**.** Let Θ be the classification hypothesis class associated with the K-class classification problem with approximating function c. The posterior distribution class $\mathfrak{P}^c$ is the Gibbs distribution family with inverse temperature parameter β [9]

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

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

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

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:

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

Its derivative with respect to $\mathbf{P}^c(\theta \mid \boldsymbol{x})$ is:

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

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 [5].

Definition (Classification confidence)**.** Let $D\in\mathcal{D}$ be a dataset associated with a realization $\boldsymbol{x}\sim\underline{\mathrm{X}}$. Let $F_j(\cdot;\gamma)$ be the j-th component of the score vector returned by the discriminant of the classifier. The cost function driving posterior selection will be the negative confidence in the prediction:

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

where θ_i is the class label associated with the i-th sample in the dataset.

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,

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

Proof. See Appendix A.1. □

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

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

Proof. See Appendix A.1. □

Theorem 3.6.3 (PA kernel for classification)**.** Let $\boldsymbol{x}'$ and $\boldsymbol{x}''$ be N-sized realizations of $\underline{\mathrm{X}}$. Let Θ be the hypothesis class represented by classifier c under $\mathcal{D}$. With no prior information about Θ, the posterior agreement kernel for supervised K-class classification tasks has the following expression:

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

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

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

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, Θ 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:

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

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

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.

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

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

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

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 3.6.4 (Properties of the PA kernel)**.** $\mathrm{PA}(\boldsymbol{x}', \boldsymbol{x}''; \beta)$ has the following properties.

P1 (Non-negativity) $\mathrm{PA}(\boldsymbol{x}', \boldsymbol{x}''; \beta) \geq 0 \;\; \forall \boldsymbol{x} \sim \underline{\mathrm{X}}$ and $\beta \in \mathbb{R}^+$.

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 noise instantiations are not indexed and no reference noiseless experiment can be performed.

P3 (Concavity) $\mathrm{PA}(\boldsymbol{x}', \boldsymbol{x}''; \beta)$ is a concave function of $\beta \in \mathbb{R}^+ \; \forall \boldsymbol{x} \sim \underline{\mathrm{X}}$. This means that the kernel optimization problem will have a unique solution [6].

Proof. See Appendix A.2. □

Theorem 3.6.5. The posterior agreement kernel for classification tasks complies with the desired properties of a robustness metric (see Properties 3.2).

Proof. See Appendix A.2. □

3.6.3 Implementation

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

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

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

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

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

```
from pametric import PosteriorAgreementBase, LogitsDataset

# Initialization
pa_metric = PosteriorAgreementBase(pa_epochs, beta_0)

# Within the training or evaluation logic:
logits_dataloader = DataLoader(LogitsDataset([logits0, logits1], y))
pa_results = pa_metric(logits_dataloader)
logPA = pa_results["logPA"]
```

Listing 3.1: PA metric implementation.

[2] https://lightning.ai/docs/torchmetrics/stable/

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

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

The complete code implementation along with the set of unit tests conducted to ensure consistency between training and data processing strategies can be found in the `pa-metric` repository[3].

[3]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 adversarial and domain generalization settings. Evidence supporting its suitability against baseline accuracy measures will be provided, thus establishing PA as a reliable algorithm selection criterion in these scenarios.

4.1 PA as a robustness metric

4.1.1 Empirical behaviour

Starting from the simplest possible setting, we will explore the behavior of the metric under different levels of sample mismatch. More specifically, we will assess the performance of perfect, random, and constant binary classifiers by manipulating a Bernoulli sample and simulating different levels of prediction confidence.

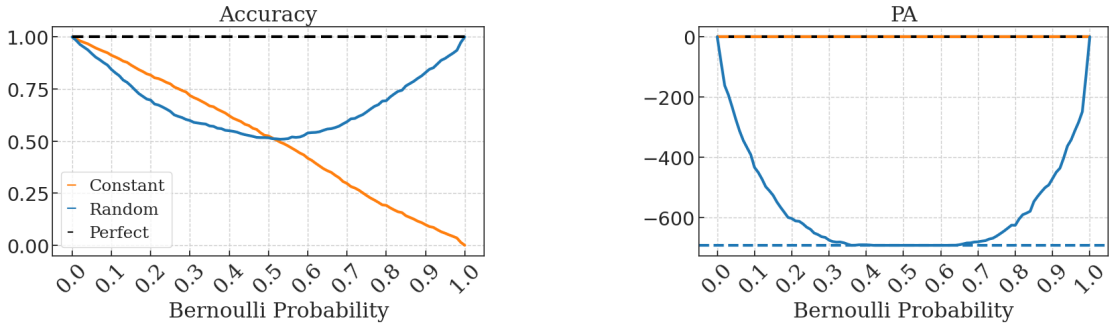


Figure 4.1: Evolution of performance and robustness for the three classifiers

A $N = 1000$ Bernoulli sample was generated with a symmetrical confidence level of prediction $\pm\Delta/2$. For $p = 0.5$, the metric achieves its minimum value $N \log 1/2$ for the random classifier, as $\beta = 0$, and tends to its maximum value of 0 as $\beta \longrightarrow \infty$.

Classifier	Accuracy	β	PA
Perfect	1.000	12.331	-0.0088218
Constant	0.525	12.331	-0.0088218
Random	0.516	0.000	-693.14

Table 4.1: Comparison of classifier performance metrics for $p = 0.5$.

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

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

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

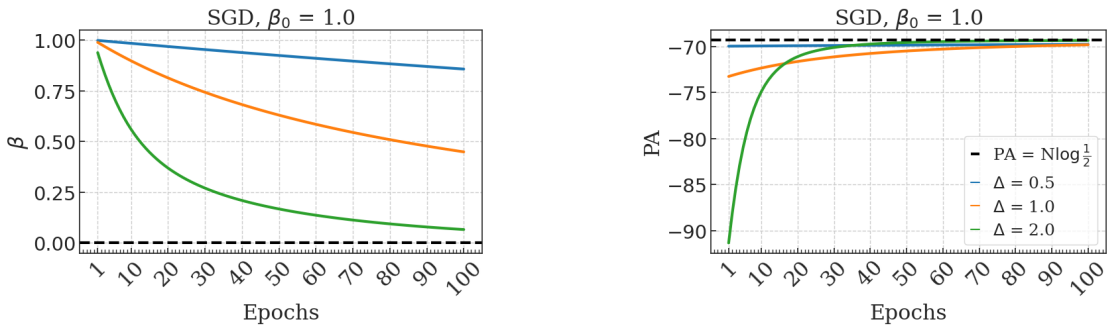


Figure 4.2: Evolution of PA kernel optimization under different levels of prediction confidence. An illustration of the original log-odds and its associated posterior distribution can be found in Appendix B.1.1.

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

4.1.2 Robustness assessment to sampling randomness

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

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

HERE THE PLOT OF THE LEVENSHTEIN DISTANCE PLOT HERE.

A sentiment classifier analysis. In this case, the random nature of the noise perturbations is shown to exploit specific vulnerabilities of language models, which paradoxically have been shown to be robust to more highly crafted perturbations, such as changes in language or replacement of complete words.

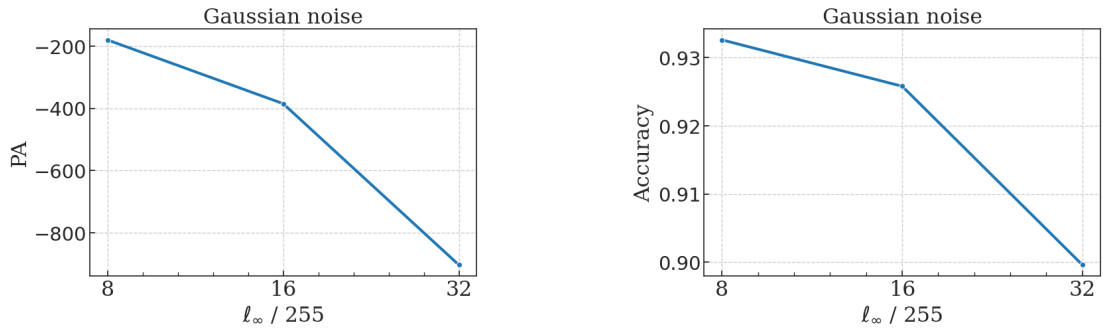


Figure 4.3: PA and accuracy of CIFAR10 classification for increasing levels of white noise intensity.

In this second example, a 10.000 observation sample of CIFAR10 images was perturbed with white noise at different levels of intensity. The magnitude of the perturbation is expressed in the same terms as those of an adversarial attack (see Section 3.3) 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 $\pm 3\sigma$.

As expected, PA is highly sensitive to the presence of white noise, and is able to capture the generalization capabilities of the model in a much more informative way than accuracy. We can obtain a higher understanding of the degree of sensitivity of the metric if we adjust the perturbation so that if only affects a certain ratio of observations.

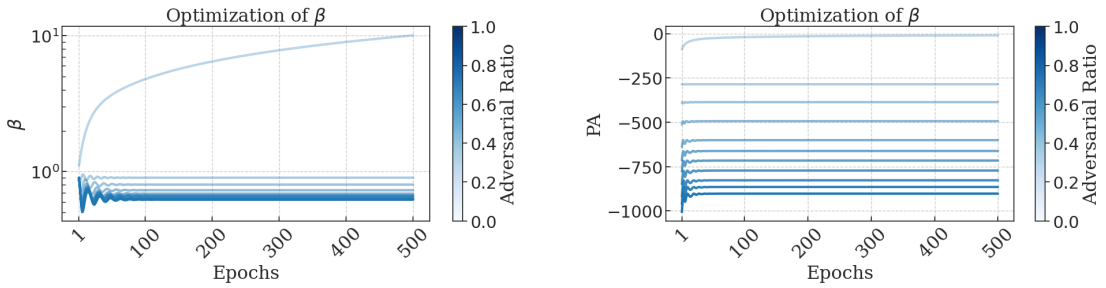


Figure 4.4: PA kernel optimization in the CIFAR10 gaussian noise setting for different ratio of perturbed samples. Perturbation magnitude is $\ell_\infty = 32 \ / \ 255$.

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

4.2 Adversarial setting

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

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

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

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

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

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

The results provided in this section have been obtained using the CIFAR10 dataset [28], which is widely regarded as a standard benchmark for robustness evaluation. CIFAR10 is a balanced dataset containing 60.000 coloured 32 $\times$ 32 pixel images belonging to 10 different classes. We will consider a pre-trained WideResNet-28-10 as a baseline, undefended model and compare it to some state-of-the-art robust models provided by the RobustBench [15] library under PGD [33] and FMN [38] attacks, both run for a thousand steps (see Section 3.3). The PGD attack power will be specified in terms of ℓ_∞, 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 4.5: Original and adversarially-perturbed CIFAR10 sample of class `horse`. Both perturbations succeed at misleading an undefended, pre-trained WideResNet-28-10 net.

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

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

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 AR and converge faster to the the AR = 1 robustness value than any accuracy-based metric, in light of the results obtained in the previous section. We can quantify the model discriminability over increasing AR by computing the adversarial ratio gap $\Delta\,\mathrm{AR}$.

Definition (Adversarial ratio gap)**.** Let γ_+ and γ_- be two models, not necessarily different. Let AR_+ and AR_- be the adversarial ratio values such that

$$\mathrm{PA}^{\gamma_+}\Big|_{\mathrm{AR}_+} = \mathrm{PA}^{\gamma_-}\Big|_{\mathrm{AR}_-}$$

the aversarial ratio gap ($\Delta\,\mathrm{AR}$) is obtained as

$$\Delta\,\mathrm{AR} = \mathrm{AR}_+ - \mathrm{AR}_- \,.$$

Before delving into the results, it is worth exploring the immediate consequences of the previous claim, namely the fact that the maximum posterior agreement will be achieved when gibbs distributions are highly peaked on the predicted class, at least for moderately aggressive attacks.

This is because most adversarial samples will not succeed at misleading the model and thus drive the inverse temperature to infinity. The divergence of β^* is only limited by the set of misleading adversarial samples, that for being perturbed from the original class are still expected to assign a significant confidence to the original prediction, even if not the maximum anymore. Table 4.2 illustrates this claim by showing that $\beta^* > 1$ for all robust models, resulting in a substantial decrease of the entropy between initial and optimal posteriors.

Defense	β^*_{PGD}	ΔH_{PGD}	β^*_{FMN}	ΔH_{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 $\Delta H = H(\beta^*) - H(\beta)$ for different models, obtained for FMN and $\ell_\infty = 8/255$ PGD attacks, both at AR $= 1$. Entropy values are estimated using the average posterior distribution over correctly classified samples, which constitute the largest proportion of the dataset. Figures B.6-B.11 show the initial and optimal average posteriors from which these values were computed.

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

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

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

where

$$\begin{aligned}
\Xi_{\text{ERR}} &= N\tau\rho \log\left(1 - 2\delta_{\text{ERR}}\right), \\
\Xi_{\text{MIS}} &= N(1-\tau)\rho \log\left(1 - 2\delta_{\text{MIS}}\right), \\
\Xi_{\text{ADV}} &= N\tau(1-\rho) \log \delta_{\text{ADV}},
\end{aligned}$$

where τ is the accuracy of the model in the original data and $\rho \equiv \text{AFR (P)}$. Variables δ_{ERR}, δ_{MIS} and δ_{ADV} account for the average probability assigned to classes other than the predicted class for the three aforementioned cases (see illustration in Figure B.14). Ξ_{SAM} aggregates the first two terms and will be interpreted as the sampling randomness contribution.

Proof. See Apendix B.2. □

Figures B.15 and B.16 compare the true and approximated PA values under increasing adversarial ratio for PGD and FMN attacks, respectively. It is clear that penalizations are overestimated,

given that the average posterior probability was used and differences by defect are more significant than those by excess due to the nonlinear nature of the logarithm in the range $[0, 1]$. Besides, β^* is fixed to its lowest possible value; that is, when AR $= 1$. This makes the approximation on the FMN attack less reliable for smaller adversarial ratio settings, as β^* decreases significantly due to the effectiveness of the attack.

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

4.2.1 Adversarial robustness assessment with PA

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

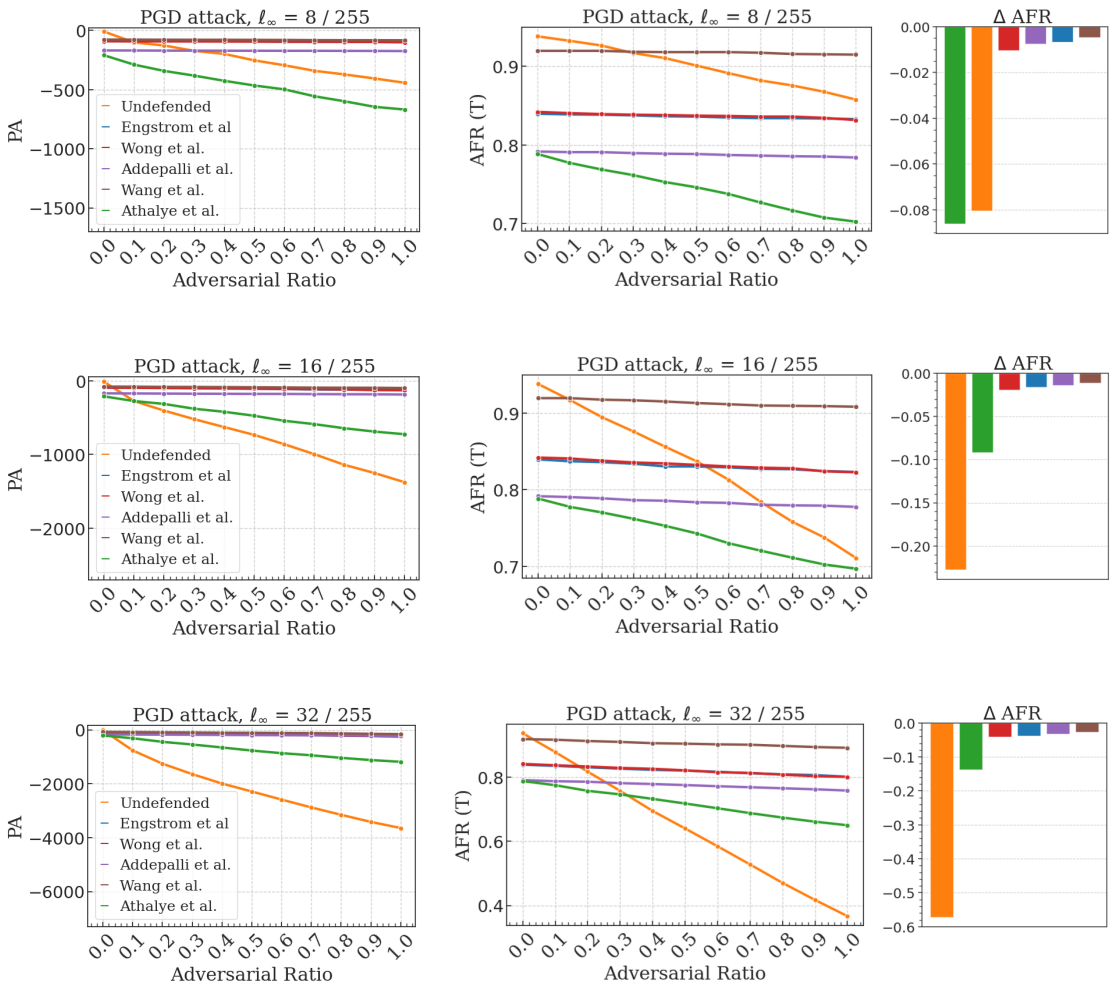


Figure 4.6: PA, AFR(T) and the AFR variation against increasing adversarial ratio at different perturbation norm bounds. The aforementioned undefended net and several RobustBench robust models are considered under a 1000 step PGD attack.

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

From both PA and AFR stems the fact that **Athalye et al.** is significantly less robust to PGD attacks than its RobustBench counterparts, as its performance decreases way more significantly with increasing AR. It is interesting to see, however, than the rate at which its performance decreases is inversely proportional to the attack power, which indicates that the principle by which robustness is achieved is more effective for large-norm perturbations.

A fundamental difference between these two models, that cannot be inferred from a purely performance-based metric, is the nature of the shift in the probabilistic output of the model, which is the source of the robust and non-robust behaviour observed. Figure 4.7 **(right)** shows the optimal β^* value for each model, which is an indication of the entropy of the posterior distribution and discriminates the two non-robust models from the rest and from each other. The **Undefended** model provides overconfident predictions that maximize disagreement in misleading and misclassified samples, whereas **Athalye et al.** provides uncertain predictions that minimize disagreement in adversarial samples but have the opposite effect in correctly classified ones. This insight clarifies the unintuitive behaviour observed earlier, by which **Athalye et al.** robustness value decreases at a lower rate with increasing attack power, despite maintaining a constant decrease in performance of $\Delta\,\text{AFR} \sim 0.1$.

Figure 4.7 **(left)** illustrates the previous reasoning by displaying the average posterior probability assigned to the predicted class by each model, conditioned on the type of prediction assigned. This discrimination yields three groups of observations, namely original samples that are correctly classified by the robust model, original samples that are misclassified, and perturbed samples that, having their associated unperturbed sample been correctly classified, have been able to mislead the model. These three cases are relevant from the adversarial robustness perspective, as they illustrate the trade-off between high-confident original predictions and adversarial vulnerability, which has been already stated in previous chapters. **Wang et al.** acts as a reference for an ideal robust behaviour, in which original samples are predicted with high confidence and adversarially misleading predicted labels are only slightly more likely than the rest. Equivalent representations for the remaining models can be found in Figure B.17.

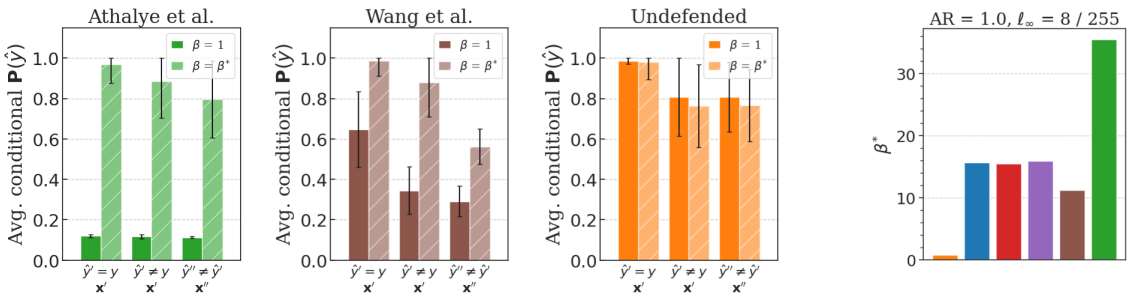


Figure 4.7: (**left**) Average posterior probability of the predicted class for correctly classified original samples, misclassified original samples, and misleading adversarial samples, respectively. (**right**) Optimal β^* value for each model. Results obtained through a PGD attack with $\ell_\infty = 8/255$.

With respect to robust models, we observe a significant difference in the discriminative power of PA and accuracy-based metrics that does not immediately derive from the informativeness of the optimal posterior. As remarked before, AFR (P) constitutes our baseline robustness metric, as

by definition represents the ratio of predictions that remained constant under adversarial perturbations, and therefore ranks models by their predictive capabilities against these attacks. The value of ΔAFR 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.

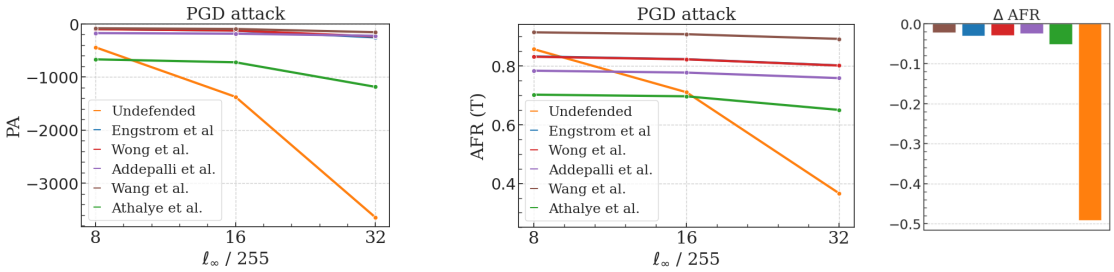


Figure 4.8: PA, AFR(T) and the AFR variation against increasing attack power for AR = 1. The aforementioned undefended net and several RobustBench robust models are considered under a 1000 step PGD attack.

Finally, Figure 4.8 shows that PA is also discriminative with respect to increasing attack power, expressed through the maximum allowed ℓ_∞ 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 ℓ_∞ is much more significant that the decrease in performance. This can be explained by the fact that the metric is sensitive to the overall posterior shift and not only the position of the maximum. When increasing the attack power, confidence in the predicted class will decrease in general, even when the sample does not succeed at misleading the model, and therefore the overall overlap between posteriors will be reduced even at comparable performance levels. This observation further illustrates the independent discriminability power offered by PA (see Properties 3.2), which constitutes the cornerstone argument of this work.

In order to widen the scope of the analysis, analogous results will be obtained for FMN attacks, which are expected to be more effective than PGD attacks for being unbounded, which translates into an overall decrease in β^* (see Table 4.2). Figure 4.9 shows the evolution of PA against increasing adversarial ratio for the same models, and compares it with the assessment provided by AFR.

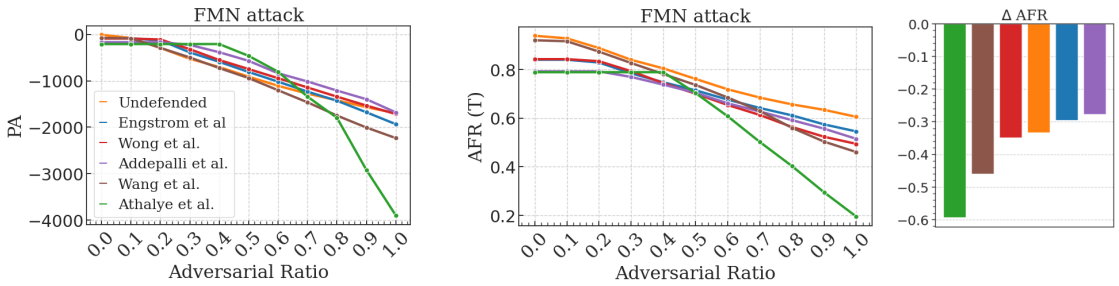


Figure 4.9: PA, AFR(T) and the AFR variation against increasing adversarial ratio. The aforementioned undefended net and several RobustBench robust models are considered under a 1000 step FMN attack.

As expected, the effectiveness of the FMN attack is superior to that of PGD attacks, as the decrease in performance is substantially more significant for all models, especially the ones previously considered robust. It is likely that these models have been defended with a compression strategy that succeeds at filtering out small perturbations, which are the ones employed by FMN, and for that reason maintain their performance at low adversarial ratio values [16]. In particular, **Athalye et al.** remains maximally robust until at least 40% of the samples are perturbed, at which point the defensive strategy is neutralized and a constant fraction of the additional perturbed samples succeeds at misleading the model, which translates into a linear decrease in performance and PA.

PA proves to be very discriminative among robust models and to represent the phase transition entailed by the collapse of the defense strategy better than AFR does, which can be observed in more detail in Figure B.13. A significative result is that PA is not so directly aligned with ΔAFR, in contrast to the PGD case, which shows again that the decrease in performance is not the main driver of the robustness assessment provided by PA, but instead can be interpreted as a consequence of a misalignment in the posterior distributions of adversarial samples, which are the ones driving the metric after the AR thresold is reached.

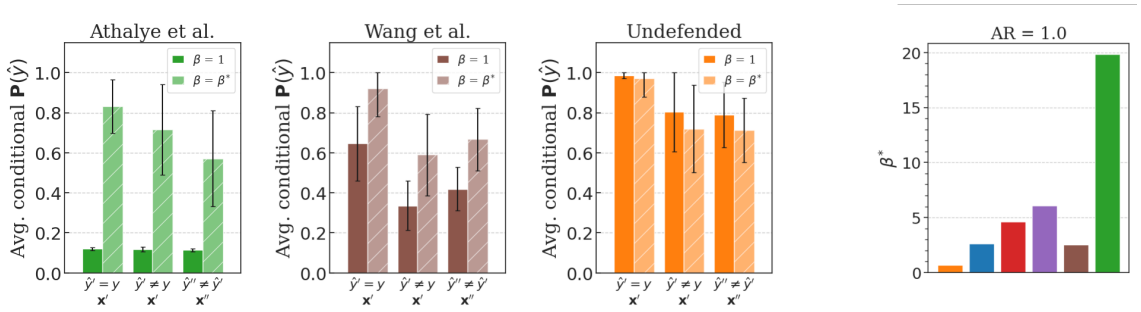


Figure 4.10: (**left**) Average posterior probability of the predicted class for correctly classified original samples, misclassified original samples, and misleading adversarial samples, respectively. (**right**) Optimal β^* value for each model. Results obtained through a FMN attack.

Figure 4.10 gives insight into the probabilistic output of the model and the informativeness of the optimal posterior for the **Undefended**, **Athalye et al.** and **Wang et al.** models, in analogous way to the PGD experiments. The first two models display a very similar behaviour for $\beta = 1$, but optimal posteriors are less informative due to the increased number of misleading samples, which translates into a smaller β^*. The response of the **Wang et al.** model further illustrates the higher effectiveness of FMN attacks, as adversarial perturbations are on average more misleading than outlier samples in the original dataset, which did not occur in the PGD case. Analogous representations for the remaining models can be found in Figure B.18, which show that the **Addepalli et al.** is the only robust model that maintains the same behaviour under both attacks.

Overall, we recognize that PA has a higher discriminative power than AFR, especially considering the evolution of each metric over increasing adversarial ratio, as seen in Figures 4.6 and 4.9. In particular, Figures B.12 and B.13 compare the evolution of PA with that of AFR (P), which is the baseline metric for robustness, and show the susceptibility of the latter to dataset variability. This is an important consideration, as PA not only improves the discriminability in terms of the scale of the differences between models, but also provides a more stable assessment across varying levels of perturbed sample presence, under which AFR (P) exhibits significant fluctuations that alter the ranking of the models at every step, as illustrated in Table 4.3.

These observations lead to the conclusion that PA offers a more reliable assessment of adversarial robustness, which in general aligns with the decrease in performance on perturbed samples, but

Defense	AR = 0.0		AR = 0.2		AR = 0.4		AR = 0.6	
	PA	AFR_P	PA	AFR_P	PA	AFR_P	PA	AFR_P
Addepalli et al.	**-169.45**	1	**-172.51**	**0.9956**	**-175.46**	0.9920	**-177.63**	0.9896
Wong et al.	-91.47	1	-97.68	0.9960	-102.90	0.9920	-109.21	**0.9876**
Engstrom et al.	-89.20	1	-94.21	0.9964	-104.61	**0.9904**	-110.34	0.9888
Wang et al.	-77.83	1	-81.95	0.9976	-84.58	0.9956	-89.39	0.9916

Table 4.3: Comparison of PA and AFR_P for a PGD attack with $\ell_\infty = 16 \ / \ 255$ across different adversarial ratio values. The worst robustness score is emboldened for every case. PA displays higher consistency and discriminative power across varying AR with respect to to AFR_P.

that relies heavily on the informativeness of the posterior distribution and the confidence in the predictions for both original and perturbed samples.

4.2.2 Interpretability of PA in the adversarial setting

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

Table 4.4 shows the contribution of each subset of observations to the final approximated PA value for a PGD attack. N_{ERR}, N_{MIS} and N_{ADV} are the number of (pairs of) contributing samples, and Ξ_{ERR}, Ξ_{MIS} and Ξ_{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.15 and B.16. The parameters $2\delta_{\text{MIS}}$ and δ_{ADV} account for the average probability assigned to classes other than the predicted class for misclassified original samples and misleading adversarial samples, respectively, and help interpret the informativeness of the distibution as well as the value of each individual penalization.

For instance, a large $2\delta_{\text{MIS}}$ value indicates robustness to sampling randomness, as it represents higher average uncertainty in misclassified predictions. A model with a high performance on test data entails a more negative penalization $\log(1-2\delta_{\text{MIS}})$, for being misclassified samples more likely to be equivalently misclassified under adversarial perturbations, but at the same time makes misclassifications less likely, and therefore the number of terms added to Ξ_{MIS}. The existing trade-off between standard and robust generalization arises when following this reasoning towards the minimization of Ξ_{MIS}, because reducing the number of misclassified samples will drive β^* to higher values and therefore decrease adversarial uncertainty δ_{ADV}. As outlined before, δ_{ADV} indicates robustness to adversarial perturbations, as it represents the average prediction uncertainty on adversarial misleading samples, and entails a penalization of $\log(\delta_{\text{ADV}})$.

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

Defense	N_{MIS}	$2\delta_{\text{MIS}}$	Ξ_{SAM}	N_{ADV}	δ_{ADV}	Ξ_{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 4.4: Approximated PA contributions for a PGD attack with $\ell_\infty = 8/255$ and AR = 1.0. The number of originally misclassified and adversarially misleading samples is $N_{\text{MIS}} = \lfloor N(1-\tau)\rho \rfloor$ and $N_{\text{ADV}} = \lfloor N\tau(1-\rho) \rfloor$, respectively. The penalization argument $2\delta_{\text{ERR}}$ has not been included for being negligible in all cases.

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

The same reasoning can be followed to explain the discrimination made by PA between **Addepalli et al.** and the other robust models. **Addepalli et al.** experiences a comparable drop in performance, and for displaying a reduced confidence in mismatching predictions is assigned a smaller Ξ_{SAM} contribution than some of these models. Nevertheless, such uncertainty is also observed for original samples, which lowers accuracy on the original dataset and thus increases random sampling penalization Ξ_{SAM}. In that sense, it can be argued that **Addepalli et al.** is more robust than **Wong et al.** and **Engstrom et al.** to adversarial perturbations, which also stems from the baseline AFR (P) and ΔAFR 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 Ξ_{ADV} is driven by the decrase in performance under attack ΔAFR, as N_{ADV} penalization terms are added. Nevertheless, the value of each of these terms is $\log(\delta_{\text{ADV}})$, which penalizes models that achieve maximum posterior agreement by increasing confidence on adversarially misleading examples. This is a clear distinctive trait with respect to accuracy-based metrics, whose penalizations are reduced to a binary decision.

Defense	N_{MIS}	$2\delta_{\text{MIS}}$	Ξ_{SAM}	N_{ADV}	δ_{ADV}	Ξ_{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 4.5: Approximated PA contributions for a FMN attack with AR = 1.0. The number of originally misclassified and adversarially misleading samples is $N_{\text{MIS}} = \lfloor N(1-\tau)\rho \rfloor$ and $N_{\text{ADV}} = \lfloor N\tau(1-\rho) \rfloor$, respectively. The penalization argument $2\delta_{\text{ERR}}$ has not been included for being negligible in all cases with the exception of the **Athalye et al.** model, which amounts to 0.36.

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

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

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

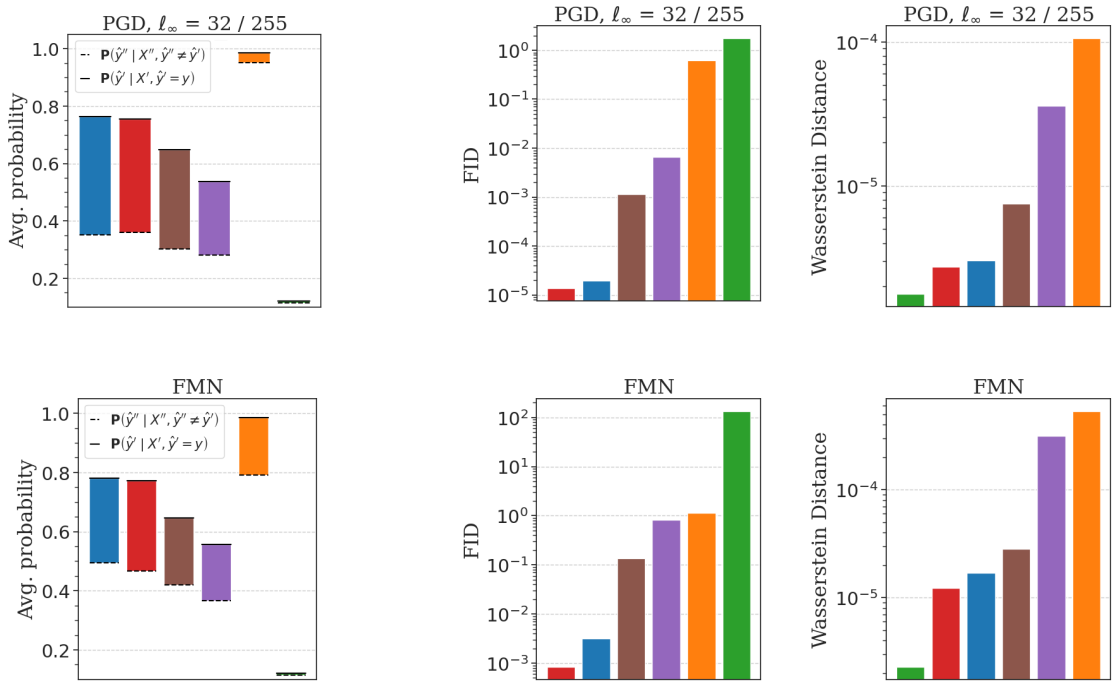


Figure 4.11: The two metrics considered are FID, which amounts to the group-based dissimilarity in the feature space, and Wasserstein distance, which measures the average distance between probability distributions.

Figure 4.11 displays some of these metrics for PGD and FMN attacks, together with a condensed visualization of the average probability gap existing amidst predictions on original and adversarial examples. As we can see, the discriminative power of these metrics stems from the overall difference between posteriors at $\beta = 1$. This result is not surprising, since the discriminant function of the classifier is constituted by a small subset of layers in the network, and is not expected to represent a complex discrimination that cannot be inferred from the distribution of samples in the feature space. As expected, probability-based metrics tend to concide in their assessment in general terms, and feature-space-based metrics as well. Besides, it is also not surprising that feature space distances are more sensitive to the success rate of the attack, given that the inductive bias remains constant under any condition, whereas confidence-based metrics are susceptible to the overall shift in the posterior, which in general favours non-informative posteriors. A deeper insight into the evolution of these metrics for PGD and FMN attacks under increasing adversarial ratio can be found in Figures B.19 and B.20.

The results obtained in this section showcase the rationale behind the maximization of posterior agreement, which can be thought of a surrogate version of the maximization of the mutual information between original and perturbed datasets [7]. From an information-theoretic perspective, the maximization of mutual information effectively distillates the information contained in the posterior distributions to that which is relevant for robustness assessment. This property is the source of the discriminative power displayed by PA and constitutes a fundamental difference with respect to the baseline accuracy-based metrics considered.

4.3 Domain generalization setting

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

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

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

Experiments will be performed by means of the DiagViB-6 dataset framework [17], which comprises MNIST images of size 128x128 within an augmentation pipeline enabling the modification of six specific image factors: shape, hue, lightness, position, scale and texture. Several variations in the `diagvibsix` library[1] have been implemented with the purpose of this project so that datasets can be built with a specific configuration of factors for each sample, which allows for a wide range of experiments in the data shift assessment and model selection settings. In an analogous way to the adversarial case, datasets will be incrementally perturbed by including only a fraction of the shifted samples, which in this context we will call shift ratio (SR).

Following the notation introduced in Section 3.4, Definition 4.3 provides a characterization of source and target domains and the randomness entailed by each dataset. The control over these aspects is the rationale behind this experimental setup, since it is through synthetic image manipulation that we can maximize invariant feature learning possibilities during training while providing optimal robustness assessment conditions in validation and testing. Since changes in image factors can be independently introduced to each sample, the shifted dataset contains the same samples and in the same order as the original dataset, thus removing sampling randomness contributions from the robustness score. Table 4.6 stipulates the specific factors conforming each environment in this experiment and Figure 4.12 illustrates them with some examples.

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

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

where X_j represents the random variable associated to domain j, being j the number of shifted factors with respect to the original MNIST sample. Datasets are generated by considering four different realizations of the experiment, namely τ_0^{train}, τ_1^{train}, τ^{val} and τ^{test}, each sampling from disjoint subsets of MNIST. Following the notation introduced in Chapter 3, we can define:

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

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

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 testing datasets entail each a single noise instantiation, which means that distribution shift is the only accountable source of randomness. Overall, two sets of 40 000 images for training, two sets of 20 000 images for validation, and six sets of 10 000 images for testing are generated.

# Shift Factors	0	1	2	3	4	5
Hue	red	**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.

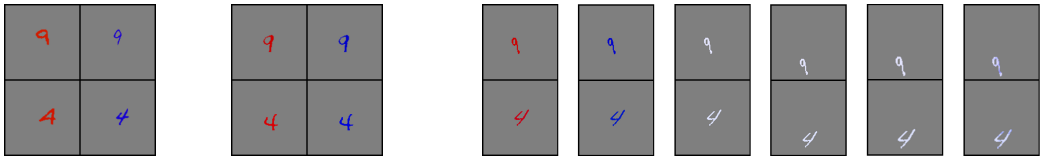


Figure 4.12: Illustration of the training, validation and test datasets. Samples for each training environment belong to different MNIST subsets, whereas samples of validation and test are corresponding.

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

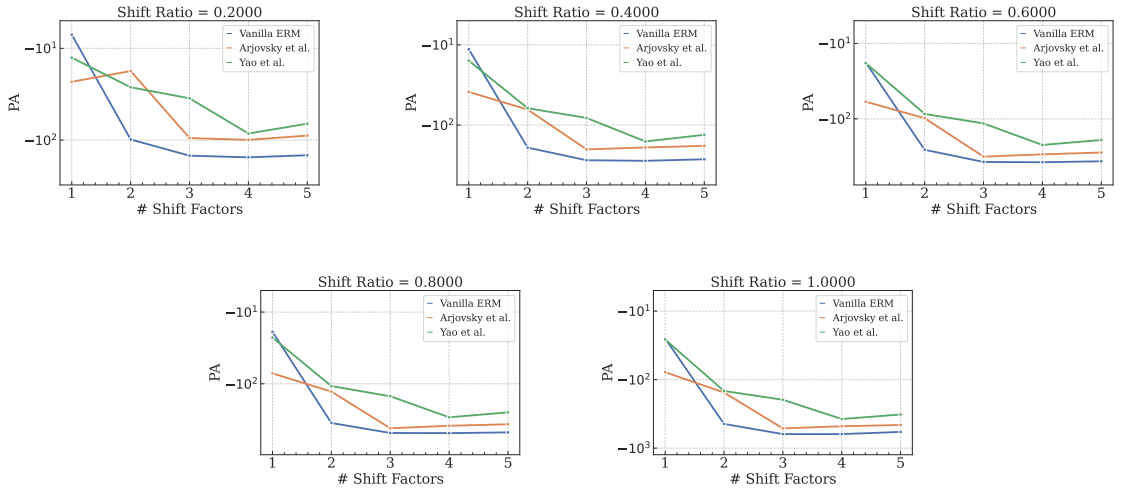


Figure 4.13: Evolution of PA under increasing levels of shift power. Weights maximizing validation accuracy were selected for ERM, IRM and LISA algorithms. Results for incremental presence of shifted samples indicate that PA is able to differentiate weak and robust models.

The unexpected increase in PA after four shifted factors seems inconsistent with the alleged non-increasing behaviour of PA, as per Section 3.2. Nevertheless, this phenomenon only highlights that robustness does not stem from the data generation process but instead from the latent representation of the model and the features selected for the construction of its inductive bias, as discussed in the introductive chapter. In this particular case, Table 4.7 shows that there is a clear discontinuity in the feature representation of the data when the texture factor is shifted, which results in a different discriminator function that ultimately leads to a predictive behaviour that aligns slightly better with the original predictions.

# Shift Factors	1	2	3	4	5
ERM	0.9978	0.9303	0.9562	0.9561	**0.6661**
IRM	0.9967	0.9018	0.9296	0.9374	**0.5585**
LISA	0.9980	0.9431	0.9431	0.9641	**0.7130**

Table 4.7: Pairwise cosine similarity between feature space representations of original and augmented images, for each of the shifted datasets. The abrupt decrease in similarity for the fifth environment indicates a discontinuity in the feature representation of images, which leads to non-comparable predictive outcomes.

As shown in Figure 4.14, PA-based model selection behaves differently in robust and non-robust algorithms. Following the reasoning derived in previous sections, it is likely that the PA assessment on ERM is mostly driven by standard generalization error, due to the fact that ERM is agnostic of the environment to which each sample belongs to and therefore to the accountable source of randomness represented by the environment shift. The risk minimization problem under these conditions should give rise to less informative predictive outcomes, given that the domain-invariant features associated with the task are less accessible, and thus reduce the penalization weight of mismatching samples. This intuition is supported by the fact that ERM optimization yields models that performs worse in all test datasets than the ones obtained through IRM, as can be seen in Figure **??**. Besides, results show that the selected model for ERM performs better an all metrics for the first test environment, which contains the same factor configuration as one of the validation environments, and maintains or slightly decreases its performance for increasing levels of shift. In addition, the PA-selected model appears to encode a slight variation of the decision rule from that

of the accuracy-selected model, given that samples near to the decision boundary are now more likely classified as nines than fours, which increases specificity as much as it decreases sensitivity.

Regarding PA-based model selection in robust learners, we observe the opposite behaviour. Since models are more likely to represent domain-invariant features, each with its distinctive strategy, predictive outcomes are expected to be more informative and thus increase the penalization contribution of mismatching samples, which will be the ones containing domain-invariant features that are not sufficiently considered in the inductive bias of the model. PA model selection under these conditions is thus more likely to favor sets of weights that perform better under increasing levels of shift, at the expense of losing predictive power on the environments in which it operated. Results align with this interpretation, as we observe a notable increase in performance across all shifted datasets with respect to the accuracy-selected model, especially in the LISA case, at the expense of significantly decreasing performance on the first environment.

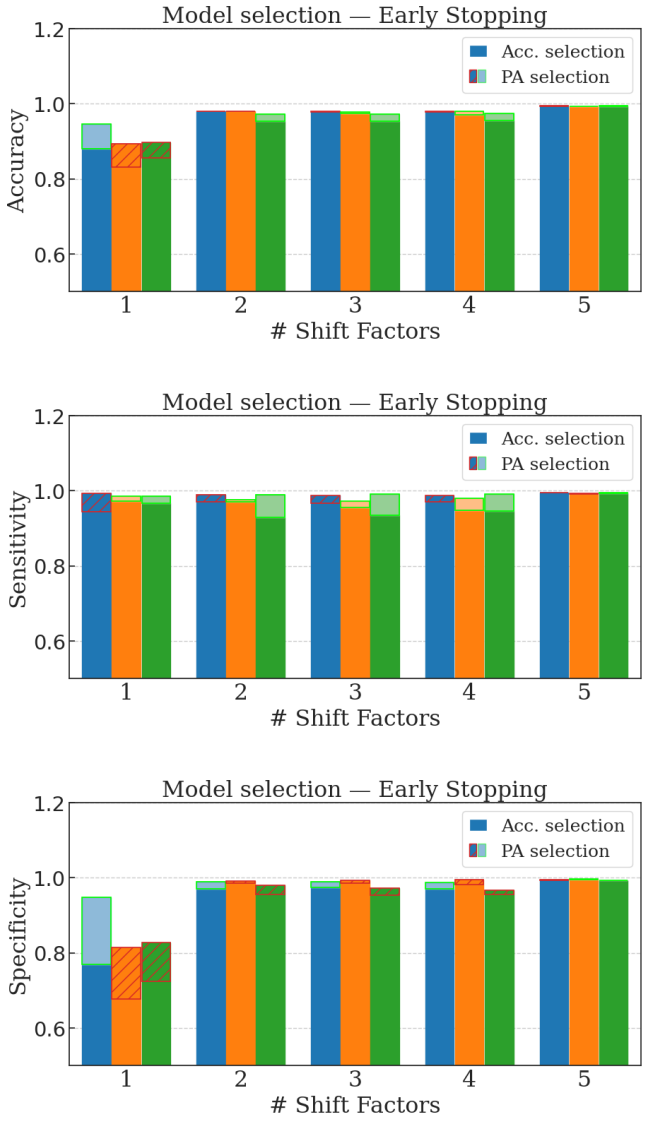


Figure 4.14: Accuracy, sensitivity and precision displayed by sets of ResNet18 weights on shifted test datasets, obtained through ERM, IRM and LISA training procedures. Accuracy-based selection is compared to PA-based selection, both operating with a validation dataset composed of samples of environments 0 and 1.

Chapter 5

Model selection

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

5.1 Model selection under controled experimental conditions

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

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

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

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

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

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

HERE PLOT OF THE POSTERIORS.

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

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

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

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

HERE DATASETS.

- Results to show: Table with accuracy and F1 of the selected models under all conditions. Maybe compare validation accuracy, etc

- Single plot. x axis is the configuration of the validation and model selection dataset, and y axis is the percentage of increase in performance of the accuracy-selected model With respect to the PA-selected model.

5.2 Vulnerabilities of PA for GO/SO

In the preceding chapters, evidence was provided supporting PA as a suitable robustness metric that effectively captures generalization capabilities under both sampling randomness and covariate shift. So far, experiments in the domain generalization setting have been conducted under synthetic

conditions in which distribution shift is the only accountable source of randomenss between $\boldsymbol{x}$ and $\boldsymbol{x}''$. These experiments have shown that PA successfully discriminates robust from non-robust learners and also provides increased early-stopping performance compared with current baseline metrics.

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

Epoch-wise model selection under these conditions entails a fundamentally different approach, especially regarding experiments performed in the previous chapter. Evaluating PA on a model selected by performance standards (i.e. validation accuracy) and
- Here interpretation of the validation accuracy
can be used as an early stopping criterion for model selection. Nevertheless, real world applications do not usually have a
More especifically, we aim at characterizing the response of the metric under different variations in the inductive bias of the model, which implicitly shifts according to the availability of certain features in the data. In this regard, two
1. Why epochwise is different than general model selection. KEY: Validation accuracy. The main problem in any robustness measurement is that we are not able to distinguish sampling randomness from robustness to adversarial shifts
2. We want to explore the difference between validation accuracy and PA for different inductive bias.
3. Results show that
If the goal is to show that it performs better than accuracy, that can be easily shown, but the improvement is non-significative. . But the principal challenge is theoretical: real-world covariate shift encompasses both distribution shift and sampling randomness, and PA (in the way we have been handling it) does not distinguish between them. When we compare the evolution of the training in between epochs, we will select the most robust features, but they won't necessarily be the most generalizable ones, they will depend on the nature of the data. => Here follows experiments on GO/SO. Explain the conclusions derived by those experiments. 1) Even when features are very similar, PA keeps increasing. 2) Model with more generalization opportunities is the least robust in high-shift scenarios. THIS IS THE KEY. The fact that robustness is so disentangled from accuracy (cite theoretical properties again) turns against us because what we want is actually to perform good. The reason is that SO entail a reduction in the effective feature space that the model navigates, which makes it more robust by definition. => Show the simple GO/SO experiment with paired samples. No sampling noise between X' and X", only distribution shift. . Alternative way of thinking about it is forgetting about images. We are comparing different models trained on different instantiations of the features that represent our data. The randomness associated with the sampling process does not present homeocedasticity, as noise magnitude increases with the distance between domains. The more generalizable, the more room for variation. In such case, PA would select the best model when the goal is to make features converge. In some sense it is, but overfitting to examples or subpopulation shifts would yield the same response.
For instance, I try to justify why PA does not always select the best performing model by reformulating the problem so that the inputs are image features rather than images themselves. Every epoch represents an instantiation of the "feature extraction" experiment, and in such case a model overfitting to the shortcut opportunities in the data is also considered robust, since we don't control how much of the robustness "value" is attributed to sampling randomness and how much to

the distribution shift. Sampling randomness can be modelled with the taus, and the key is that it presents heterocedasticity, in the sense that its "spread" is reduced when the model learns (also if it learns the wrong features) and that affects the PA value.
The most "elegant" way I have been able to devise (in the sense that it compiles both a measure of sampling noise and a measure of distribution shift) is comparing the first PCA direction in the feature space between both environments. The spread of each label in the PCA direction accounts for sampling randomness, whereas the MSE between components of the same sample in different environments accounts for distribution shift.

5.3 Model selection on benchmark datasets

In light of the results obtained in the previous sections, we will finally assess the model selection capabilities of PA on benchmark datasets. In particular, several WILDS [27] datasets will be considered, as they provide a comprehensive set of domain generalization tasks that are representative of real-world scenarios. Each of the datasets under consideration entails a specific configuration of learning opportunities that will shift the inductive bias towards suboptimal representations for out-of-distribution generalization. The performance displayed by PA in the previous chapter will be considered when evaluating its behaviour under these conditions.

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

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.

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

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:

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

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 PA as a robustness metric

B.1.1 Empirical behaviour

Results on the empirical behaviour of the PA metric also serve as an exploration of the optimization landscape. As seen in Theorem 3.6.4, the posterior agreement kernel is concave for $\beta \in \mathbb{R}^+$, which implies that the optimization problem has a unique maximum.

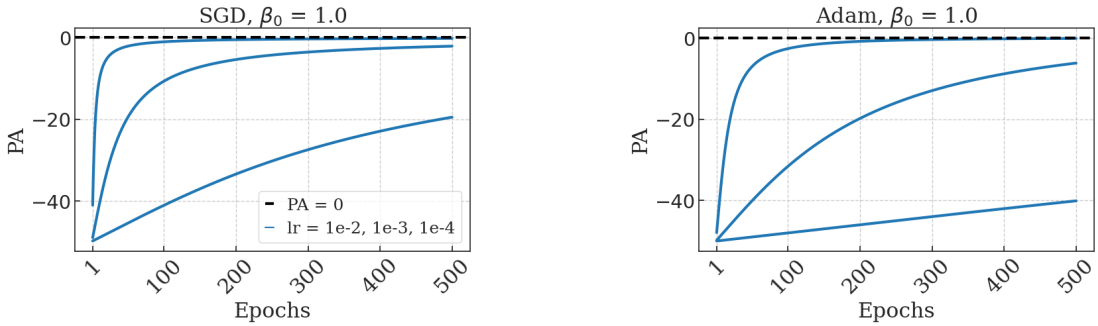


Figure B.1: Evolution of the β optimization for a robust sample.

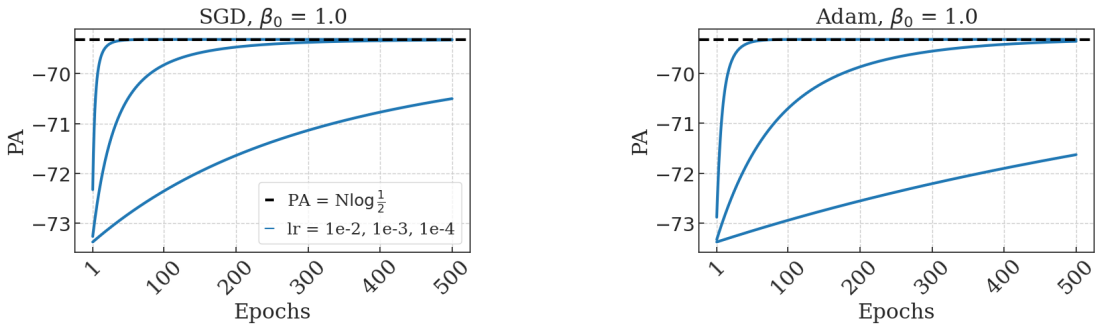


Figure B.2: Evolution of the β optimization for a non-robust sample.

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

β^* 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.

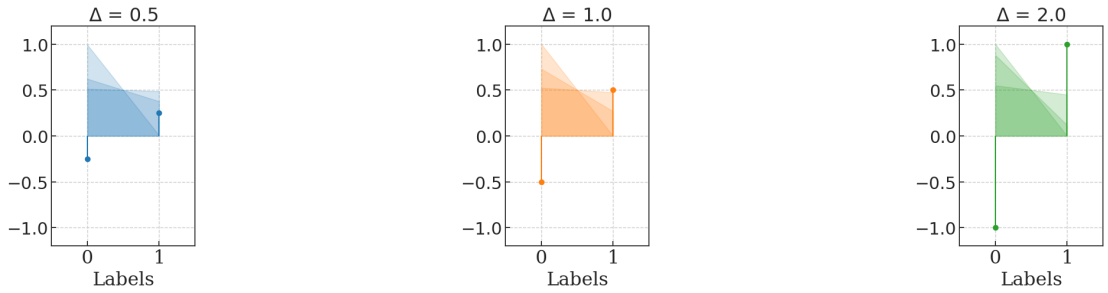


Figure B.3: Logit distributions associated with the behaviour observed in Figure 4.2.

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 β 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.

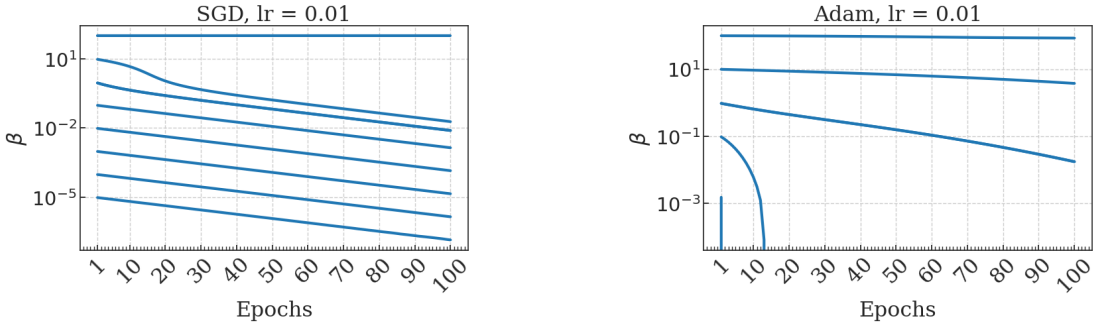


Figure B.4: Evolution of β optimization for different initial values for a non-robust classifier.

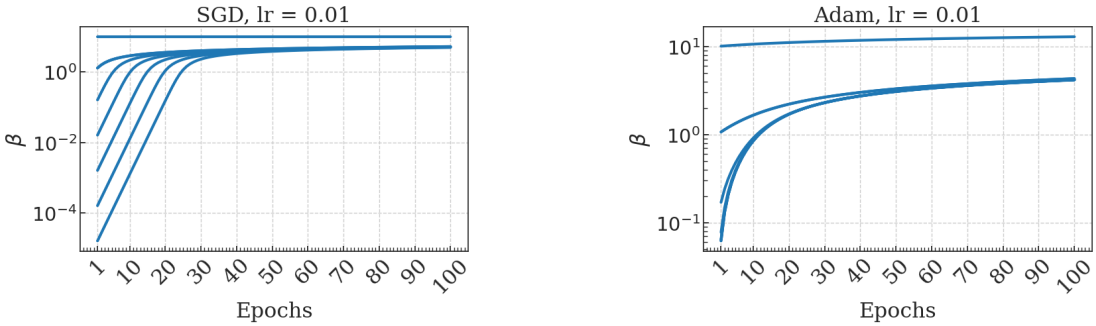


Figure B.5: Evolution of β optimization for different initial values for a robust classifier.

B.2 Adversarial setting

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

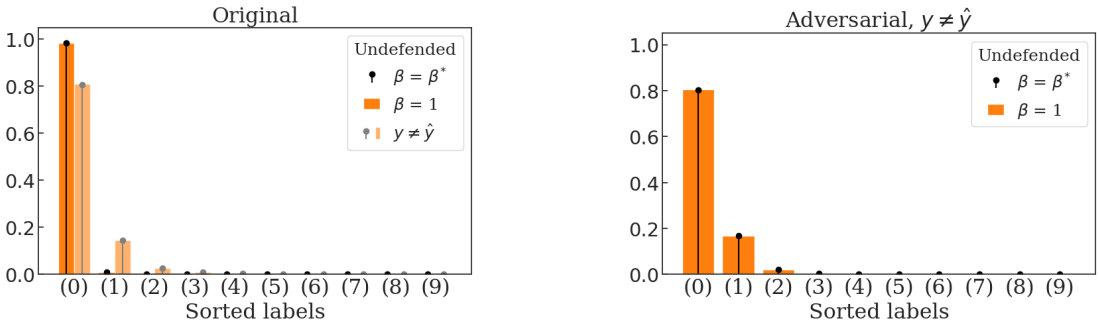


Figure B.6: Average $\mathbf{P}(\hat{y}' \mid \mathbf{x}', \hat{y}'' = \hat{y}' = y)$, $\mathbf{P}(\hat{y}' \mid \mathbf{x}', \hat{y}'' = \hat{y}' \neq y)$ and $\mathbf{P}(\hat{y}'' \mid \mathbf{x}'', \hat{y}'' \neq \hat{y}')$, respectively. **Undefended** model under PGD attack, ℓ_∞=8/255.

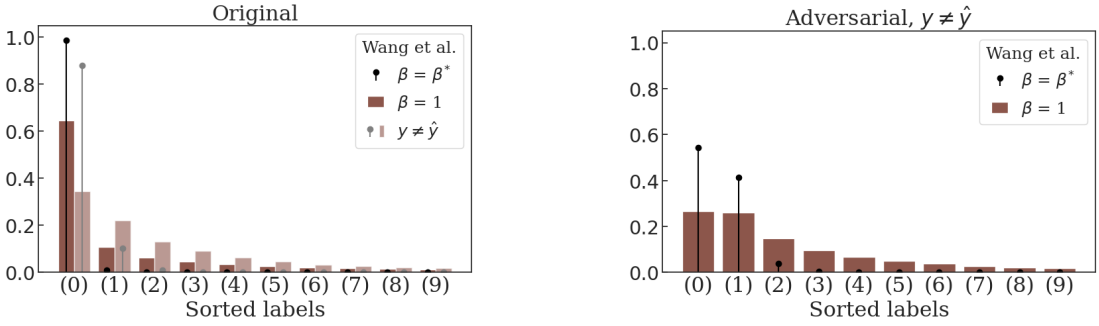


Figure B.7: Average $\mathbf{P}(\hat{y}' \mid \mathbf{x}', \hat{y}'' = \hat{y}' = y)$, $\mathbf{P}(\hat{y}' \mid \mathbf{x}', \hat{y}'' = \hat{y}' \neq y)$ and $\mathbf{P}(\hat{y}'' \mid \mathbf{x}'', \hat{y}'' \neq \hat{y}')$, respectively. **Wang et al.** model under PGD attack, ℓ_∞=8/255.

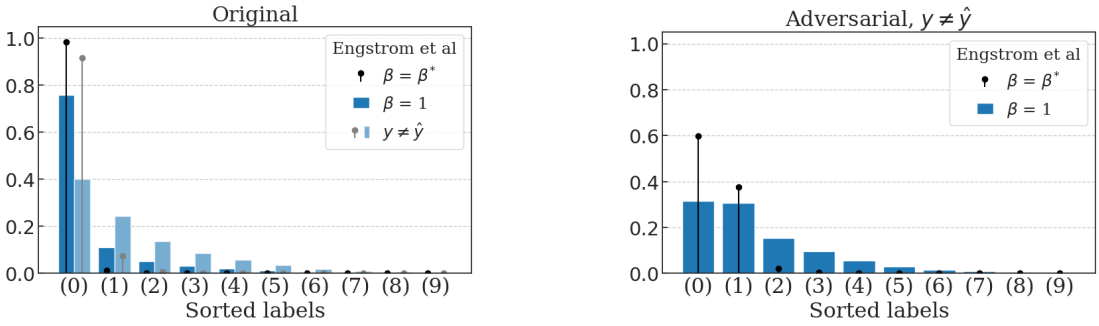


Figure B.8: Average $\mathbf{P}(\hat{y}' \mid \mathbf{x}', \hat{y}'' = \hat{y}' = y)$, $\mathbf{P}(\hat{y}' \mid \mathbf{x}', \hat{y}'' = \hat{y}' \neq y)$ and $\mathbf{P}(\hat{y}'' \mid \mathbf{x}'', \hat{y}'' \neq \hat{y}')$, respectively. **Engstrom et al.** model under PGD attack, ℓ_∞=8/255.

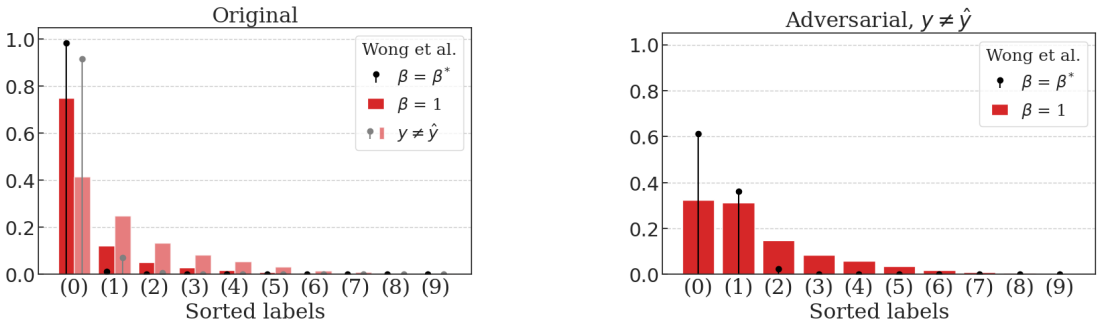


Figure B.9: Average $\mathbf{P}(\hat{y}' \mid \mathbf{x}', \hat{y}'' = \hat{y}' = y)$, $\mathbf{P}(\hat{y}' \mid \mathbf{x}', \hat{y}'' = \hat{y}' \neq y)$ and $\mathbf{P}(\hat{y}'' \mid \mathbf{x}'', \hat{y}'' \neq \hat{y}')$, respectively. **Wong et al.** model under PGD attack, ℓ_∞=8/255.

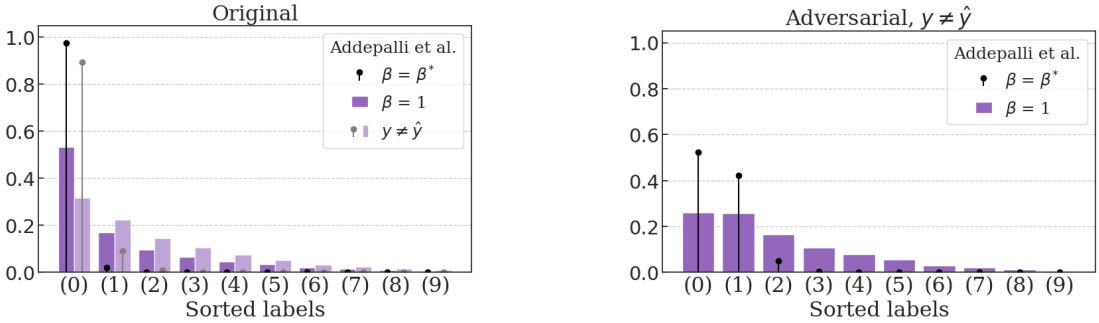


Figure B.10: Average $\mathbf{P}(\hat{y}' \mid \mathbf{x}', \hat{y}'' = \hat{y}' = y)$, $\mathbf{P}(\hat{y}' \mid \mathbf{x}', \hat{y}'' = \hat{y}' \neq y)$ and $\mathbf{P}(\hat{y}'' \mid \mathbf{x}'', \hat{y}'' \neq \hat{y}')$, respectively. **Addepalli et al.** model under PGD attack, ℓ_∞=8/255.

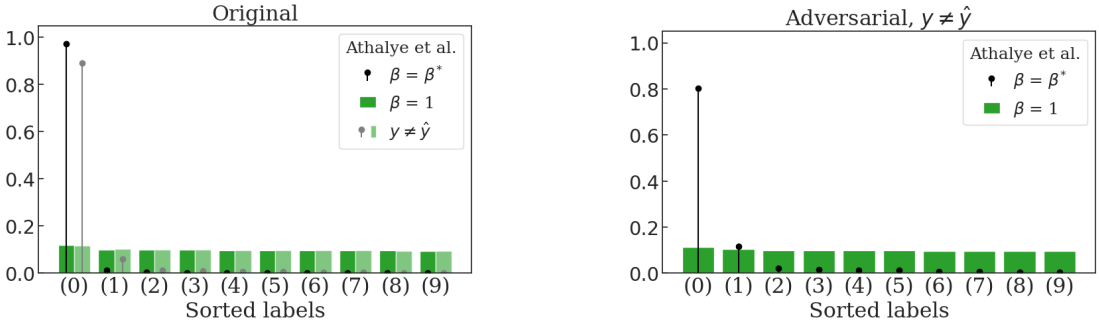


Figure B.11: Average $\mathbf{P}(\hat{y}' \mid \mathbf{x}', \hat{y}'' = \hat{y}' = y)$, $\mathbf{P}(\hat{y}' \mid \mathbf{x}', \hat{y}'' = \hat{y}' \neq y)$ and $\mathbf{P}(\hat{y}'' \mid \mathbf{x}'', \hat{y}'' \neq \hat{y}')$, respectively. **Athalye et al.** model under PGD attack, ℓ_∞=8/255.

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

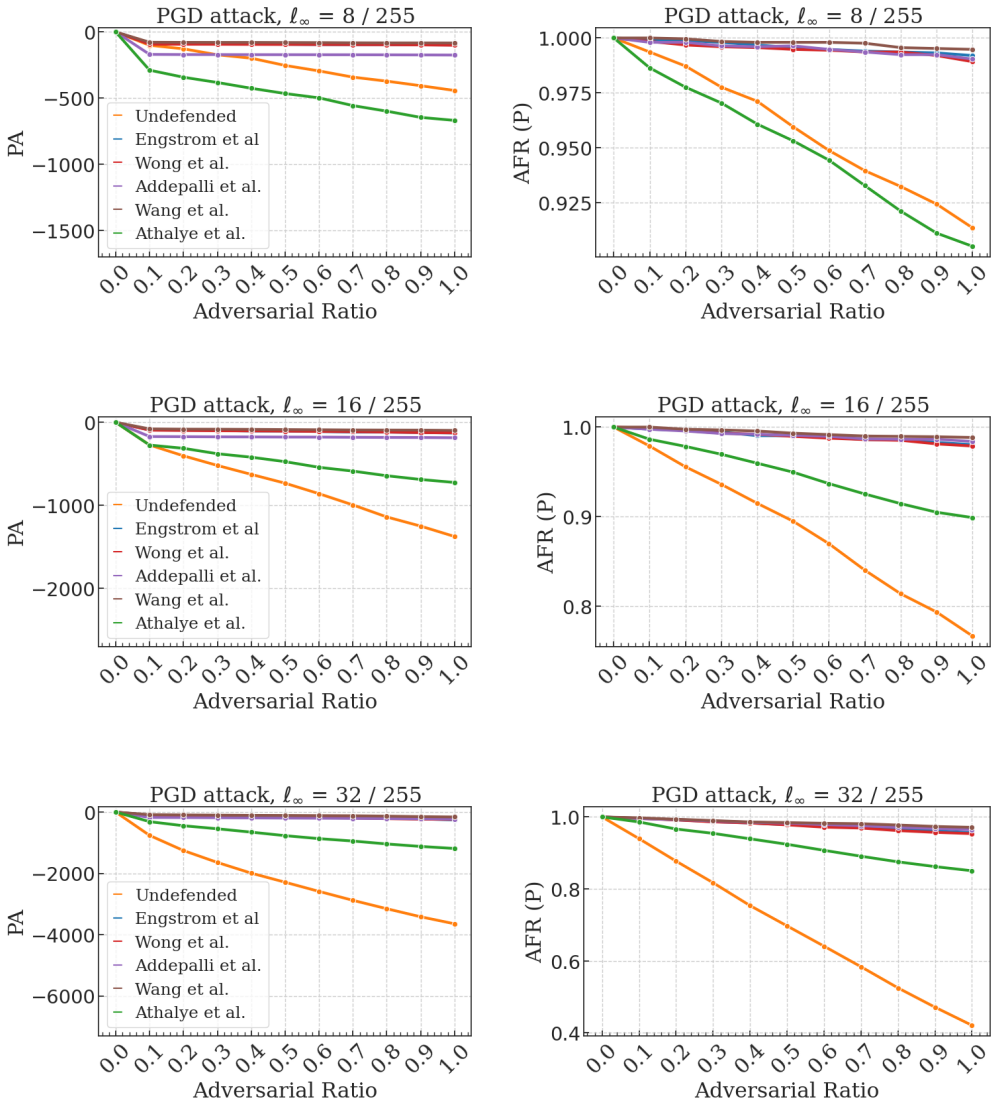


Figure B.12: PA and AFR(P) variation under increasing adversarial ratio at different perturbation norm bounds. The undefended net and several RobustBench robust models are considered against a 1000 step PGD attack.

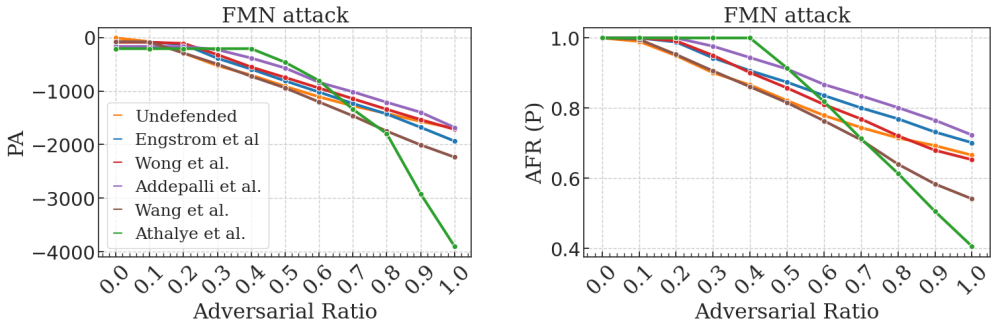


Figure B.13: PA and AFR(P) variation under increasing adversarial ratio. T he undefended net and several RobustBench robust models are considered against a 1000 step FMN attack.

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

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

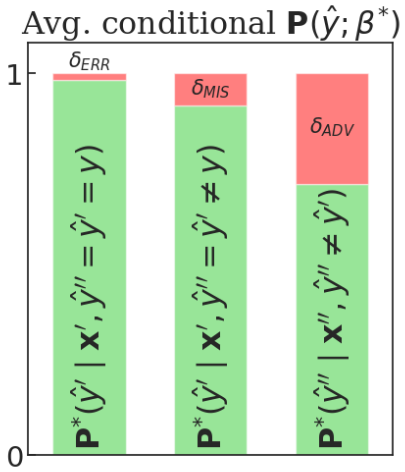


Figure B.14: Illustrative representation of the terms and posterior values constrained considered for the PA approximation.

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

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

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

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

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

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. To avoid notation clutter, we will define:

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

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

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

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

which yields

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

where δ_{ERR} represents the lack of confidence when successfully predicting original samples.

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

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

which yields

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

where δ_{MIS} represents the missing prediction confidence on misclassified original samples.

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

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

which yields

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

given that $\delta_{\text{ERR}} \approx -2\delta_{\text{ERR}}\delta_{\text{ADV}}$. δ_{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 ℓ_∞=8/255 PGD attack computed on the CIFAR10 data.

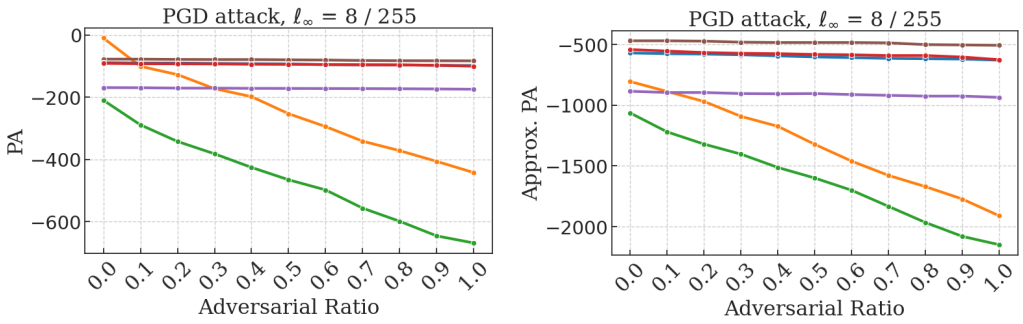


Figure B.15: True and approximated PA values under increasing adversarial ratio for a PGD attack with ℓ_∞=8/255.

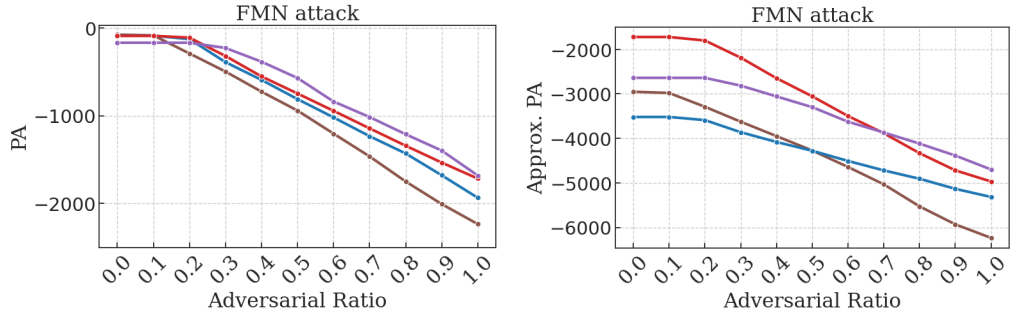


Figure B.16: True and approximated PA values under increasing adversarial ratio for a FMN attack.

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.

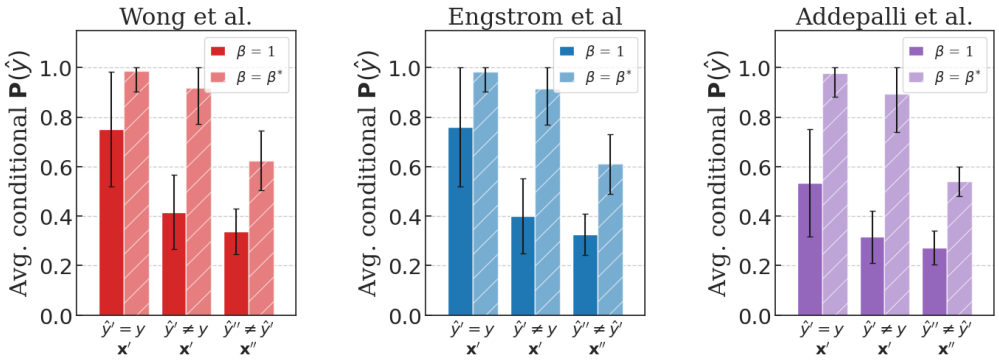


Figure B.17: Average posterior probability of the predicted class for correctly classified original samples, misclassified original samples, and misleading adversarial samples, respectively. Results have been obtained through a PGD attack with ℓ_∞=8/255 and sorted by increasing β^*.

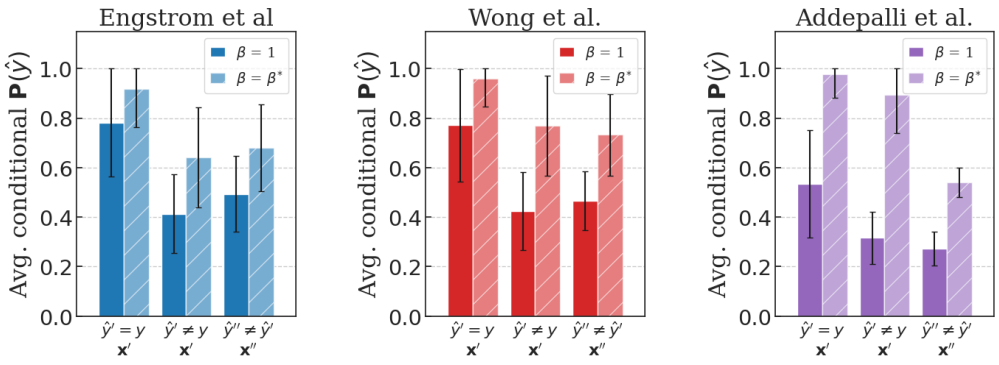


Figure B.18: Average posterior probability of the predicted class for correctly classified original samples, misclassified original samples, and misleading adversarial samples, respectively. Results have been obtained through a FMN attack and sorted by increasing β^*.

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.

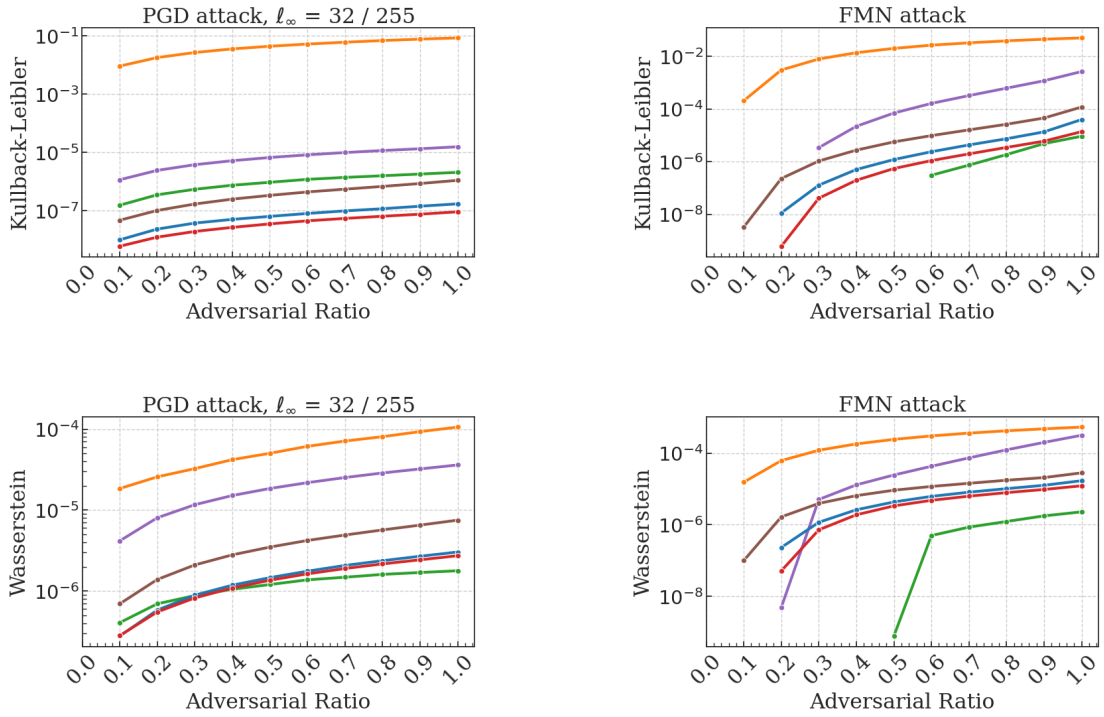


Figure B.19: Comparison of FMN and PGD attacks using probability-based distances.

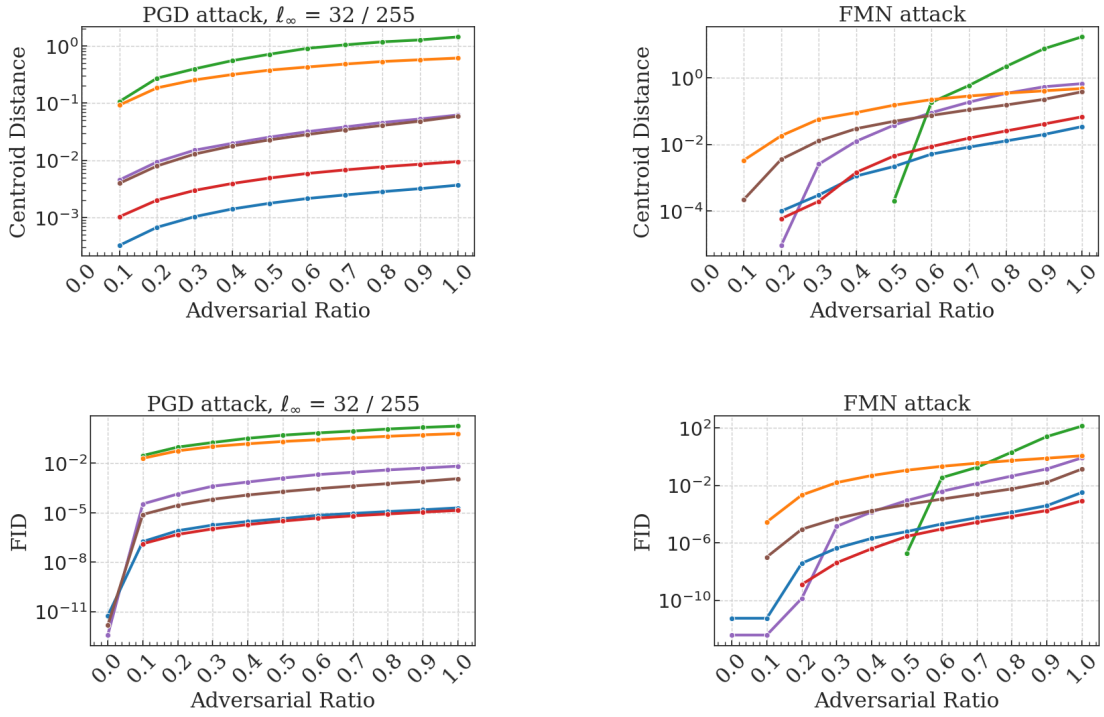


Figure B.20: Comparison of FMN and PGD attacks using feature-space-based distances.

Bibliography

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN.

[3] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent Advances in Adversarial Training for Adversarial Robustness.

[4] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from Several Related Classification Tasks to a New Unlabeled Sample.

[5] Anton Bovier. *Statistical Mechanics of Disordered Systems: A Mathematical Perspective*. Cambridge University Press.

[6] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press.

[7] Joachim M. Buhmann. Data Science Algorithms and the Rate-Distortion Tradeoff.

[8] Joachim M. Buhmann. Information theoretic model validation for clustering.

[9] Joachim M. Buhmann. Posterior Agreement for Model Robustness Assessment in Covariate Shift Scenarios.

[10] Joachim M Buhmann, Morteza Haghir Chehreghani, Mario Frank, and Andreas P Streich. Information Theoretic Model Selection for Pattern Analysis.

[11] Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks.

[12] George Casella and Roger L. Berger. *Statistical Inference*. Wadsworth Group Duxbury, second edition.

[13] Morteza Haghir Chehreghani, Alberto Giovanni Busetto, and Joachim M Buhmann. Information Theoretic Model Validation for Spectral Clustering.

[14] Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified Adversarial Robustness via Randomized Smoothing.

[15] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: A standardized adversarial robustness benchmark.

[16] 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.

[17] 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.

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

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

[20] 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.

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

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

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

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

[25] 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.

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

[27] 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.

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

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

[30] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[50] 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.

[51] 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.

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

[53] 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.

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

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

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

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

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

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

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

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

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

Title of work:

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

Thesis type and date:

Master Thesis, September 2024

Supervision:

Dr. Joao Borges de Sa Carvalho, Dr. Alessandro Torcinovich
Prof. Dr. Joachim M. Buhmann

Student:

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

Statement regarding plagiarism:

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

Declaration of Originality:

`http://www.ethz.ch/faculty/exams/plagiarism/confirmation_en.pdf`

Zurich, 16. 8. 2024: _____