Rethinking Robustness in Machine Learning: A Posterior Agreement Approach

Anonymous authors
Paper under double-blind review

Abstract

The robustness of algorithms against covariate shifts is a fundamental problem with critical implications for the deployment of machine learning algorithms in the real world. Current evaluation methods predominantly match the robustness definition to that of standard generalization, relying on standard metrics like accuracy-based scores, which, while designed for performance assessment, lack a theoretical foundation encompassing their application in estimating robustness to distribution shifts.

In this work, we set the desiderata for a robustness metric, and we propose a novel principled framework for the robustness assessment problem that directly follows the Posterior Agreement (PA) theory of model validation. Specifically, we extend the PA framework to the covariate shift setting by proposing a PA metric for robustness evaluation in supervised classification tasks. We assess the soundness of our metric in controlled environments and through an empirical robustness analysis in two different covariate shift scenarios: adversarial learning and domain generalization. We illustrate the suitability of PA by evaluating several models under different nature and magnitudes of shift, and proportion of affected observations. The results show that the PA metric provides a sensible and consistent analysis of the vulnerabilities in learning algorithms, even in the presence of few perturbed observations.

1 Introduction

Real-world data analysis problems are often formulated as (potentially intractable) combinatorial optimization tasks, *e.g.*, inferring data clusterings, image segmentations, orderings, embeddings, or parameter estimation of dynamical systems, etc. The stochastic nature of the input data and the computational complexity require resorting to approximated, probabilistic estimates. However, such solutions often suffer from instability in generalization to new observations that contain the same signal but systematically different noise perturbations, a phenomenon often referred to as *covariate shift* (Quiñonero-Candela et al., 2008). Covariate shift has gained increasing relevance, especially with the rise of deep learning and its astonishing improvements for a wide variety of predictive tasks. The focus of the research community is progressively "shifting" towards modelings covering different levels of heterogeneity in the data. Consequently, new experimental settings have been implemented in order to test the robustness of machine learning models under nontrivial perturbations of the signal.

The focus of this work is on image classification tasks in two main categories of covariate shift:

- *adversarial (intentional) shift*: the data is crafted ad-hoc by an adversary with vicious intentions, to specifically hinder the output quality of the algorithm (Carlini & Wagner, 2017b; Biggio & Roli, 2018).
- *out-of-distribution (unintentional) shift*: the data is subject to different initial conditions during its measurement (*e.g.*, lighting conditions, orientation, and so on)[1] (Koh et al., 2021; Wang et al., 2023a).

[1]In some cases, out-of-distribution is used to define a shift in the target set (Fang et al., 2022).

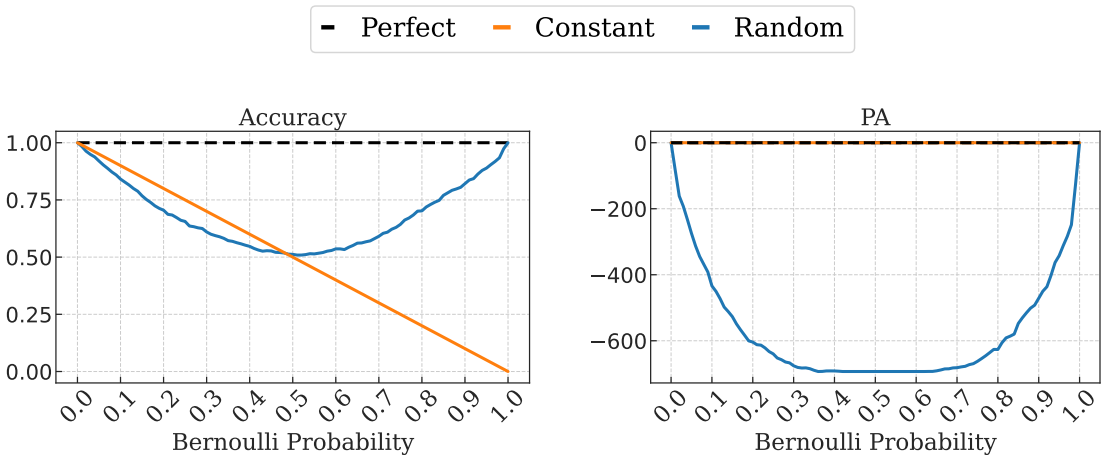


Figure 1: Evolution of accuracy and our proposed metric, Posterior Agreement (PA), displayed by random, constant, and perfect classifiers under a Bernoullian sampling of size 1000. The random classifier sample is generated by permuting the original so that the number of mismatched observations depends on the Bernoulli probability. Accuracy does not comply with the desired properties of a robustness metric and provides an inconsistent assessment that is exclusively driven by task performance, while PA discriminates the robust from the unrobust classifier, in particular, detecting the robustness of a constant classifier.

Under these settings, the usual standard robustness evaluation procedures consist of comparing the accuracy performance of the model under increasing levels of shift (Cinà et al., 2025; Koh et al., 2021). This procedure is essentially the same as that commonly used to evaluate the generalization capabilities of models in the absence of (measurable/accountable) shift, that is, when data is subject to the randomness entailed by the sampling process only. In the following, we argue that robustness assessment requires a paradigm shift in the way it is approached.

First of all, to better characterize the concept of robustness metric, we list two properties that it should possess:

1. *non-increasing*: the metric should be non-increasing with respect to the response of the model against increasing levels of the shift power.
2. *shift-sensitive*: 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 properties described above serve as guiding principles in the robustness evaluation process. However, robustness itself may be susceptible to different interpretations. For example, the concept of *shift power* can be encoded in several ways. In adversarial learning (Madry et al., 2018), it usually coincides with the norm of the difference between a data observation and its perturbed counterpart. On the other hand, in domain generalization (Eulig et al., 2021), shifts usually encode perceptual or semantic alterations that are specific to each dataset and that are hard or even impossible to characterize in a unified way. Furthermore, the shift power may also describe different aspects of the perturbation, such as the number of affected observations, often neglected but equally relevant to time and performance efficiency. In both settings, a principled approach to measuring robustness is required and is currently missing.

In general, Property 1 is satisfied by accuracy-based metrics. However, as accuracy results from thresholding over the predictions, it is unable to encapsulate any information regarding model confidence. In standard validation procedures, devising prediction-confidence-based metrics increases their discriminative power, for instance, when comparing models with similar predictive capabilities but different confidence. In the covariate shift setting, this distinction is decisive, not only for detecting low shift power perturbations but also for

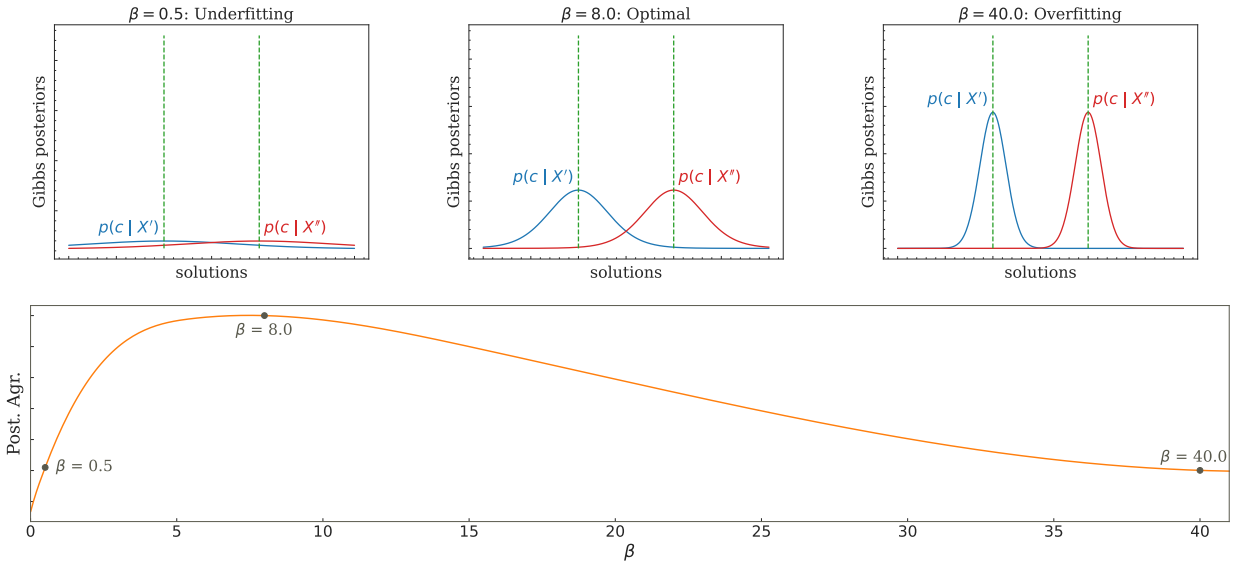


Figure 2: The optimization over the inverse temperature parameter β is required to make the two posteriors insensitive to noise perturbations. In the top figures, the curves represent two posteriors $p(c \mid X')$, $p(c \mid X'')$ in this case over a real-valued parameter c, (*e.g.*, a distribution's mean). When β is too low (top left), the posteriors tend to be uniform and uninformative. When the β is too high (top right) the posteriors are too peaked and therefore sensitive to even small noise perturbation. The optimal β (top center) maximizes the posterior agreement (bottom) guaranteeing informativeness and stability in the distributions that can be then used for model selection or, in our case, for robustness assessment.

providing a more consistent robustness assessment that does not heavily depends on the stochasticity of the dataset.

Additionally, some accuracy metrics involve a comparison with the target variables in the test set (*e.g.*, the attack success rate in adversarial learning), thus not complying with Property 2. For example, a *constant classifier* outputting the same prediction, regardless of the input, is robust by definition, since its output is independent of any shift in the data. Similarly, a *perfect classifier* that always outputs the correct prediction for any input is also robust. An accuracy metric would, however, discriminate the two models as differently robust since their performance would differ (*cf.* Figure 1 left).

In this work, we overcome these weaknesses and propose a new metric complying with the aforementioned properties, from the principled perspective of the Posterior Agreement (PA) framework (Buhmann, 2010; Buhmann et al., 2018; Gronskiy, 2018). PA is an alternative model validation procedure rooted in information-theoretic thinking, in particular the rate-distortion theory (Cover & Thomas, 1999). In this setting, the learning algorithm is thought of as a lossy compression procedure with the aim of resolving the hypothesis class as precisely as possible, given the stochastic data source, a critical requirement for a metric aligned with the two proposed properties.

The proposed PA metric provides a unique and unified framework for robustness assessment in the covariate shift setting, as it relies on a concept of robustness that does not stem from the nature of the data or the shift but instead from the consistency of the probabilistic response of the model. Therefore, PA is a confidence-based metric that does not depend on a model's classification performance, thus aligning with the foundational properties outlined before. As an illustrative example, Figure 1 depicts the difference between accuracy and PA in the robustness assessment of the constant/perfect classifiers described before.

Following epistemologically grounded thinking, we assume two datasets $X', X'' \in \mathcal{X}$, derived from the same experiment, both with ideally identical signal, but different noise realizations. Given a hypothesis set $\mathcal{C}$, the posterior distributions $p(c \mid X')$, $p(c \mid X'')$, for $c \in \mathcal{C}$, should become as similar as possible when adapting the

machine learning algorithms and, as a consequence, the functional form of the posterior $p(\cdot \mid \cdot)$ is optimized for succinctness and stability. Traditionally, posteriors have been designed by defining a cost function and then by calculating or approximating the respective Gibbs distribution. The spread of such a posterior is controlled by a width parameter β, also known as inverse computational temperature. An optimal control of β should render the posterior as much as possible insensitive to noise perturbation (*cf.* Figure 2).

In this article, we first adapt the PA framework to be applied to covariate shift settings without modifying its fundamental principles. In particular, we formulate PA without any requirements of supervised information, making it a suitable metric for robustness evaluation in contexts where the supervision is very scarce or null (*e.g.*, medical applications, or autonomous driving). We then conduct a robustness analysis on two current settings of interest, namely adversarial learning in the evasion attack scenario and domain shift under targeted covariate shifts. In particular, we analyze the performance of several models in terms of PA under different nature and magnitudes of perturbation and proportion of affected observations. Not only does PA succeed in documenting the worsening of model performance in great detail, but it also detects an ongoing shift even when the number of targeted observations is low. Additionally, these findings were extended by leveraging PA for robust model selection with early stopping. In particular, experiments included the perturbation of data with spurious factor co-occurrences. The results show that PA-based selection succeeds at mitigating the spurious correlations existing between shifted environments.

In summary, *(i)* we propose the PA metric, to assess model robustness in covariate shift scenarios and motivate it theoretically, *(ii)* we conduct a robustness analysis on several settings, illustrating the findings emerging from an evaluation with PA, *(iii)* we summarize our findings and discuss a new notion of model robustness *in the PA sense* arising from the presented results.

2 Related Work

Adversarial Learning Adversarial robustness is a fundamental research area in machine learning due to the susceptibility of deep neural networks to small perturbations, leading to erroneous predictions (Szegedy et al., 2014; Goodfellow et al., 2015). A huge plethora of effective and efficient attacks such as FGSM (Goodfellow et al., 2015; Yuan et al., 2019), PGD (Kurakin et al., 2018; Madry et al., 2018), BIM (Kurakin et al., 2018), C&W (Carlini & Wagner, 2017b), DeepFool (Moosavi-Dezfooli et al., 2016), FMN (Pintor et al., 2021), and others (Carlini & Wagner, 2017a; Croce & Hein, 2019; Modas et al., 2019; Zheng et al., 2023), have been developed to evaluate the adversarial robustness of machine learning models. Adversarial examples are often measured in terms of their distance from the original input, using ℓ_p norms (Carlini & Wagner, 2017b). Their use allows for a consistent and well-defined way to measure the magnitude of the perturbation and to compare different adversarial attacks (Carlini & Wagner, 2017b; Cinà et al., 2025). In this work, we also include analyses on the ratio of perturbed observations in a dataset. We believe that this aspect can help in better understanding the behavior of models against attacks, as later shown in the experimental section.

Assessing model robustness proved to be necessary for developing more reliable machine learning systems and for ensuring their resilience *before* deployment. Typically, white-box models are employed to analyze a model's worst-case scenarios and avoid relying on *security by obscurity* (Carlini et al., 2019; Eisenhofer et al., 2021; Däubener et al., 2020). The assessment is commonly performed in the literature by scoring models with accuracy-based metrics such as *attack success rate* and *adversarial accuracy*. While the definition of these two scores is not consolidated, in general, they estimate the model performance over a test set in terms of un/successful attacks. Additionally, Weng et al. (2018) approach robustness evaluation as a Lipschitz constant estimation problem. The method is, however, susceptible to gradient masking (Goodfellow, 2018), causing the metric to overestimate the size of the perturbation needed to fool a model. In addition, the bound is obtained by exploiting the ReLU property of neural networks, therefore narrowing its scope to such models. Wang et al. (2023b) uses, instead, the converging time to an adversarial sample for estimating the robustness of a model. This approach requires the estimation of a Jacobian with respect to the input, which may be infeasible in the case of large-sized data, and is applicable only to differentiable models.

Domain Generalization Domain generalization is a field of paramount importance in machine learning due to both the limitation of acquiring a diverse and large enough set of training samples (Wolpert & Macready, 1997), and the nature of current methods to capture spurious correlations from training data that lead to catastrophic loss of performance in out-of-distribution settings (Wiles et al., 2022; Geirhos et al., 2020). Over the years, several benchmark datasets have been introduced specifically for domain generalization. The PACS dataset (Yu et al., 2022) represents shifts in visual styles between real photos and artistic media like paintings, cartoons, and sketches, reflecting variations in colour, texture, and abstractness. VLCS (Torralba & Efros, 2011) captures object-centric and scene-centric distribution shifts, with variations in object variability, scene context, and annotation styles across other previously established datasets. The WILDS benchmark (Koh et al., 2021) embodies real-world shifts, covering temporal and geographical changes, as well as differences in image resolution and quality, which are typical in practical applications. Each benchmark encapsulates unique distribution shifts, providing diverse challenges for evaluating domain generalization methods. More recently, the DiagViB-6 dataset (Eulig et al., 2021) has also been proposed for the systematic analysis of data distribution shifts over multiple generative factors, covering hue, position, lightness, scale, and texture, and allowing for accumulation of multiple shifts. While several new benchmarks have been proposed to address growing requirements for both more realistic and challenging evaluation settings, accuracy on the unseen domains remains the fundamental metric for estimating the model robustness.

3 Methodology

In this section, we devise a tractable version of the posterior agreement for the validation of classification models in covariate shift settings. Similar to Buhmann et al. (2018), we start from the work of Rose (1998) for deterministic annealing, and adapt it in the context of PA to a classification task.

3.1 Setting

Let $D_X = (x_i)_{i=1}^N$ be a dataset[2] of i.i.d. measurements (observations) drawn from a random variable X with support $\mathcal{X}$. A K-class *classifier* can be defined as the composition of two functions:

- a function $F : \mathcal{X} \to \mathbb{R}^K$, mapping observations to a vector of *discriminant functions*, *i.e.*, $x \mapsto F(x) = (F_1(x), \dots, F_K(x))$. The discriminant functions are parametrized by a set of parameters θ, which is fixed, as we are conducting an evaluation.
- a *decision rule* $f : \mathbb{R}^K \to \{1, \dots, K\}$, applied to choose the class for each observation, based on the discriminant functions. Usually, f is the *Maximum A Posteriori (MAP)* rule:

$$f(F(x)) = \arg\max_j F_j(x) \tag{1}$$

A classifier is defined as $c = f \circ F$.

Note that, since F is fixed, the hypothesis set is spanned by all possible decision rules over $\mathcal{X}$. Restricting the process to D_X, gives rise to a finite hypothesis set[3] $\mathcal{C}$, containing all possible mappings from D_X to $\{1, \dots, K\}$. Therefore, $|\mathcal{C}| = K^N$.

Each classifier is associated with a score evaluating its *confidence* in explaining the data:

$$R(c, X; \theta) = \sum_{i=1}^{N} F_{c(x_i)}(x_i; \theta). \tag{2}$$

Such score can be used to define a posterior $p(c \mid X)$ over the hypothesis set $\mathcal{C}$. In particular, we aim to find a probability distribution p such that $R(c_1, X; \theta) > R(c_2, X; \theta) \iff p(c_1 \mid X) > p(c_2 \mid X), \forall c_1, c_2 \in \mathcal{C}$,

[2]Note that, we do not need to specify the targets of D_X for assessing the robustness of a model.

[3]The finiteness of $\mathcal{C}$ is a mathematical expedient to efficiently deal with the subsequent derivations. (*cf.* the discussion section).

a property that we will rewrite as $\phi_{\mathcal{C}}(p)$. As many distributions fulfill this requirement, we narrow the possibilities by applying the *Maximum Entropy Principle (MEP)* (Jaynes, 1957), which states to search for the maximally uninformative (*i.e.*, entropy-maximizing) distribution which best encodes the observed data.

The following optimization objective

$$\underset{p(c \mid X)}{\text{maximize}} \quad H[p] \tag{3a}$$

$$\text{subject to} \quad \phi_{\mathcal{C}}(p), \tag{3b}$$

$$p(c \mid X) \geq 0, \tag{3c}$$

$$\sum_{c \in \mathcal{C}} p(c \mid X) = 1, \tag{3d}$$

$$\mathbb{E}_{C|X}[R(C, X)] = \mu \tag{3e}$$

gives the required solution. Here, μ is a hyperparameter ensuring that the expectation is finite. We set the Lagrangian without the inequality constraints

$$\mathcal{L}(p, \alpha, \beta) = H[p] \tag{4}$$

$$+ \alpha \left(1 - \sum_{c \in \mathcal{C}} p(c \mid X)\right) \tag{5}$$

$$+ \beta(\mathbb{E}_{C \sim p}[R(C, X) - \mu]), \tag{6}$$

and verify that they hold for the found solution. The derivative with respect to $p(c)$ is

$$\frac{\partial \mathcal{L}}{\partial p(c \mid X)} = -1 - \log p(c \mid X) - \alpha + \beta R(c, X). \tag{7}$$

Equating it to 0 and solving for $p(c)$ gives

$$p(c \mid X) = \frac{\exp(\beta R(c, X))}{\exp(1 + \alpha)}. \tag{8}$$

Setting $\exp(1 + \alpha) = \sum_{c \in \mathcal{C}} \exp(\beta R(c, X))$ and $\beta \geq 0$ ensures that $p(c \mid X)$ is a Gibbs distribution that satisfies the constraints 3b and 3c with the inverse temperature parameter β. The exact value of β depends on μ, and can be found by enforcing the posterior agreement principle, discussed in the next section.

The summation over the hypothesis set can pose a drawback in the computation of the posterior. The following result provides an efficient factorization for $p(c \mid X)$.

Theorem 1.

$$p(c \mid X) = \prod_{i=1}^{N} p_i(c(x_i) \mid X), \tag{9}$$

where

$$p_i(k \mid X) = \frac{\exp(\beta F_k(x_i))}{\sum_{j=1}^{K} \exp(\beta F_j(x_i))} \tag{10}$$

is the probability that x_i is assigned to class k.

Proof. cf. Appendix A.1 □

The energy term $F_k(x_i) \in \mathbb{R}$ quantifies the degree of membership of observation i to a class k (the higher, the more probable). $F_k(x_i)$ can be, for example, the logit output of a multi-class prediction model (*e.g.*, a neural network).

3.2 Posterior Agreement

We are given two i.i.d. datasets X', X'' with $|X'| = |X''| = N$ and for $i = \{1, \dots, N\}$ x'_i and x''_i are two realizations sampled from the same ideal (*i.e.*, noiseless) process $x_i^{(0)}$. Given a classification model, we can estimate its agreement in the prediction results through the *expected posterior agreement kernel*

$$k = \mathbb{E}_{X',X''}\left[\log\left(\sum_{c\in\mathcal{C}} \frac{p(c \mid X')p(c \mid X'')}{p(c)}\right)\right], \tag{11}$$

where the expectation is computed over the random variables of X' and X''. A notable difference with the previous versions of the kernel is the addition of the term $p(c)$. It is a prior over $\mathcal{C}$, used to account for the model complexity. In the following, we will work under the assumption that we do not have access to any prior information, that is $p(c) = 1/|\mathcal{C}|$, $\forall c \in \mathcal{C}$. Often, the computation of the expected value over X' and X'' is infeasible, therefore the *empirical posterior agreement kernel* is adopted as

$$k(X', X'') = \log\left(\sum_{c\in\mathcal{C}} \frac{p(c \mid X')p(c \mid X'')}{p(c)}\right), \tag{12}$$

which estimates the overlap between the Gibbs posteriors defined as explained above.

The computation depends on the inverse temperature hyperparameter $\beta \in \mathbb{R}_{\geq 0}$. Similar to the MEP, we search for the β that maximizes the overlap between the posteriors. The *Posterior Agreement (PA)* metric is then computed as

$$\mathrm{PA}(X', X'') = \underset{\beta}{\text{maximize}} \quad \frac{1}{N}k(X', X''). \tag{13a}$$

$$\text{subject to} \quad \beta \geq 0 \tag{13b}$$

$1/N$ is a scaling correction factor since PA increases proportionally with the size of the dataset.

In the following, we propose and discuss an operative formula to compute the empirical posterior agreement kernel.

Theorem 2. *With no prior information available, the empirical posterior agreement kernel $k(X', X'')$, can be rewritten as:*

$$k(X', X'') = \log\left(|\mathcal{C}| \prod_{i=1}^{N} \sum_{j=1}^{K} p_i(j \mid X')p_i(j \mid X'')\right). \tag{14}$$

Proof. cf. Appendix A.2. □

By optimizing Program 13 with this kernel, we can see the effect of β in determining the optimal overlap between posteriors. An increment in β corresponds in peaking the distributions $p_i(\cdot \mid X)$, $X = X', X''$ toward their MAP, awarding *matching distributions* that share the same MAP. On the contrary, decreasing β flattens the distributions, and mitigates penalizations due to *mismatching distributions* with different MAPs. Therefore, the optimization operates a tradeoff between these opposite sub-objectives, searching for the best solution that explains the agreement between the two data sources.

We conclude by characterizing the posterior agreement metric as computed with the kernel of Theorem 2.

Theorem 3. *Under no prior information available, PA has the following properties.*

1. *Boundedness:* $0 \leq PA(X', X'') \leq \log K$.
2. *Symmetry:* $PA(X', X'') = PA(X'', X')$.
3. *Concavity:* $PA(X', X'')$ *is a concave function in* $\beta < +\infty$.

Proof. cf. Appendix A.3. □

4 Experimental Results

In this section, we present a comprehensive analysis of the empirical results. In particular, we study two covariate shift scenarios, adversarial learning and domain generalization, and set up a comparison of the scores obtained by different models, using PA and accuracy-based metrics. Our main purpose is to highlight the differences between performance- and robustness-based evaluation criteria. Therefore, we compare several defences and learning techniques with different capabilities, to encompass, as much as possible, the variety of cases that can arise during an evaluation process. For further technical information, the reader is referred to our code implementation[4].

For measuring model robustness, we use the PA metric with no prior information. In particular, we use the original test set as X' and its perturbed version as X''. For a better visualization of the results, we adopt a logarithmic scale, and we omit the $|\mathcal{C}|$ and the $1/N$ constants in PA to increase its dynamic range, therefore $\text{PA} \in [-N\log(K), 0]$.

4.1 Adversarial Robustness

For the adversarial robustness scenarios, we carry out our experiments with the CIFAR-10 (Krizhevsky et al., 2009) dataset, widely adopted in the machine learning security literature as a benchmark for robustness evaluation. The dataset contains 60 000 colour images of 32×32 pixels equally distributed in 10 classes. The analyzed models are trained on the training set (50 000 images), and the PA evaluation is performed on the test set (10 000 images). In particular, we perturb the test set with two different evasion attack methods: Projected Gradient Descent (PGD) (Madry et al., 2018) and Fast Minimum Norm (FMN) (Pintor et al., 2021). The max attack power is specified in terms of the ℓ_∞ norm and is set in advance ($\ell_\infty = [8/255, 16/255, 32/255]$) for the PGD attack, while it is automatically searched for each observation with the FMN attack. Both attacks are run for 1000 steps on the test set to generate its perturbed version for subsequent PA computation.

We consider several models with different robustness properties. In particular, we score an undefended WideResNet-28-10 (Zagoruyko & Komodakis, 2016) and five defended models: a ResNet18 (He et al., 2016a) defended by Addepalli et al. (2022), a ResNet50 (He et al., 2016a) defended by Engstrom et al. (2019), a WideResNet-28-10 defended by Wang et al. (2023c), a PreActResNet18 (He et al., 2016b) defended by Wong et al. (2020) and a 3-layer CNN (LeCun et al., 2015) defended by a JPEG pre-processing Das et al. (2017) and a Reverse Sigmoid post-processing Lee et al. (2018). All models except for the latter are implemented with the RobustBench library (Croce et al., 2021).

We perform a comparison of PA with an accuracy-based metric, the Attack Failure Rate ($\text{AFR}_T = 1 - \text{ASR}_T$) and evaluate the performance of the models under investigation to verify the effectiveness of the PGD and FMN attacks. The evaluation is performed over ten different adversarial ratios (ARs) of attacked examples in the test set, ordered by increasing attack power. In particular, a dataset with $AR = p$ contains the attacked observations under the $10p$-decile, ordered by ℓ_∞, and the remaining unattacked ones. $AR = 1$ corresponds to an entirely attacked dataset. The β parameter is searched with an Adam (Kingma & Ba, 2015) optimization procedure, run for 500 epochs.

PGD Attack As illustrated in Figure 3 (left), PA consistently discriminates the undefended model, which significantly decreases its performance with increasing adversarial ratio and attack power. As expected, the rate at which the performance decreases is faster the more powerful the attack is. The JPEG + RS model also displays poor robustness scores, compared to the other models, especially for low-norm attacks, which is consistent with the JPEG compression defence employed.

The fundamental difference between PA and AFR_T arises in the scoring of the undefended model, which overperforms according to AFR_T, especially for low ℓ_∞. To understand the causes of such a difference, in Figure 4, we display the distribution in prediction confidence of three example models. In particular, we compare the correctly predicted observations before the attack ($p(\hat{y}' \mid X', \hat{y}' = y)$, full colour) with

[4]Metric: https://anonymous.4open.science/r/pa-metric/README.md
Experiments: https://anonymous.4open.science/r/pa-covariate-shift/README.md

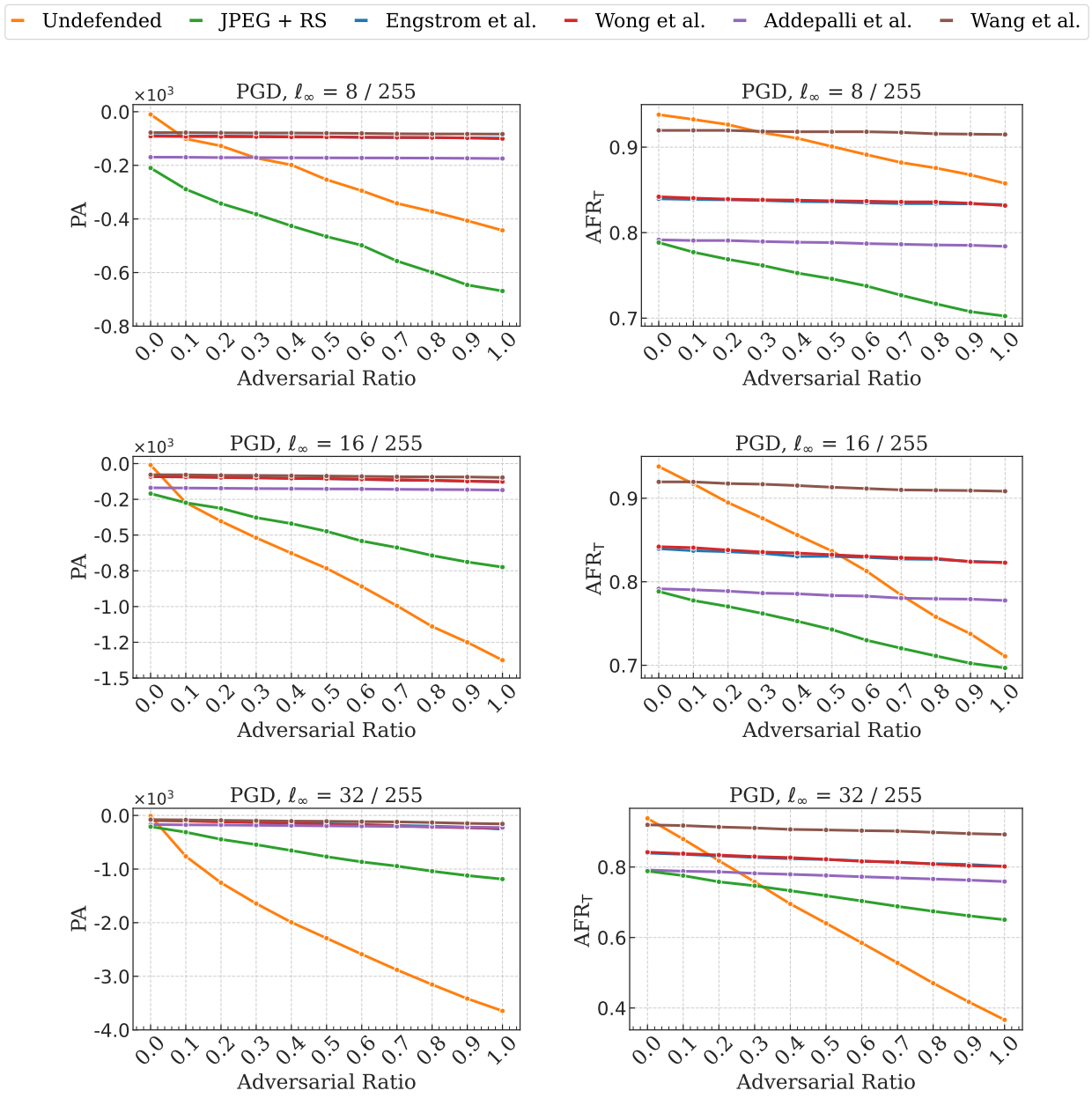


Figure 3: PA (left) and AFR$_T$ (right) scores against increasing AR and ℓ_∞, for the PGD attack. The tendencies are similar with the only exception of the undefended model, which is overperforming according to AFR$_T$.

the successfully attacked ones ($p(\hat{y}'' \mid X'', \hat{y}'' \neq \hat{y}')$, streaked colour). We look at two weak models (the undefended and JPEG + RS) and the most robust one (Wang et al., 2023c), in the case of $AR = 1$ (*cf.* Appendix B for an analysis of all models). Standard errors (black bars) are also included. Additionally, we include the values of the optimized β parameters after convergence for all the models. We analyze the behaviour of the three PGD plots:

- The undefended model outputs, on average, high-confidence predictions. The standard errors overlap, hinting that the model cannot detect that an attack has happened, therefore not lowering enough

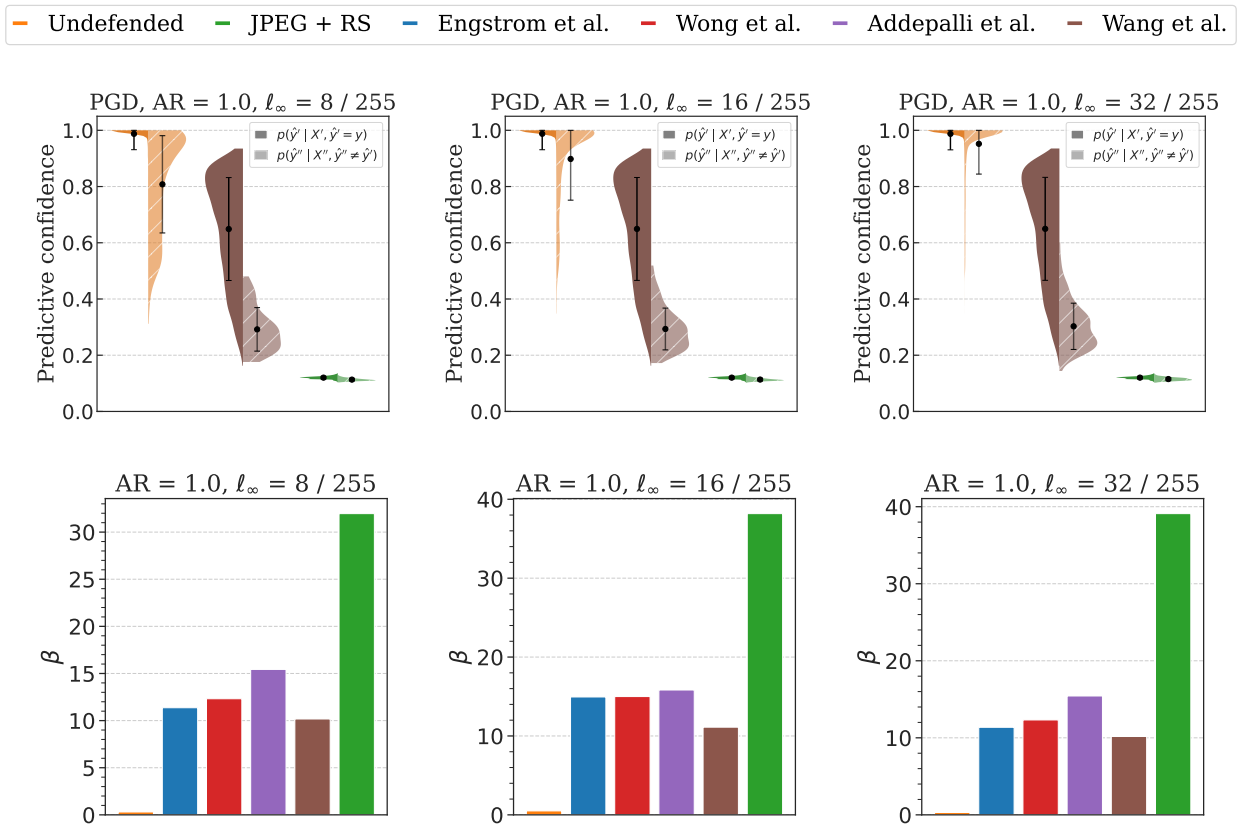


Figure 4: (top) Distribution of the predictive confidence for three example models. The robust one (Wang et al., 2023c) is significantly decreasing its confidence, while the weak ones (undefended and JPEG + RS) are not. (bottom) Final β values for $AR = 1$. Anomalous β values identify the weak models.

its confidence. AFR_T scores are high overall. The maximization of PA entails flat posteriors, to mitigate the effect of the mismatches between the pre- and post-attack distributions. Therefore, β converges to a low value. Consequently, we obtain low PA scores overall.

- The JPEG + RS model provides low confidence predictions, due to the reverse sigmoid defense used, and does not reduce its predictive confidence in the presence of misleading adversarial examples. As a result, PA is maximized when posterior distributions are highly informative, to peak the matches (*i.e.*, β converges to a high value). However, mismatching predictions induce a penalty on the robustness score that drives PA to a low value, overall. AFR_T scores are low, as well.

- Wang et al. (2023c) is sensible to the attack and significantly lowers its overall confidence, as indicated by non-overlapping standard errors. Both AFR_T and PA are high, overall.

In conclusion, while both metrics report similar trends and rankings, AFR tends to overestimate the robustness of the undefended model, especially when a few number of samples have been attacked, while PA provides a more consistent assessment, unbiased by the nature of the experiment (*e.g.*, ℓ_∞ and AR).

Additionally, note that for $AR = 0$, PA scores do not converge to the theoretical result of 0. In the absence of shift, the optimal β tends toward infinity to peak all coinciding distributions over their MAPs (*cf.* proof of Theorem 3, Property 2). However, the Adam optimizer progressively scales down the learning rate, dampening, in turn, the convergence rate. The obtained PA scores then reflect the "peakedness" of the initial model predictions, that is, they provide information about the predictive confidence of each model

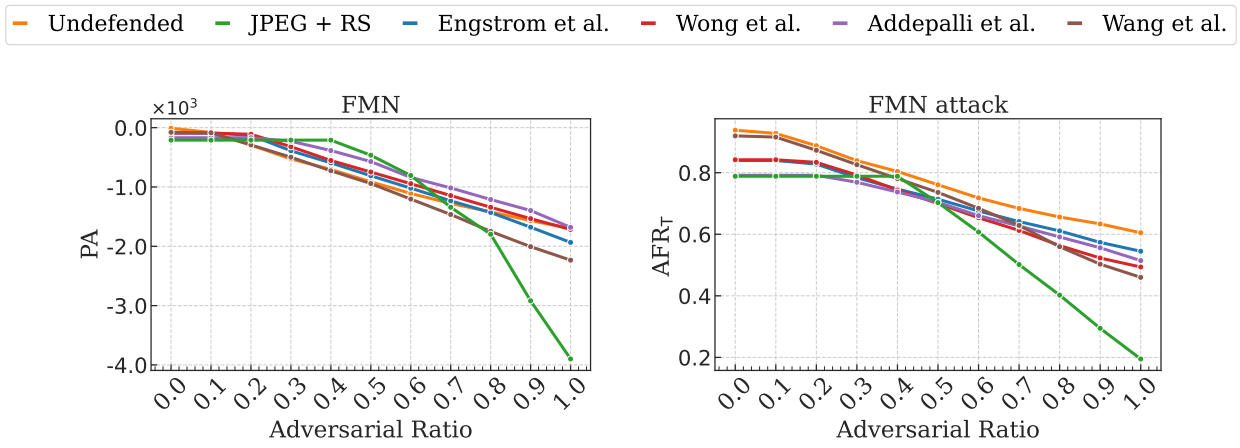


Figure 5: PA (left) and AFR_T (right) scores against increasing AR and ℓ_∞, for the FMN attack. Again, the undefended model robustness is overestimated, according to AFR_T. For $AR \in [0.3, 0.6]$ JPEG + RS model is more robust than the others, with a similar trend in performance.

before the attack, with the most confident models scoring better. The ranking of PA coincides with that of AFR_T, hinting that in the absence of shift, the most confident models are also the more accurate[5].

FMN Attack FMN does not require an ℓ_∞ limit, and automatically finds, for each observation, the minimum distance needed to evade the classification. We found that for $AR \gtrsim 0.5$, the observations are perturbed far over the shift power against which all models are robustified. We, therefore, focus on the cases with $AR \leq 0.5$, and report, for completeness, the extended results in Appendix B. In Figure 5, the PA and AFR_T scores are presented. Again, the undefended model robustness is overestimated by AFR_T. For some ARs, the JPEG + RS model performs better than the other models overall, showing that the JPEG compression defensive mechanism effectively filters out the perturbations and reduces the attack transferability performance of FMN. In Figure 6 (right), we can note the increment in the corresponding β, indicating that the two distributions contain more MAP matches. Wang et al. (2023c) results instead in being the weakest model and, therefore, less suitable against this attack. Accordingly, its β is low. For completeness, in Appendix B, we include the predictive confidence distributions of all models for $AR \in \{0.5, 1\}$.

4.2 Domain Generalization

Following the analysis on adversarial robustness, we perform a series of experiments on several domain generalization settings to assess both model discriminability and model selection when comparing PA to accuracy-based metrics. Again, AFR_T is taken as the comparison reference. In addition, Adam is run for 1000 epochs to search for the β parameter.

In this scenario, we conduct our experiments through a modified version of the DiagViB-6 dataset (Eulig et al., 2021) that comprises distorted and upsampled coloured images of size 128×128 from the MNIST dataset (LeCun, 1998). Each specific image factor is controllable, leading to changes in texture, hue, lightness, position, and scale, and observations paired to assemble X' and X''. Furthermore, we control the specific selected samples, differently from the common setting. In particular, we establish two unique domains for the training data $\{e_0^1, e_0^2\}$ under a pair of original instantiations of each image factor, where all factors from e_0^1 are different from e_0^2. We induce a second pair of domains for the validation data, following precisely the factor configuration of the training domains. Finally, a set of six additional domains, $\{e_0, e_1, \ldots, e_5\}$, is generated, with each e_i corresponding to images where i factors have been modified with respect to e_0^1. Our final dataset comprises two sets of 40 000 images for training, two sets of 20 000 images for validation, and

[5]However, this is not always necessarily true. We want to remark that task performance and predictive confidence are two distinct aspects (*cf.* Appendix C).

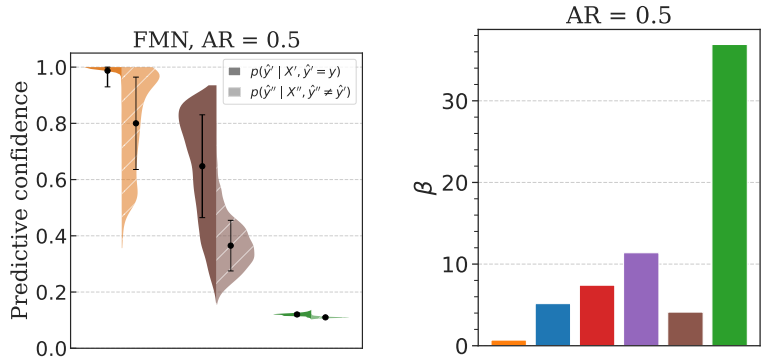


Figure 6: Predictive confidence distributions (left) and β plots (right) for the FMN attack. The results are related to $AR = 0.5$ not to include perturbations against which the models are not robust. The models perform similarly to the PGD case, with a less marked difference in the standard error for Wang et al. (2023c) model.

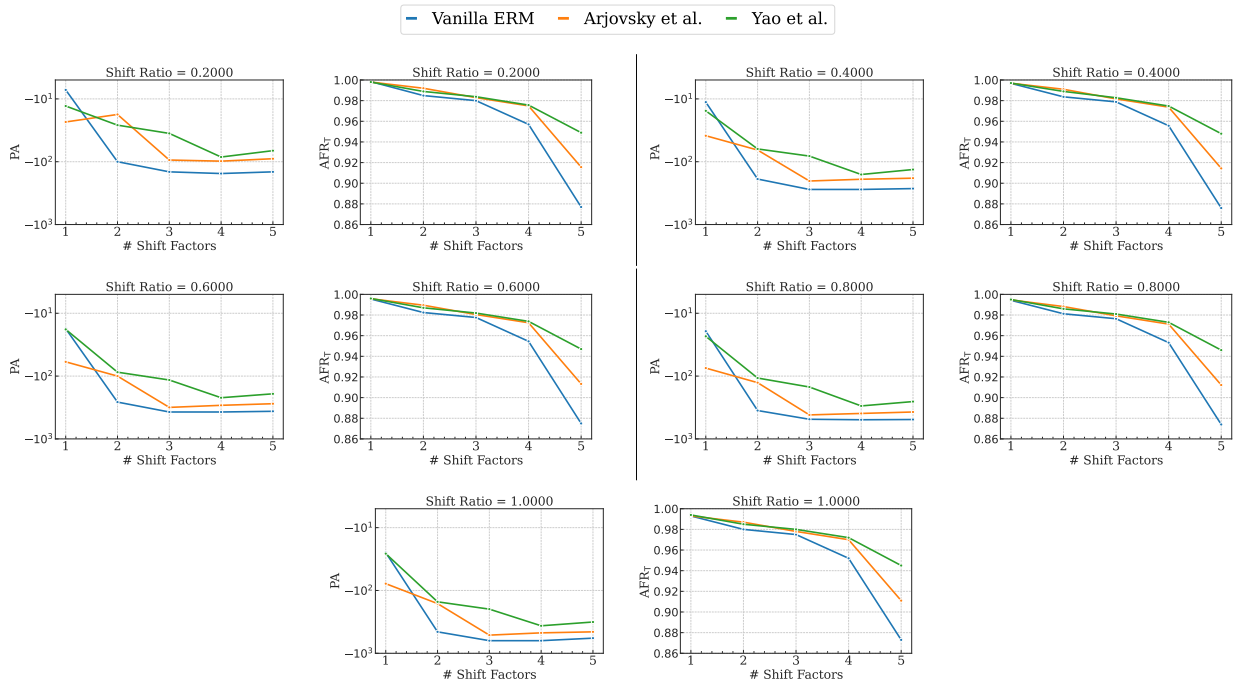


Figure 7: Effect of domain shift in the PA performance on a weak and robust model, over five distinct levels of the shift ratio. Each plot depicts the response of both models to cumulative levels of distribution shift, from one to five shifted factors. At each shift ratio, PA is able to differentiate between the two models.

six sets of 10 000 images for testing. Note that the original MNIST images used for generating the datasets do not overlap. For more details on how the dataset was adjusted to our setting, refer to Appendix E.

For the experiment, we consider a ResNet50 model trained with three different algorithms: vanilla Empirical Risk Minimization (ERM), Invariant Risk Minimization (IRM) (Arjovsky et al., 2019) and Learning Invariant Predictors with Selective Augmentation (LISA) (Yao et al., 2022).

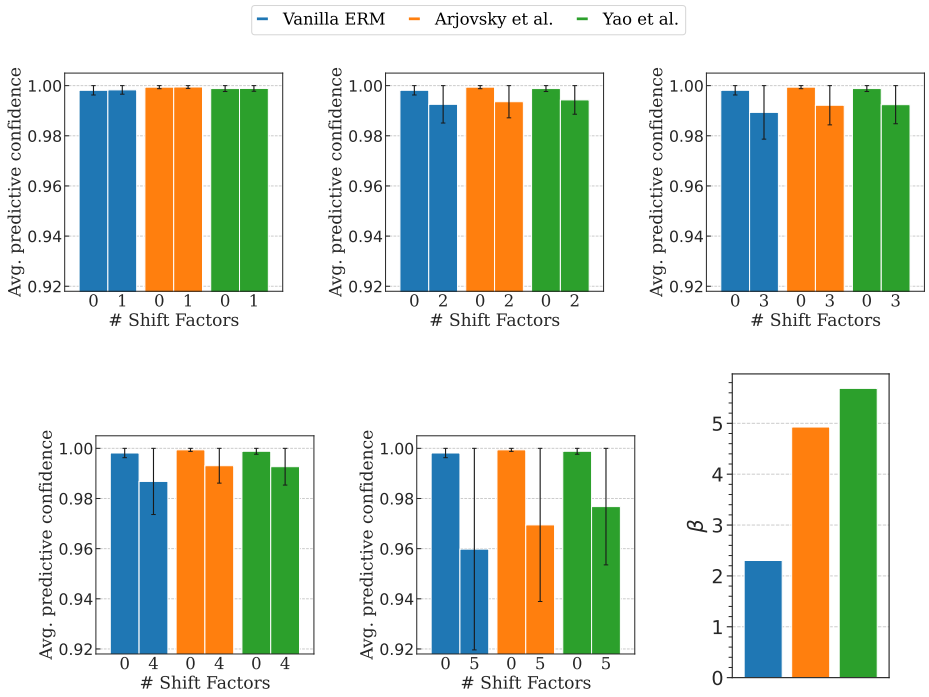


Figure 8: Average predictive confidence for three models—(blue) Vanilla ERM, (orange) IRM, and (green) LISA, under a $SF = 1$. Each panel shows results at a different number of shift factors ($\#SF$), illustrating how the models' predictive confidence varies as the distribution shift increases. The left column remains fixed, and the right column changes across $\#SF = 1, 2, 3, 4, 5$. The final bar plot on the right depicts the β for 5 shift factors. We see that higher model confidence leads to an adjusted higher β.

Model discriminability In Figure 7, we showcase the PA evaluation results over cumulative distribution shifts. Each plot presents the variation of PA over five cumulative implementations of the shift factors (SF), in five distinct test sets corresponding to increasing amounts of shift ratio (SR). In particular, we subject a subset of the test set to the related cumulative distribution shift to an SR corresponding to 20%, 40%, 60%, 80%, and 100% of the original dataset. Similarly to AR, $SR = 1$ entails a change of all the samples in the dataset.

As depicted, and already well established in Koh et al. (2021), the overall trend of robustness starts with ERM as the weakest model, followed by IRM, and then by LISA as the most robust. However, PA is also able to identify more nuanced robustness behaviors. In particular, when $SF = 1$, ERM achieves a higher PA score because it relies on domain-specific features already encountered during training, even when the number of mismatching predictions is higher. PA is the only metric that was able to discriminate ERM. Conversely, the domain invariance imposed by IRM slightly hinders PA under these mild shifts, even though its overall AFR remains comparable or better.

This subtle source-versus-target domain difference clearly illustrates why PA is more suitable than AFR_T for evaluating robustness: the former captures how models respond to domain shifts that partially overlap with training distributions, whereas the latter might mask these fine-grained effects. Consequently, IRM's robust design trades off some PA in low-shift regimes but does not sacrifice performance, distinguishing its behavior from both ERM and LISA in a way that PA can clearly reveal.

As shown in Figure 8, predictive confidence varies across models under distribution shifts. ERM starts with high confidence but declines as the number of shift factors increases, while IRM remains stable, and LISA consistently exhibits the highest confidence, which suggests superior robustness. Higher confidence models require a greater β for posterior alignment, as seen in the rightmost plot. Conversely, lower confidence models

benefit from reduced β to flatten distributions and minimize mismatches. This effect is most pronounced with five shift factors, where LISA maintains strong confidence while ERM struggles. This behavior aligns with adversarial settings, where models with lower confidence require a higher β to align their distributions. However, this is not strictly a matter of distribution alignment but rather *mutual information* maximization: such models only push β to higher values if their predictions agree. PA demonstrates superior performance because it effectively navigates the trade-off between model performance and informativeness. In particular, models with lower confiPA effectively captures these dynamics, and through optimal selection of β can enhance robustness under shifting distributions.

We also attempted to better understand the impact of controlled domain shifts and their impact on robustness assessment through PA (*cf.* Appendix D for more details). A phenomenon observed in this work is the decrease in classification performance as the number of shift factors increases, while the PA metric remains comparable or slightly improves. Specifically, at $SF = 5$, the feature space differs significantly (as shown in Figure 11) from those at $SF \leq 4$, leading to lower accuracy yet roughly similar (or slightly improved) robustness. This observation highlights that in more complex settings, where the effects of the covariate shift are directly controlled (as in the case of DiagViB-6), the robustness of the model does not stem solely from how out-of-distribution the samples are. Instead, as noted by Geirhos et al. (2020), it is influenced by compounded learning phenomena related to which features are captured during training. Consequently, future research should account for these interactions when designing evaluation protocols. For example, one possibility is to test the model on *all* possible combinations of shift factors, thereby more comprehensively revealing how different combinations and sequences of shifts affect robustness.

Model Selection While our metric has been primarily intended for evaluation, in this section, we investigate whether PA can also effectively guide *model selection*. To that end, we now examine a setting where the inductive bias is deliberately "poisoned" by shortcut opportunities (SO), *i.e.*, spurious correlations between the predicted factor F_P and the learning factor F_L. As illustrated in Figure 12 (Appendix E), we systematically control these co-occurrences to create zero or partial Generalization Opportunities (GO). Under Zero-GO (ZGO), each instance of F_P is exclusively paired with one instance of F_L, strongly encouraging models to overfit to the shortcut. Partial GOs (1/2/3-CGO) break some of these exclusive pairings, while Zero-SO (ZSO) allows all factor combinations. For more details on the experiment setup, refer to Appendix E.

In contrast to the previous experiments, we exclude LISA here because its data-augmentation/interpolation strategies would undermine the carefully constructed SO/GO configurations as the algorithm augments the data based on the same image transformations present in this experiment. We therefore restrict our study to ERM and IRM, trained on source environments. Additionally, following the model-selection findings from earlier sections, we conduct in-distribution validation: the validation datasets match the same shift configuration as the training datasets, ensuring that model selection is not confounded by unseen shifts. The results in Table 1 show that *robustness-driven* model selection (based on PA) substantially improves the test performance, in particular, in settings with partial GOs (*i.e.*, 1-CGO or 3-CGO). In these cases, shortcut learning is partially mitigated, so identifying models that best maintain consistent predictions across slightly varied conditions (rather than just maximizing raw accuracy) proves crucial. Notably, IRM tends to benefit more from these robustness-based criteria, converging to models that can resist spurious correlations and better exploit the available GOs. Overall, this approach demonstrates how a focus on robust performance through PA, rather than pure in-distribution accuracy can significantly enhance generalization.

5 Discussion

The PA metric is an attempt at assessing robustness under an epistemologically grounded approach. Such an assessment is agnostic of the underlying data generation process and even the nature of the data itself, therefore it is versatile and can be applied to explain diverse scenarios with a single theory, as illustrated in the experimental study. Additionally, PA is general in its scope, as it only requires the (probabilistic) outputs of the model. Our proposed version requires only the logits of a model[6] to optimize a single parameter β,

[6]N.B. The model itself does not need to be differentiable, as we do not optimize its parameters.

	Test 1		Test 2		Test 3		Test 4		Test 5	
	Acc.	**ΔAcc.**	**Acc.**	**ΔAcc.**	**Acc.**	**ΔAcc.**	**Acc.**	**ΔAcc.**	**Acc.**	**ΔAcc.**
ERM										
ZGO	53.2	±0.01	54.6	±0.01	55.7	±0.01	66.7	±0.01	66.6	±0.01
1-CGO	62.9	+9.5	64.7	+10.2	60.8	+0.3	62.9	+2.2	64.2	+0.5
2-CGO	69.1	+9.4	71.2	+7.8	71.9	+2.2	76.2	-2.4	77.0	-2.8
3-CGO	73.1	+16.6	85.6	+3.6	70.1	+9.7	71.4	+6.4	72.1	+6.7
ZSO	99.6	±0.01	92.8	-0.1	89.9	±0.01	89.9	+0.2	85.9	±0.01
IRM										
ZGO	50.1	+5.9	50.5	+4.9	52.8	+9.5	64.4	+1.1	69.4	+1.2
1-CGO	63.0	+7.0	65.9	+7.6	59.4	+2.2	59.0	+1.8	59.0	+1.8
2-CGO	69.0	+10.6	69.7	+10.0	65.8	+4.7	77.0	+13.0	65.1	+12.6
3-CGO	79.5	+11.6	83.0	+9.8	73.6	+10.9	79.5	+11.0	72.2	+11.3
ZSO	99.4	+0.1	93.4	+1.3	89.2	+0.2	87.0	+1.6	87.0	+1.6

Table 1: Test performance under increasing levels of shift for models selected through different configurations of factor co-occurrence in the hue-based learning factor experiment. Specifically, the performance of models selected through validation accuracy (Acc) and the difference between accuracy-based and PA-based selection (ΔAcc) are reported. PA is able to select models that perform better than the accuracy-selected model in most cases.

however, variants are possible (*cf.* Buhmann (2010); Buhmann et al. (2018)). By relying on the MEP, PA provides an estimate of model's robustness which is as neutral as possible with respect to missing information, while sticking to the available observations. In addition to that, its Bayesian perspective, allows us to give a robustness estimation of the entire hypothesis set and not just the specific model.

The experimental results show that PA evaluates models at a finer level, providing higher discriminability and consistency in covariate shift settings than accuracy-based metrics. The discriminative power of AFR_T relies, instead, solely on test performance and does not fulfill the desired properties of a robustness metric. Additionally, by relying on predictive confidence rather than preditions, PA favors classifiers that exhibit desirable inductive biases, and grasps sensitivity to the nature of the shift at an even greater detail than in standard analyses.

Our proposed version of PA has, however, limitations. The theoretical framework has been developed by considering only the case of a finite hypothesis set. While this does not pose a severe drawback, and the experimental results attest to the validity of this approach, a characterization of PA for continuous hypothesis sets would be much more general in its scope. Additionally, we have assumed a uniform probability $p(c)$, to obtain a more tractable formula for the metric, ignoring the complexity of the hypothesis set. Adding this information would improve the robustness estimation process, allowing for more fine-grained discrimination between models. In future, we plan to extend some of the presented theoretical results to more general kernels.

6 Conclusions

In this work, we have set some desiderata for designing a metric measuring robustness against forms of covariate shift such as adversarial noise and domain generalization. Following theoretically grounded thinking, we proposed PA, a metric derived from the Posterior Agreement framework. We have conducted a comparison against commonly used accuracy metrics, presenting analyses based on novel aspects of the shift. Our study shows that PA is consistent with the desired robustness properties. Additionally, by estimating the overlap between posteriors over the hypothesis space, PA provides a more fine-grained evaluation and a better discriminative power than hard-counting metrics, such as accuracy. In conclusion, our work lays the basis for a more sound model robustness assessment, *in the PA sense*.

References

Sravanti Addepalli, Samyak Jain, Gaurang Sriramanan, and R. Venkatesh Babu. Scaling adversarial training to large perturbation bounds. In *European Conference on Computer Vision*, 2022.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2019.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2154–2156, 2018.

Joachim M. Buhmann. Information theoretic model validation for clustering. In *IEEE International Symposium on Information Theory, ISIT 2010, June 13-18, 2010, Austin, Texas, USA, Proceedings*, pp. 1398–1402. IEEE, 2010.

Joachim M. Buhmann, Julien Dumazert, Alexey Gronskiy, and Wojciech Szpankowski. Posterior agreement for large parameter-rich optimization problems. *Theor. Comput. Sci.*, 745:1–22, 2018.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec)*, 2017a.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (S&P)*, 2017b.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness, 2019.

Antonio Emanuele Cinà, Jérôme Rony, Maura Pintor, Luca Demetrio, Ambra Demontis, Battista Biggio, Ismail Ben Ayed, and Fabio Roli. Attackbench: Evaluating gradient-based attacks for adversarial examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 1999.

Francesco Croce and Matthias Hein. A randomized gradient-free attack on relu networks. In *Pattern Recognition*, 2019.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks*, 2021.

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. *CoRR*, abs/1705.02900, 2017.

Sina Däubener, Lea Schönherr, Asja Fischer, and Dorothea Kolossa. Detecting adversarial examples for speech recognition via uncertainty quantification. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2020.

Thorsten Eisenhofer, Lea Schönherr, Joel Frank, Lars Speckemeier, Dorothea Kolossa, and Thorsten Holz. Dompteur: Taming audio adversarial examples. In *USENIX Security Symposium*, 2021.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations, 2019.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021.

Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Ian J. Goodfellow. Gradient masking causes CLEVER to overestimate adversarial perturbation size. *CoRR*, abs/1804.07870, 2018.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Alexey Gronskiy. *Statistical Mechanics and Information Theory in Approximate Robust Inference*. PhD thesis, ETH Zurich, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, pp. 770–778. IEEE Computer Society, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision, ECCV*, volume 9908 of *Lecture Notes in Computer Science*, pp. 630–645. Springer, 2016b.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

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 Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018.

Yann LeCun. The mnist database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*, 1998.

Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. Deep learning. *Nat.*, 521(7553):436–444, 2015.

Taesung Lee, Benjamin Edwards, Ian M. Molloy, and Dong Su. Defending against model stealing attacks using deceptive perturbations. *CoRR*, abs/1806.00054, 2018.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the Association for Computational Linguistics, ACL*, pp. 142–150. The Association for Computer Linguistics, 2011.

A Madry, A Makelov, L Schmidt, D Tsipras, and A Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Sparsefool: a few pixels make a big difference. In *Conference on computer vision and pattern recognition (CVPR)*, 2019.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Conference on computer vision and pattern recognition (CVPR)*, 2016.

Maura Pintor, Fabio Roli, Wieland Brendel, and Battista Biggio. Fast minimum-norm adversarial attacks through adaptive norm constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 20052–20062, 2021.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 12 2008.

Kenneth Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proc. IEEE*, 86(11):2210–2239, 1998.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528, 2011.

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. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8052–8072, 2023a. doi: 10.1109/TKDE.2022.3178128.

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. *ACM Trans. Multim. Comput. Commun. Appl.*, 19(5s):172:1–172:17, 2023b.

Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning, ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 36246–36263. PMLR, 2023c.

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. In *International Conference on Learning Representations, ICLR*, 2018.

Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dj Dvijotham, and Ali Taylan Cemgil. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations*, 2022.

David H. Wolpert and William G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.

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

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pp. 25407–25437. PMLR, 2022.

Samuel Yu, Peter Wu, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Pacs: A dataset for physical audiovisual commonsense reasoning. In *European Conference on Computer Vision*, 2022.

Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*. BMVA Press, 2016.

Yang Zheng, Luca Demetrio, Antonio Emanuele Cinà, Xiaoyi Feng, Zhaoqiang Xia, Xiaoyue Jiang, Ambra Demontis, Battista Biggio, and Fabio Roli. Hardening rgb-d object recognition systems against adversarial patch attacks. *Inf. Sci.*, 651:119701, 2023.

A Proofs of theorems

A.1 Proof of Theorem 1

We first require the following lemma. For ease of reading, we slightly abuse the notation and consider c to be a mapping from the object indices i instead of the measurements x_i, that is $c : \{1, \dots, N\} \to \{1, \dots, K\}$.

Lemma 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,*

$$\sum_{c \in \mathcal{C}} \prod_{i=1}^{N} \mathcal{E}_{i,c(i)} = \prod_{i=1}^{N} \sum_{j=1}^{K} \mathcal{E}_{ij} \tag{15}$$

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

$$\sum_{c \in \mathcal{C}} \prod_{i \leq N} \mathcal{E}_{i,c(i)} = \sum_{c \in \mathcal{C}} \mathcal{E}_{1,c(1)} = \sum_{j \leq K} \mathcal{E}_{1,j} = \prod_{i \leq N} \sum_{j \leq K} \mathcal{E}_{i,j}. \tag{16}$$

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

$$\begin{aligned}
&\sum_{c \in \mathcal{C}} \prod_{i \leq N+1} \mathcal{E}_{i,c(i)} = \\
&= \sum_{(c_1,\dots,c_{N+1}) \in \{1,\dots,K\}^{N+1}} \prod_{i \leq N+1} \mathcal{E}_{i,c_i} \\
&= \sum_{\substack{(c_1,\dots,c_N) \in \{1,\dots,K\}^N \\ c_{N+1} \leq K}} \prod_{i \leq N+1} \mathcal{E}_{i,c_i} \\
&= \sum_{(c_1,\dots,c_N) \in \{1,\dots,K\}^N} \sum_{c_{N+1} \leq K} \prod_{i \leq N+1} \mathcal{E}_{i,c_i} \\
&= \sum_{(c_1,\dots,c_N) \in \{1,\dots,K\}^N} \sum_{c_{N+1} \leq K} \left(\mathcal{E}_{N+1,c(N+1)} \prod_{i \leq N} \mathcal{E}_{i,c_i} \right) \\
&= \left(\sum_{c_{N+1} \leq K} \mathcal{E}_{N+1,c(N+1)} \right) \sum_{(c_1,\dots,c_N) \in \{1,\dots,K\}^N} \prod_{i \leq N} \mathcal{E}_{i,c_i} \\
&= \left(\sum_{c_{N+1} \leq K} \mathcal{E}_{N+1,c(N+1)} \right) \prod_{i \leq N} \sum_{j \leq K} \mathcal{E}_{i,j} \\
&= \left(\sum_{j \leq K} \mathcal{E}_{N+1,j} \right) \prod_{i \leq N} \sum_{j \leq K} \mathcal{E}_{i,j} \\
&= \prod_{i \leq N+1} \sum_{j \leq K} \mathcal{E}_{i,j}.
\end{aligned} \tag{17}$$

□

We are ready to prove the theorem

Theorem 1

$$p(c \mid X) = \prod_{i=1}^{N} p(c(x_i) \mid X), \tag{18}$$

where

$$p(k \mid X) = \frac{\exp(\beta F_k(x_i))}{\sum_{j=1}^{K} \exp(\beta F_j(x_i))} \tag{19}$$

is the probability that x_i is assigned to class k.

Proof. The Gibbs distribution is

$$p(c \mid X) = \frac{\exp\left(\beta \sum_{i \leq N} F_{c(x_i)}(x_i)\right)}{\sum_{c \in \mathcal{C}} \exp\left(\beta \sum_{i \leq N} F_{c(x_i)}(x_i)\right)}. \tag{20}$$

The numerator can be rewritten as follows:

$$\exp\left(\beta \sum_{i \leq N} F_{c(x_i)}(x_i)\right) = \prod_{i \leq N} \exp\left(\beta F_{c(x_i)}(x_i)\right). \tag{21}$$

We now apply Lemma 1:

$$\sum_{c \in \mathcal{C}} \prod_{i \leq N} \exp\left(\beta F_{c(x_i)}(x_i)\right) = \prod_{i \leq N} \sum_{k \leq K} \exp\left(\beta F_k(x_i)\right) \tag{22}$$

Putting these results together yields that

$$\begin{aligned} p(c \mid \theta, X) &= \frac{\prod_{i \leq N} \exp\left(\beta F_{c(x_i)}(x_i)\right)}{\prod_{i \leq N} \sum_{k \leq K} \exp\left(\beta F_k(x_i)\right)} && (23) \\ &= \prod_{i \leq N} \frac{\exp\left(\beta F_{c(x_i)}(x_i)\right)}{\sum_{k \leq K} \exp\left(\beta F_k(x_i)\right)} && (24) \\ &= \prod_{i \leq N} p(c(x_i) \mid X). && (25) \end{aligned}$$

□

A.2 Proof of Theorem 2

Theorem 2 *With no prior information available, the empirical posterior agreement kernel $k(X', X'')$, can be rewritten as:*

$$k(X', X'') = \log\left(|C| \prod_{i=1}^{N} \sum_{j=1}^{K} p_i(j \mid X') p_i(j \mid X'')\right). \tag{26}$$

Proof.

$$\begin{aligned} k(X', X'') &= \log\left(\sum_{c \in \mathcal{C}} \frac{p(c \mid X') p(c \mid X'')}{\pi(c)}\right) \\ &= \log\left(|\mathcal{C}| \sum_{c \in \mathcal{C}} \prod_{i=1}^{N} p_i(c(x_i) \mid X') \prod_{i=1}^{N} p_i(c(x_i) \mid X'')\right) \\ &= \log\left(|\mathcal{C}| \sum_{c \in \mathcal{C}} \prod_{i=1}^{N} p_i(c(x_i) \mid X') p_i(c(x_i) \mid X'')\right) \\ &= \log\left(|\mathcal{C}| \prod_{i=1}^{N} \sum_{k=1}^{K} p_i(k \mid X') p_i(k \mid X'')\right). && (27) \end{aligned}$$

The last step is obtained by applying Lemma 1. □

A.3 Proof of Theorem 3

Theorem 3 *Under no prior information available, the following properties hold for the empirical posterior agreement kernel:*

1. *Boundedness:* $0 \leq PA(X', X'') \leq \log K$.
2. *Symmetry:* $PA(X', X'') = PA(X'', X')$.
3. *Concavity:* $PA(X', X'')$ *is a concave function in* $\beta < +\infty$.

Proof.

Property 1 (Boundedness) When all predictions match, $\beta \longrightarrow +\infty$ and posteriors converge to Kronecker deltas centred at their respective MAPs, $\hat{y}'_i$ and $\hat{y}''_i$, coinciding for each $i = 1, \ldots, N$:

$$\begin{aligned}\mathrm{PA}(X', X''; \beta^*) &= \frac{1}{N} \log \left(|C| \prod_{i=1}^{N} \sum_{j=1}^{K} \delta_{j\hat{y}'_i} \delta_{j\hat{y}''_i} \right) = \frac{1}{N} \log \left(|C| \prod_{i=1}^{N} 1 \right) = \frac{1}{N} \left(\log(|C|) + \sum_{i=1}^{N} \log(1) \right) \\ &= \frac{\log(|C|)}{N} = \frac{\log(K^N)}{N} = \log(K).\end{aligned}$$

When none of the predictions matches, $\beta \longrightarrow 0$ and posteriors converge to a uniform distribution:

$$\begin{aligned}\mathrm{PA}(X', X''; \beta^*) &= \frac{1}{N} \log \left(|C| \prod_{i=1}^{N} \sum_{j=1}^{K} \frac{1}{K} \cdot \frac{1}{K} \right) = \frac{1}{N} \log \left(|C| \prod_{i=1}^{N} \frac{1}{K} \right) = \frac{1}{N} \left(\log(|C|) + \sum_{i=1}^{N} \log(K^{-1}) \right) \\ &= \frac{\log(|C|)}{N} - \log(K) = \log(K) - \log(K) = 0.\end{aligned}$$

Property 2 (Symmetry) Trivial, as it follows from the commutativity of the product.

Property 3 (Concavity) First, note that

$$k\left(X', X''\right) = \log \left(|C| \prod_{i=1}^{N} \sum_{j=1}^{K} p_i\left(j \mid X'\right) p_i\left(j \mid X''\right) \right) \propto \sum_{i=1}^{N} \log \left(\sum_{j=1}^{K} p_i\left(j \mid X'\right) p_i\left(j \mid X''\right) \right), \tag{28}$$

where the posteriors $p_i(j \mid X)$, $X \in \{X', X''\}$ are the Gibbs distributions over classes for each observation[7]:

$$p_i(j \mid X) = \frac{\exp(\beta R(j, x_i))}{\sum_{k=1}^{K} \exp(\beta R(k, x_k))}. \tag{29}$$

Since the sum of concave functions is concave, we will focus on proving only the concavity of the log term. In particular, we will show that

$$\lambda(\beta) = -\log \left(\sum_{j=1}^{K} p(j \mid X') p(j \mid X'') \right) \tag{30}$$

is convex.

[7]Note that we do not specify the form of the cost function, therefore the theorem can be applied with any cost function $R(j, x_i)$.

Let us define $R(j, x') = R'_j$ and $\exp(\beta R(j, x')) = \exp(\beta R'_j) = e'_j$, with $x' \in X'$. Similar notation is used for $x'' \in X''$. $f(\beta)$ can be therefore rewritten as

$$\lambda(\beta) = -\log\left(\frac{\sum_{j=1}^{K} e'_j e''_j}{\sum_{k=1}^{K} e'_k \sum_{p=1}^{K} e''_p}\right) = -\log\left(\sum_{j=1}^{K} e'_j e''_j\right) + \log\left(\sum_{k=1}^{K} e'_k \sum_{p=1}^{K} e''_p\right) = -\lambda_1(\beta) + \lambda_2(\beta). \tag{31}$$

First, let us focus on the first term. In particular,

$$\frac{d}{d\beta}\lambda_1(\beta) = \frac{\sum_{j=1}^{K}(R'_j + R''_j)e'_j e''_j}{\sum_{k=1}^{K} e'_k e''_k}. \tag{32}$$

The second derivative is

$$\frac{d^2}{d\beta^2}\lambda_1(\beta) = \frac{\sum_{j=1}^{K}(R'_j + R''_j)^2 e'_j e''_j}{\sum_{k=1}^{K} e'_j e''_j} - \left(\frac{\sum_{j=1}^{K}(R'_j + R''_j)e'_j e''_j}{\sum_{k=1}^{K} e'_j e''_j}\right)^2. \tag{33}$$

Therefore,

$$\frac{d^2}{d\beta^2}\lambda_1(\beta) > 0 \iff \left(\sum_{k=1}^{K} e'_j e''_j\right)\left(\sum_{j=1}^{K}(R'_j + R''_j)^2 e'_j e''_j\right) - \left(\sum_{j=1}^{K}(R'_j + R''_j)e'_j e''_j\right)^2 > 0 \tag{34}$$

Using the distributive property of the product over the sum, the expression becomes

$$\sum_{k=1}^{K}\sum_{j=1}^{K}(R'_j + R''_j)^2 e'_j e''_j e'_k e''_k - \sum_{k=1}^{K}\sum_{j=1}^{K}(R'_j + R''_j)(R'_k + R''_k)e'_j e''_j e'_k e''_k > 0 \tag{35}$$

$$\iff \sum_{k=1}^{K}\sum_{j=1}^{K}[(R'_j + R''_j) - (R'_k + R''_k)](R'_j + R''_j)e'_j e''_j e'_k e''_k > 0 \tag{36}$$

Let $\Delta_{(jj),(kk)} = (R'_j + R''_j) - (R'_k + R''_k)$ define the difference in the cost attributed to reference class j and the cost attributed to class k, accumulated over X', X'', and $E_{jk} = e'_j e''_j e'_k e''_k$.

Overall, the sum can be expressed as:

$$\sum_{k=1}^{K}\sum_{j=1}^{K}[(R'_j + R''_j) - (R'_k + R''_k)](R'_j + R''_j)e'_j e''_j e'_k e''_k = \sum_{k=1}^{K}\sum_{j=1}^{K}(R'_j + R''_j)E_{jk}\Delta_{(jj),(kk)} = \sum_{k=1}^{K}\sum_{j=1}^{K} S_{(jj),(kk)} \tag{37}$$

Note that $\Delta_{(jj),(jj)} = 0 \implies S_{(jj),(jj)} = 0$. Moreover, $\Delta_{(jj),(kk)} = -\Delta_{(kk),(jj)}$ and $E_{jk} = E_{kj}$. Then, the previous term can be rewritten as

$$\sum_{j=1}^{K}\sum_{k<j}^{K} S_{(jj),(kk)} + S_{(kk),(jj)} = \sum_{j=1}^{K}\sum_{k<j}^{K}(R'_j + R''_j)E_{jk}\Delta_{(jj),(kk)} + (R'_k + R''_k)E_{kj}\Delta_{(kk),(jj)} \tag{38}$$

$$= \sum_{j=1}^{K}\sum_{k<j}^{K} E_{jk}\Delta_{(jj),(kk)}[(R'_j + R''_j) - (R'_k + R''_k)] = \sum_{j=1}^{K}\sum_{k<j}^{K} E_{jk}\Delta^2_{(jj),(kk)} \tag{39}$$

The last term is strictly positive for $E_{jk} > 0 \implies e'_j, e''_j > 0$, for $j = 1, \ldots, K$, which is always possible for $\beta > 0$. Therefore, the first term is convex.

We proceed equivalently with the second term:

$$\lambda_2(\beta) = \log\left(\sum_{j=1}^{K} e'_j \sum_{k=1}^{K} e''_k\right) = \log\left(\sum_{k=1}^{K}\sum_{j=1}^{K} e'_j e''_k\right) \tag{40}$$

$$\frac{d}{d\beta}\lambda_2(\beta) = \frac{\sum_{k=1}^{K}\sum_{j=1}^{K}(R'_j + R''_k)e'_j e''_k}{\sum_{k=1}^{K}\sum_{j=1}^{K} e'_j e''_k} \tag{41}$$

$$\frac{d^2}{d\beta^2}\lambda_2(\beta) = \frac{\sum_{k=1}^{K}\sum_{j=1}^{K}(R'_j + R''_k)^2 e'_j e''_k}{\sum_{k=1}^{K}\sum_{j=1}^{K} e'_j e''_k} - \left(\frac{\sum_{k=1}^{K}\sum_{j=1}^{K}(R'_j + R''_k)e'_j e''_k}{\sum_{k=1}^{K}\sum_{j=1}^{K} e'_j e''_k}\right)^2 > 0 \tag{42}$$

$$\iff \left(\sum_{k=1}^{K}\sum_{j=1}^{K} e'_j e''_k\right)\left(\sum_{k=1}^{K}\sum_{j=1}^{K}(R'_j + R''_k)^2 e'_j e''_k\right) - \left(\sum_{k=1}^{K}\sum_{j=1}^{K}(R'_j + R''_k)e'_j e''_k\right)^2 > 0 \tag{43}$$

$$\iff \sum_{k=1}^{K}\sum_{q=1}^{K}\sum_{j=1}^{K}\sum_{i=1}^{K}(R'_j + R''_k)^2 e'_j e''_k e'_i - (R'_j + R''_k)e'_j e''_k (R'_i + R''_q)e'_i e''_i > 0 \tag{44}$$

$$\iff \sum_{k=1}^{K}\sum_{q=1}^{K}\sum_{j=1}^{K}\sum_{i=1}^{K}(R'_j + R''_k)e'_j e''_k e'_i e''_i[(R'_j + R''_k) - (R'_i + R''_q)] > 0 \tag{45}$$

Similarly, as we did before, we define $e'_j e''_k e'_i e''_i = E_{(jk),(iq)} = E_{(ik),(jq)} = E_{(jq),(ik)} = E_{(iq),(jk)}$ and $\Delta_{(jk),(iq)} = (R'_j - R'_i) + (R''_k - R''_q) = -\Delta_{(iq),(jk)}$. Then,

$$\frac{d^2}{d^2\beta}\lambda_2(\beta) = \sum_{k=1}^{K}\sum_{q=1}^{K}\sum_{j=1}^{K}\sum_{i=1}^{K} S_{(jk),(iq)} = \sum_{k=1}^{K}\sum_{q=1}^{K}\sum_{j=1}^{K}\sum_{i=1}^{K}(R'_j + R''_k)E_{(jk),(iq)}\Delta_{(jk),(iq)} \tag{46}$$

Therefore,

$$\sum_{k=1}^{K}\sum_{q<k}^{K}\sum_{j=1}^{K}\sum_{i<j}^{K} S_{(jk),(iq)} + S_{(iq),(jk)} \tag{47}$$

$$= \sum_{k=1}^{K}\sum_{q<k}^{K}\sum_{j=1}^{K}\sum_{i<j}^{K}(R'_j + R''_k)E_{(jk),(iq)}\Delta_{(jk),(iq)} + (R'_i + R''_q)E_{(iq),(jk)}\Delta_{(iq),(jk)} \tag{48}$$

$$= \sum_{k=1}^{K}\sum_{q<k}^{K}\sum_{j=1}^{K}\sum_{i<j}^{K} E_{(jk),(iq)}\Delta_{(jk),(iq)}[(R'_j + R''_k) - (R'_i + R''_q)] \tag{49}$$

$$= \sum_{k=1}^{K}\sum_{q<k}^{K}\sum_{j=1}^{K}\sum_{i<j}^{K} E_{(jk),(iq)}\Delta^2_{(jk),(iq)} \tag{50}$$

The last term is again positive for $e'_j, e''_j > 0$, $j = 1, \ldots, K$.

Last, we prove the convexity of $\lambda(\beta)$:

$$\frac{d^2}{d\beta^2}\lambda(\beta) = \left(\sum_{k=1}^{K}\sum_{q<k}^{K}\sum_{j=1}^{K}\sum_{i<j}^{K} E_{(jk),(iq)}\Delta^2_{(jk),(iq)} - \sum_{j'=1}^{K}\sum_{i'<j'}^{K} E_{(j'j'),(k'k')}\Delta^2_{(j'j'),(k'k')}\right) \tag{51}$$

By reindexing $k' = j$ and $j' = i$ it is clear that the second sum is contained in the first one, thus the negative terms nullify, and the derivative is positive. Proving that $\lambda(\beta)$ is absolutely convex in $\mathbb{R}^+$. □

Note that concavity is assured for positive β and on the limit $\beta \to \infty$ the curvature is not defined, so it is advisable to start a numerical optimization procedure at a value $\beta_0 = 0^+$, since

$$\lim_{\beta \to 0^+} \frac{d^2}{d\beta^2}F(\beta) > 0. \tag{52}$$

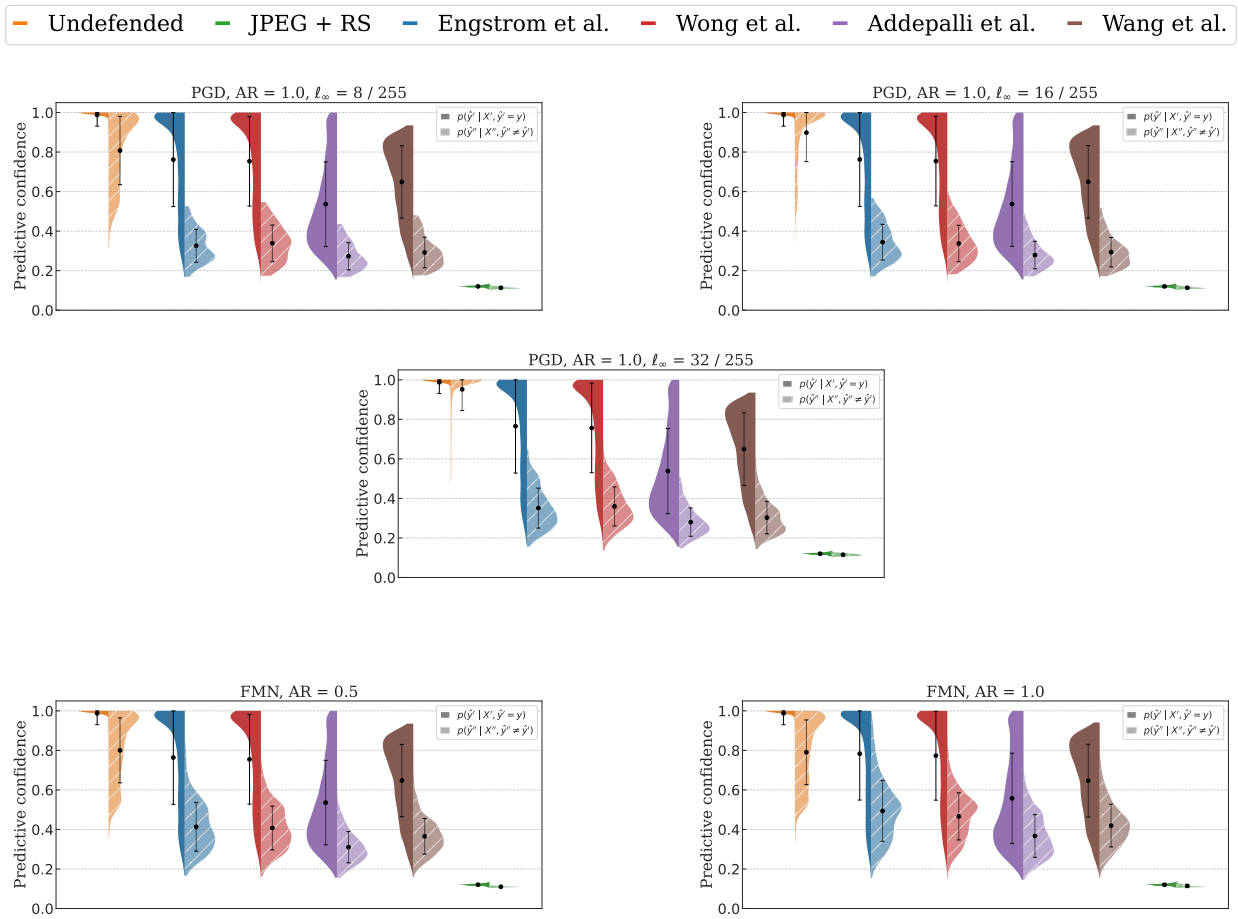


Figure 9: Adversarial setting: predictive confidence distribution for all tested models.

B Extended Results

In Figure 9, we report the predictive confidence distribution for all the tested models in the adversarial setting. Again, it can be seen that the robust models lower their average confidence on X'', showing that they are effectively detecting an ongoing attack. Addepalli et al. (2022) presents a long-tailed distribution before the attack, favouring more conservative, less confident predictions, which penalize the detection of an attack.

We include a plot of FMN for $AR = 1$. This case contains attacked images with a very high norm ($\ell_\infty > 32$). The models were not trained to resist such powerful attacks, therefore the average conditional confidence is more variable and all standard errors overlap.

C PA is not accuracy

It may be tempting to use Posterior Agreement to measure model performance instead of accuracy metrics. In Figure 10, we show an example of why this is a wrong use of PA. In particular, we display the performance of PA and of a classification model (DistilBERT-base) under three different shift strategies: (i) Levenshtein: addition, removal or substitution of characters in the sentence, (ii) Amplification: addition of adjectives reinforcing the sentence's sentiment (iii) Contradiction: addition of adjective weakening the sentence's sentiment. While in cases (i) and (iii) both metrics behave similarly, in case (ii) the addition of reinforcing adjectives has the effect of increasing the confidence in the model predictions, improving the

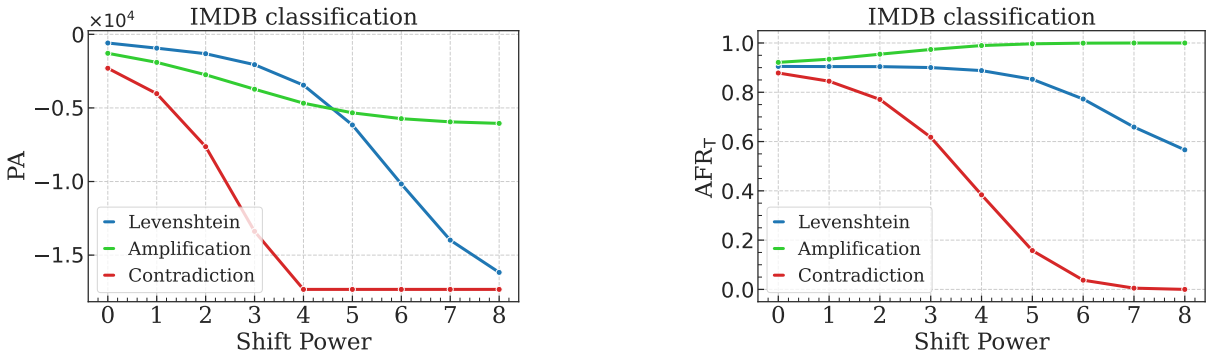


Figure 10: PA and accuracy for the IMDB sentiment classification task (Maas et al., 2011) under simple attacks. Observations are perturbed by manipulating some characters (Levenshtein) and by replacing some words with positive or negative adjectives that either encourage (Amplification) or discourage (Contradiction) the true sentiment of the review. The shift power is defined as W, being 2^W the number of replacements.

overall performance. The sensibility of the method to this shift is correctly detected by PA, that *penalizes* the model decreasing its robustness score. This example again remarks that PA and accuracy are different metrics meant to investigate distinct aspects of a model.

D Robustness and feature alignment

To better understand the impact of controlled domain shifts on robustness assessment through PA, we used ERM and IRM algorithms to train a ResNet18 model for 50 epochs on $\mathcal{D}_{\text{train}}$, using Adam with a learning rate of 10^{-2}. From each validation sample $x_0^{\text{val}}, x_1^{\text{val}}$, we selected 128 observations to form a reduced validation set $\mathcal{D}_{\text{sub}} = \{x_0^{\text{sub}}, x_1^{\text{sub}}\} \subset \mathcal{D}_{\text{val}}$. Both x_0^{sub} and x_1^{sub} share the same random seed τ^{val}, thus containing the same MNIST samples in the same order, and they differ only by a hue-based distribution shift (blue *vs* red; *cf.* Table 2).

At the end of each epoch, we compute the principal component (via PCA) of the feature representations of x_0^{sub} and x_1^{sub} separately. Each MNIST observation in $\mathcal{D}_{\text{sub}}$ thus has two projected values, one for each sample. By examining this principal component (the direction of greatest variance), we can qualitatively evaluate the inductive bias encoded after each training epoch.

Let Φ^c be the feature extractor of a classifier c after a given epoch. Denote by $\mathbf{v}_0$ and $\mathbf{v}_1$ the principal directions of the feature space for x_0^{sub} and x_1^{sub}, respectively. We compute the projections of each observation $x_{0,n}^{\text{sub}}$ and $x_{1,n}^{\text{sub}}$ onto these directions:

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

The distribution of $z_{0,n}$ and $z_{1,n}$ across different classes indicates how well the inductive bias aligns with task-relevant features: ideally, class membership should drive the principal variance in the feature space. The class-conditional variance of these projections thus gauges the discriminative strength of the model's latent features and, in turn, its robustness to sampling variability.

Figure 11 illustrates the principal component projections of the feature space representations of samples x_0^{sub}, x_1^{sub} for ERM and IRM algorithms at three different training stages. These results show that ERM is unable to encode a representation that is both discriminative and invariant to domain shifts, as it either displays a high cross-domain error or a high class-conditional variance. This indicates that its inductive bias is exclusively driven by domain-specific features or by class-specific features, and possibly the high learning rate avoids the model to converge to a more robust solution. In contrast, IRM is able to encode a representation that reduces both qualitative measures at the same time, which indicates that the inductive

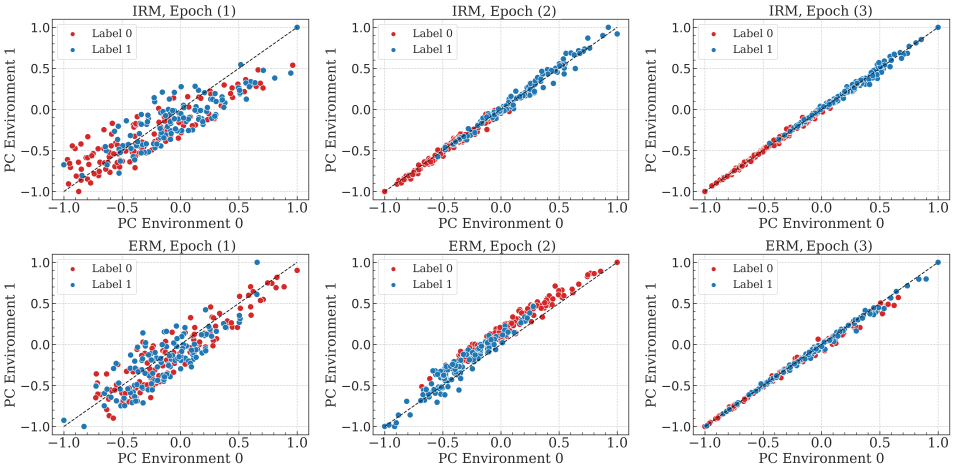


Figure 11: Normalized principal component for each environment at three different training stages. ERM (top) and IRM (bottom) algorithms are considered, and projections are colored by class membership. ERM projections display either a high cross-domain error or a high class-conditional variance. In contrast, IRM is able to encode a representation that reduces both measures at the same time, thus indicating a more robust inductive bias.

bias is able to capture the most predictive features for the task at hand without being significantly influenced by the shift in the hue factor.

These measures have been shown to qualitatively assess the suitability of the inductive bias for the aforementioned sources of randomness separately.

E Details on the DiagViB-6 Dataset

The DiagViB-6 dataset comprises both source domains ($\mathcal{S} = \{X_0, X_1\}$) and target domains ($\mathcal{T} = \{X_2, X_3, X_4, X_5\}$). Here, X_j represents the random variable corresponding to domain j, where j indicates the number of shifted factors relative to X_0. The classification task focuses on predicting the *shape factor* (the digit) using handwritten 4s and 9s from MNIST. We systematically control the *covariate shift* by varying visual attributes (*e.g.*, color/hue) while keeping class identities intact.

Data Splits and Notation For each domain X_j, we generate four disjoint subsets of MNIST by applying different random instantiations τ_0^{train}, τ_1^{train}, τ^{val}, and τ^{test}. We thus define:

- $\mathcal{D}_{\text{train}} = \{x_0^{\text{train}}, x_1^{\text{train}}\}$, where $x_j := x_j \circ \tau_j^{\text{train}}$, for $j = 0, 1$.
- $\mathcal{D}_{\text{val}} = \{x_0^{\text{val}},\ x_1^{\text{val}}\}$, where $x_j := x_j \circ \tau^{\text{val}}$, for $j = 0, 1$.
- $\mathcal{D}_{\text{test}}^{(j)} = \{x_j^{\text{test}}\}$, where $x_j^{\text{test}} := x_j^{\text{test}} \circ \tau^{\text{test}}$, for $j = 0, \ldots, 5$.

In this way, the *training* data is subject to both sampling randomness ($\tau_0^{\text{train}} \neq \tau_1^{\text{train}}$) and domain shift ($X_0 \not\sim X_1$), while the *validation* and *test* splits each use a single instantiation, ensuring that *distribution shift* (*i.e.*, changes in hue or other visual attributes) is the only significant source of randomness across X' and X''. Overall, we create 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.

Controlled Covariate Shift By design, each shift factor introduced in domain X_j modifies specific image features (*e.g.*, color channels, texture patterns) while preserving the underlying digit shape. Because each domain relies on the same MNIST base classes (4s and 9s), domain shift can be precisely *engineered* to test a

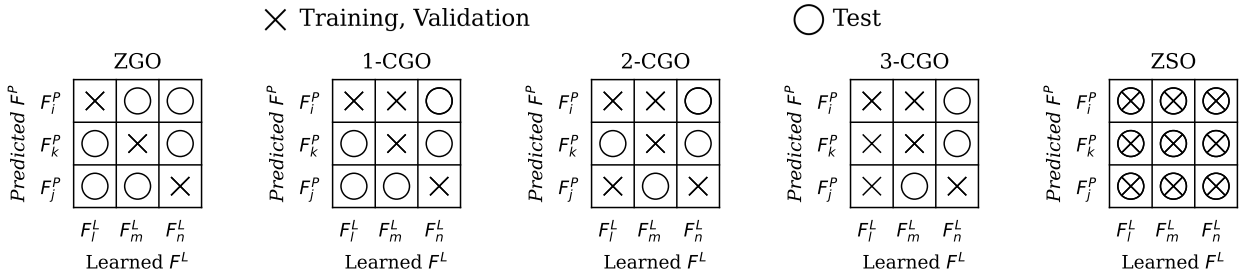


Figure 12: Representation of the co-occurrence pattern in between learning factors F^L and predicted factors F^P for the ZGO, CGO and ZSO settings that will be considered in this experiment.

# 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 2: Specific image factors associated with each environment considered in the model discriminability experiments. CC and LC account for *centered center* and *centered low*, respectively.

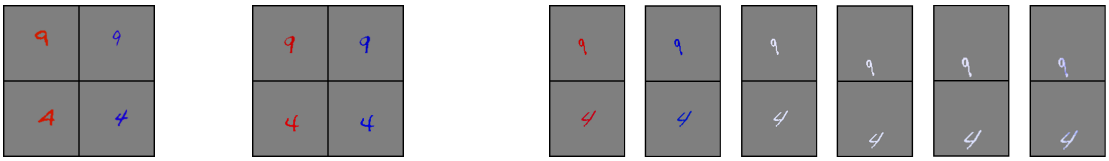
model's robustness. Hence, the DiagViB-6 setup allows for a clear separation between *sampling randomness* and *shift-induced variability*:

- Sampling randomness: Different subsets of MNIST (different seeds) across training, validation, and test.
- Domain shift: Controlled variations (*e.g.*, hue) systematically applied to form $X_1, \ldots, X_5$ from X_0.

Rationale for DiagViB-6 This setup maximizes the potential for learning *invariant features* by exposing models to progressively more challenging shifts during training. It also provides strong *validation and testing* conditions to evaluate how well the learned representations generalize to new shifts. Because the same underlying digit classes are maintained across domains, the impact of domain shift on classification can be directly attributed to feature manipulation (*e.g.*, hue, stroke thickness), rather than confounded by changes in class identity. This approach follows the methodology outlined in Geirhos et al. (2020), where controlling image factors enables a clearer view of the *shortcut learning* phenomena.

DiagViB-6 for Model Discriminability To adjust DiagViB-6 to specific parameters of our multiple environments, we have defined the values described in Table 2 as the shift factor settings for each environment. An example of training, validation, and test samples subject to different distribution shifts is depicted in Figure 13.

DiagViB-6 for Model Selection To evaluate model selection under domain generalization, we examined the impact of *hue shift* as the primary source of inductive bias. Since hue represents a significant variation in image representation, this experiment provides insight into the consistency of model selection criteria. For this experiment, two different sampling instantiations for the validation datasets are considered, denoted as $\tau_0^{\text{val}} \neq \tau_1^{\text{val}}$. This means that $\mathcal{D}_{\text{val}}^{(0)}$ and $\mathcal{D}_{\text{val}}^{(1)}$ correspond to different instantiations of the hue shift factor. The training and validation dataset configurations are detailed in Table 2. Similarly to the model discriminability experiment, training datasets are always drawn from source domains and consist of samples $\mathcal{D}_{\text{train}} = \{x_0^{\text{train}}, x_1^{\text{train}}\}$. Validation datasets can be drawn from either source or target domains, leading to different generalization scenarios.



Train Validation Test (0 - 5)

Figure 13: Samples from the training, validation and test datasets. Training samples belong to different MNIST subsets, whereas validation and test samples are the transformations of the same observation.

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

Table 3: Image factors associated with each of the environments considered in the model selection experiment. CC and UL account for 'centered center' and 'upper left', respectively.

First, when both $\mathcal{D}_{\text{val}}^{(0)}$ and $\mathcal{D}_{\text{val}}^{(1)}$ are drawn from *source domains*, two configurations are considered. If validation samples originate from the same distribution as training, they are in the *same distribution* (SD) setting. If they come from source domains but with different hue instantiations than the training data, they are in the *in-distribution* (ID) setting. Second, when $\mathcal{D}_{\text{val}}^{(0)}$ is drawn from source domains and $\mathcal{D}_{\text{val}}^{(1)}$ from target domains, the *mixed distribution* (MD) setting applies. Depending on the magnitude of the hue shift, we define two cases: a 1-*factor mixed distribution* (1F-MD), where only hue varies, and a 5-*factor mixed distribution* (5F-MD), where hue is combined with additional variations. Finally, when both $\mathcal{D}_{\text{val}}^{(0)}$ and $\mathcal{D}_{\text{val}}^{(1)}$ are drawn from *target domains*, the setup corresponds to the *out-of-distribution* (OOD) setting, where validation samples differ entirely from training.

This setup ensures that models are systematically tested under increasing levels of domain shift, from mild (ID, 1F-MD) to severe (5F-MD, OOD). By doing so, we assess the robustness of model selection criteria when generalizing to unseen hue shifts. For more details on the final dataset configuration please refer to Table 3.