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

Master Thesis

Víctor Jiménez Rodríguez

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

ETH Zurich

September 19, 2024

Experimental pathway

Posterior Agreement has been proposed as a theoretically-grounded alternative for model robustness assessment in covariate shift settings.

1. Characterization of the robustness problem and the sources of randomness that are relevant in the context of image classification tasks.
2. Properties of Posterior Agreement as a robustness metric.
3. Robustness assessment in the adversarial setting.
4. Robustness assessment in the out-of-distribution setting.
5. Model selection with early-stopping.

The robustness challenge

The robustness challenge – Introduction

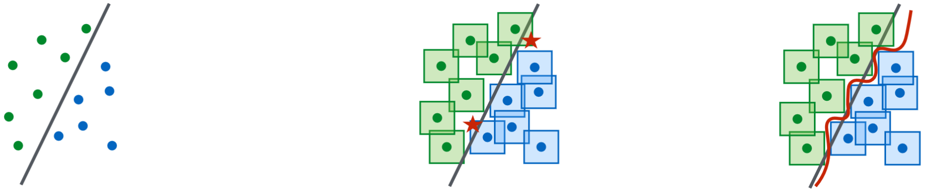
Goal: Maintain predictive power under expected variations in the data.

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

Figure: Illustrative example of the three expected sources of randomness in the context of image classification.

The robustness challenge – Challenges

- Lack of understanding of the inductive bias.
- Robust vs non-robust features.
- Generalization-complexity trade-off.



(a) Linsep points (b) Standard training (c) Adversarial training (ℓ_∞)

Figure: Standard vs adversarial training decision boundaries. [11]

Learning framework

Learning framework – Introduction

Goal: 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.

- Images $x \in \mathcal{X}$ belonging to $K = |\mathcal{Y}|$ classes.
- f^* encodes the causal structure underlying the data generation process.
- $\mathcal{F}$ is composed of NN-parametrized classifiers.

Learning framework – Model

Definition (*Classifier*): Let $K \in \mathbb{N} < \infty$ be the cardinality of $\mathcal{Y}$.

$$\begin{array}{ccccccc} f : \mathcal{X} & \longmapsto & \mathbb{R}^d & \longmapsto & \mathbb{R}^K & \longmapsto & \mathcal{Y} = \{1, \ldots, K\} \\ x & \longmapsto & z & \longmapsto & \boldsymbol{F}(z) & \longmapsto & \hat{y} = \arg\max_k F_k(z) \end{array}$$

Under a NN parametrization Γ:

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

Learning framework – Algorithm

Definition (*K-class classification problem*):

- Let L be the cross-entropy loss function.

$$L(y) = -\log F_y(z;\gamma)$$

- Let $\hat{R}(f)$ be the empirical risk.

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

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

Posterior Agreement

Posterior Agreement

Definition (*Hypothesis class*) A data science algorithm learns a function f implementing the following mapping:

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

Definition (*Posterior*) The posterior $\mathbf{P}^f \in \mathfrak{P}^f$ establishes the stochastic relation between experiment realizations and hypotheses.

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

Posterior Agreement

Definition (*Generalization error*):

- Let $\boldsymbol{x}'$ and $\boldsymbol{x}''$ be realizations of $\boldsymbol{X}$.
- Let Θ be the hypothesis class represented by f given $\boldsymbol{X}$.
- Let $-\log \mathbf{P}^f(\cdot)$ be the description length of the posterior.

The generalization error is defined as the out-of-sample description length:

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

Intuitively, a low generalization error is obtained when good quality hypothesis on $\boldsymbol{x}''$ are likely to be drawn from $\boldsymbol{x}'$.

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

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

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

Proposition: Posterior agreement based model selection criterion.

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

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

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

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

Proposition (*Posterior Agreement kernel*): With no prior information about Θ, the posterior agreement kernel for supervised K-class classification tasks has the following expression:

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

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

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

Theorem. The posterior agreement kernel for K-classification problems $\mathrm{PA}\left(\boldsymbol{x}', \boldsymbol{x}''; \beta\right)$ has the following properties $\forall \boldsymbol{x}', \boldsymbol{x}'' \sim \boldsymbol{X}$ and $\beta \in \mathbb{R}^{+}$.

- P1 (Boundedness) $\mathrm{PA}\left(\boldsymbol{x}', \boldsymbol{x}''; \beta\right) \leq N \log K$. Because $\log_2 K$ bits are needed to encode a uniform distribution over the classes for each observation.
- P2 (Symmetry) $\mathrm{PA}\left(\boldsymbol{x}', \boldsymbol{x}''; \beta\right) = \mathrm{PA}\left(\boldsymbol{x}'', \boldsymbol{x}'; \beta\right)$. Because randomness instantiations are not observable.
- P3 (Concavity) $\mathrm{PA}\left(\boldsymbol{x}', \boldsymbol{x}''; \beta\right)$ is a concave function of $\beta \in \mathbb{R}^{+}$. Then the kernel optimization problem will have a unique solution.

In particular, the maximum value $\mathrm{PA} \equiv \mathrm{PA}\left(\boldsymbol{x}', \boldsymbol{x}''; \beta^{*}\right)$ is bounded by the situations $\beta^{*} \longrightarrow 0$ and $\beta^{*} \longrightarrow \infty$. Then:

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

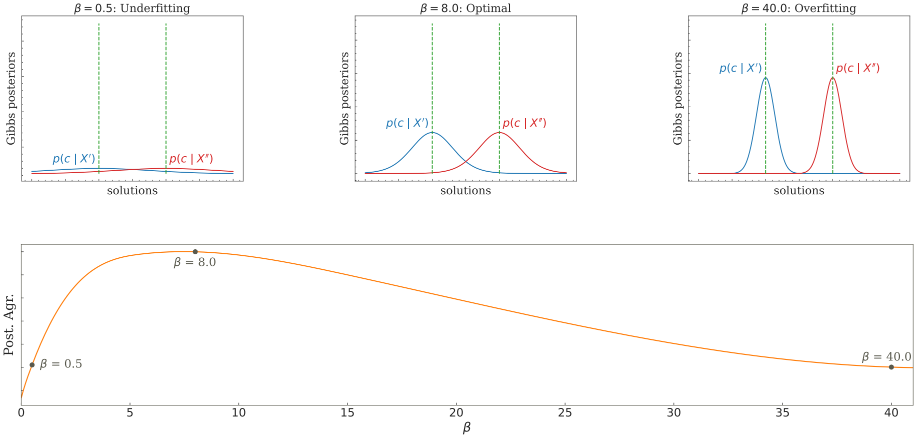


Figure: Illustration of the optimization over the inverse temperature parameter β. Posterior Agreement is maximum at a value β^* in which hypothesis selected from the posterior over $\theta \mid \boldsymbol{x'}$ are assigned a high probability by the posterior over $\theta \mid \boldsymbol{x''}$.

Robustness against covariate shift

Robustness against covariate shift – Robustness metric

Proposition. A robustness metric should possess the following properties:

- P1 (Non-increasing) The metric should be non-increasing with respect to the response of the model under increasing levels of covariate shift.
- P2 (Independent discriminability) The metric should discriminate models exclusively by their generalization capabilities against covariate shift. For instance, the metric should be independent of the task performance.

Example I. Consider the following classifiers in a balanced dataset:

Classifier	Performance	Robustness
Perfect	1.0	Max.
Constant	$1/K$	Max.
Random	$1/K$	Min.

A binary sample $\boldsymbol{y} \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(p)$ of size 1000 is generated, with $p = \mathbf{P}_Y(y = 1)$.

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

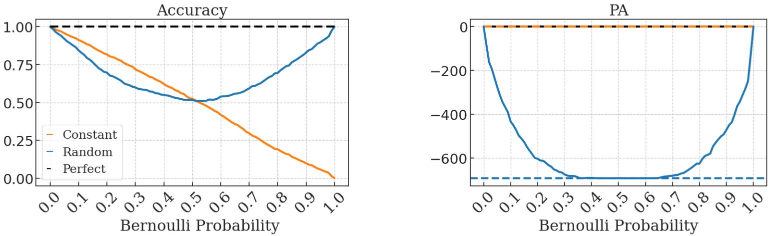


Figure: Evolution of PA and accuracy for constant, perfect and random classifiers across different values of $p \in [0, 1]$.

Example II. Consider a sentiment classifier in the IMDB dataset. $\boldsymbol{x}'$ is a sample of original reviews, and $\boldsymbol{x}''$ is formed by manipulating every $x \in \boldsymbol{x}'$.

- Levenshtein: Addition, removal or substitution of characters.
- Amplification: Addition of reinforcing adjectives.
- Contradiction: Addition of contradicting adjectives.

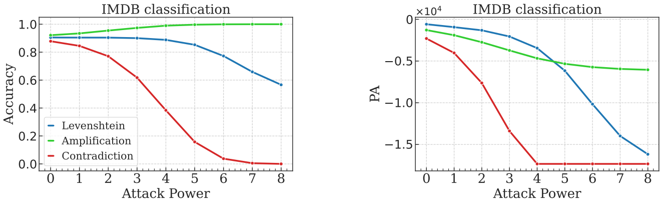


Figure: Accuracy and PA for the IMDB sentiment classification task under random and adversarial perturbations. The attack power is defined as 2^W, being W the number of modifications performed.

Adversarial setting

Adversarial setting – Attacks

Definition (*Perturbation*) Let $\mathbf{B}_p^\epsilon(x)$ be the ℓ_p-norm ball of radius ϵ centered at observation x. A perturbation Δ is defined as

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

Attack (*PGD*): Projected gradient descent.

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

Attack (*FMN*): Fast minimum norm.

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

Adversarial setting – Characterization

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

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

Definition (*Attack failure rate*) Let $\hat{\boldsymbol{y}}', \hat{\boldsymbol{y}}'' \in \mathcal{Y}^N$ be the predicted class labels for $\boldsymbol{x}'$ and $\boldsymbol{x}''$, respectively, and let $\boldsymbol{y} \in \mathcal{Y}^N$ be the true labels.

$$\mathrm{AFR}_{\mathsf{T}} = \mathrm{Accuracy}(\hat{\boldsymbol{y}}'', \boldsymbol{y}), \quad \mathrm{AFR}_{\mathsf{P}} = \mathrm{Accuracy}(\hat{\boldsymbol{y}}'', \hat{\boldsymbol{y}}').$$

The variation of AFR with respect to α is also reported:

$$\Delta\,\mathrm{AFR} = \mathrm{AFR}_{\mathsf{T}}\Big|_{\alpha=1} - \mathrm{AFR}_{\mathsf{T}}\Big|_{\alpha=0} = \mathrm{AFR}_{\mathsf{P}}\Big|_{\alpha=1} - \mathrm{AFR}_{\mathsf{P}}\Big|_{\alpha=0}$$

Adversarial setting – Experimental setup

Experimental setup. CIFAR10 [10] is a balanced dataset containing 60.000 colored 32 $\times$ 32 pixel images belonging to 10 different classes. We will consider a pre-trained WideResNet-28-10 [17] as a baseline, **Undefended** model and compare it to some state-of-the-art robust ResNet50 [7] models provided by the RobustBench [4] library under PGD [11] and FMN [12] attacks, both run for 1000 steps. The defenses applied are those proposed by **Engstrom et al.** [5], **Athalye et al.** [3], **Wong et al.** [14], **Addepalli et al.** [1] and **Wang et al.** [13].

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

Figure: Original and adversarially-perturbed CIFAR10 observation of class `horse`.

Adversarial setting – PGD

Experiment 1. A pre-trained, undefended WideResNet-28-10 and five RobustBench defended models are subject to a 1000 step PGD attack with $\ell_\infty = 8/255$.

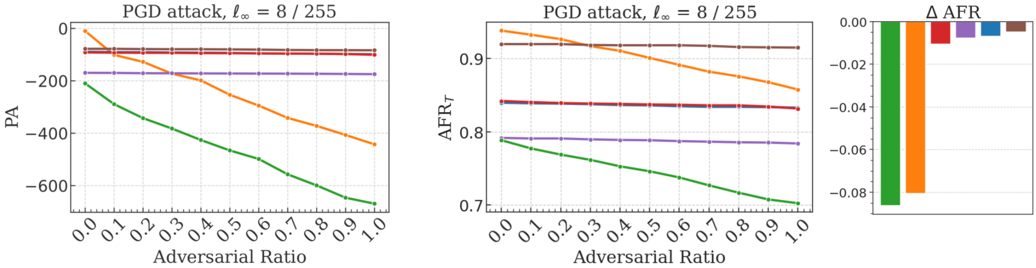


Figure: PA, AFR_T and the AFR variation against increasing adversarial ratio $\alpha \in [0, 1]$.

Adversarial setting – PGD

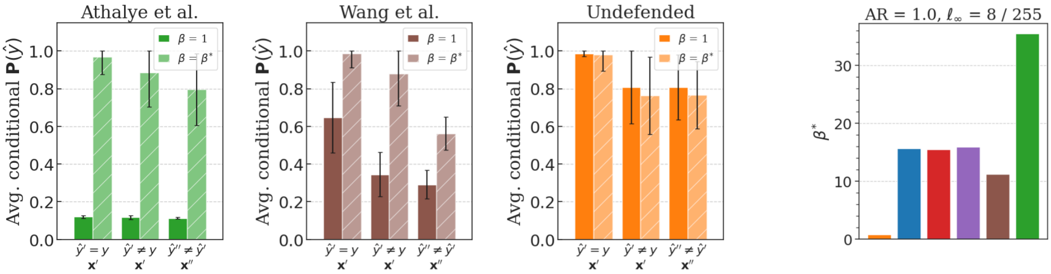


Figure: (**left**) Average posterior probability of the predicted class for correctly classified original observations, misclassified original observations and misleading adversarial observations. (**right**) Optimal β^* value achieved by each model.

Adversarial setting – FMN

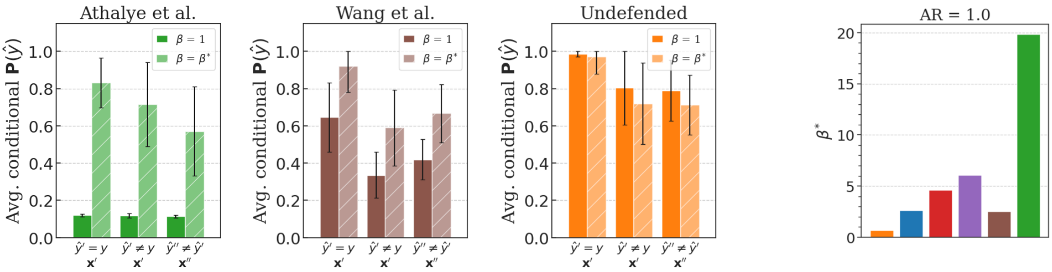


Figure: (**left**) Average posterior probability of the predicted class under FMN attack for correctly classified original observations, misclassified original observations, and misleading adversarial observations. (**right**) Optimal β^* value for each model.

Adversarial setting – PGD vs FMN

Comparison between metrics.

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

Table: Comparison of PA, $\mathrm{AFR_P}$ and $\mathrm{AFR_T}$ scores for a PGD attack with ℓ_∞=16/255 and a FMN attack across different adversarial ratio values.

Adversarial setting – PGD vs FMN

Approximated PA contributions. A surrogate version of the PA kernel can be obtained considering the average posterior for the two main robustness contributions:

- Sampling randomness contribution ζ_{SAM}, accounting for N_{SAM} misclassified observations in $\boldsymbol{x}'$ with average probability ρ_{MIS}.
- Adversarial attack contribution ζ_{ADV}, accounting for N_{ADV} misleading adversarial observations in $\boldsymbol{x}''$ with average probability ρ_{ADV}.

	PGD, ℓ_∞=16/255						**FMN**					
Defense	N_{MIS}	ρ_{MIS}	ζ_{SAM}	N_{ADV}	ρ_{ADV}	ζ_{ADV}	N_{MIS}	ρ_{MIS}	ζ_{SAM}	N_{ADV}	ρ_{ADV}	ζ_{ADV}
Wang et al.	799	0.88	**-468.62**	47	0.56	**-39.44**	435	0.59	**-1599.08**	4215	0.67	**-4637.45**
Engstrom et al.	1591	0.91	**-566.72**	67	0.61	**-63.43**	1125	0.64	**-2469.65**	2505	0.68	**-2847.99**
Wong et al.	1562	0.91	**-537.25**	90	0.62	**-88.98**	1032	0.77	**-1125.53**	2920	0.73	**-3844.40**
Addepalli et al.	2063	0.89	**-877.42**	75	0.54	**-58.92**	1507	0.74	**-1910.69**	2187	0.72	**-2788.89**
Undefended	566	0.77	**-736.63**	810	0.76	**-1173.55**	412	0.72	**-704.58**	3132	0.71	**-3906.55**
Athalye et al.	1915	0.88	**-963.85**	747	0.79	**-1183.96**	859	0.74	**-2054.23**	4679	0.57	**-3955.43**

Table: Approximated PA contributions for a PGD attack with ℓ_∞=16/255 and a FMM attack.

Out-of-distribution setting

Out-of-distribution setting – Robust learners

- **IRM** (domain alignment): Regularization term that pushes towards the minimization of the dissimilarity of feature representations originated from different source environments [2].
- **LISA** (data augmentation): Artificial observations are generated by intra-domain and/or intra-label interpolation [15].

(a) Original

(b) Mixup [18]

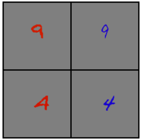
(c) CutMix [16]

Figure: Mixup and Cutmix strategies can be used to interpolate between different labels and/or domains by generating intermediate observations. [16]

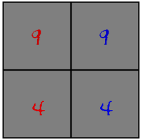
Out-of-distribution setting – Experimental setup

Experimental setup. The DiagViB-6 dataset framework [6] 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. ERM, IRM [2] and LISA [15] algorithms were used to train a ResNet18 architecture for 100 epochs on dataset D^{train} using Adam [8] optimizer with a learning rate of 10^{-3}. Accuracy on validation dataset D^{val} was monitored and weights achieving maximum performance were selected for evaluation.

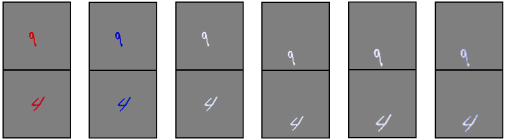
$$D^{\text{train}} = \{\boldsymbol{x}_0^{\text{train}}, \boldsymbol{x}_1^{\text{train}}\}, \quad D^{\text{val}} = \{\boldsymbol{x}_0^{\text{val}}, \boldsymbol{x}_1^{\text{val}}\},$$
$$D^{\text{test}} = \{\boldsymbol{x}_0^{\text{test}}, \boldsymbol{x}_1^{\text{test}}, \boldsymbol{x}_2^{\text{test}}, \boldsymbol{x}_3^{\text{test}}, \boldsymbol{x}_4^{\text{test}}, \boldsymbol{x}_5^{\text{test}}\}.$$



Train



Validation



Test (0 - 5)

Out-of-distribution setting – Non-paired samples

Experiment 2. D^{val} and D^{test} are each generated from a different MNIST sample.

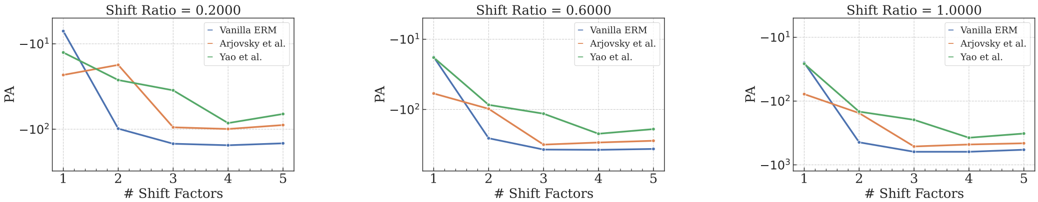


Figure: Evolution of PA under increasing levels of shift power and shift ratio α.

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

Table: Comparison of PA, $\mathrm{AFR_P}$ and $\mathrm{AFR_T}$ scores for $\alpha = 1$.

Model selection

Model selection – Challenges

- Agreement between predictive outcomes across different samples no longer guarantees that the set of features learned are relevant for the task at hand.
- A classifier overfitting to specific features during training would lower its performance on validation data and simultaneously be considered robust.
- The ultimate measure of domain adaptation capabilities is accuracy on target domains.

Experiment 1. Access to target domains for model selection purposes is sequentially increased. In most cases, $\boldsymbol{x}_0^{\text{val}}$ will belong to the source.

Experiment 2. The inductive bias of the model is artificially manipulated to encode shortcut learning opportunities.

Model selection – Experiment 1

Experimental setup. Both position and hue are considered as learning factors.

	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: Factors associated with each of the environments considered in this experiment (hue). CC and UL account for 'centered center' and 'upper left', respectively.

Model selection – Experiment 1

	Acc. Test 0			**Acc. Test 1**			**Acc. Test 2**			**Acc. Test 3**			**Acc. Test 4**			**Acc. Test 5**		
IRM	Acc.	AFR_P	PA	Acc.	AFR_P	PA	Acc.	AFR_P	PA	Acc.	AFR_P	PA	Acc.	AFR_P	PA	Acc.	AFR_P	PA
SD	99.3	99.3	99.3	71.5	71.5	**83.3**	70.6	70.6	**91.9**	65.3	65.3	**85.9**	**75.0**	**75.0**	66.5	28.8	28.8	**46.7**
ID	99.4	99.4	99.4	44.3	44.3	44.3	88.1	88.1	88.1	76.1	76.1	76.1	59.4	59.4	59.4	45.2	45.2	45.2
1F-MD	99.4	99.4	99.4	44.3	44.3	44.3	88.1	88.1	88.1	76.1	76.1	76.1	59.4	59.4	59.4	45.2	45.2	45.2
5F-MD	**99.4**	**99.4**	99.3	31.2	31.2	**83.3**	88.7	88.7	**91.9**	73.8	73.8	**85.9**	65.5	65.5	**66.5**	**50.2**	**50.2**	46.7
OOD	99.5	99.5	99.5	62.4	62.4	62.4	90.1	90.1	90.1	84.1	84.1	84.1	52.2	52.2	52.2	42.6	42.6	42.6
LISA	Acc.	AFR_P	PA	Acc.	AFR_P	PA	Acc.	AFR_P	PA	Acc.	AFR_P	PA	Acc.	AFR_P	PA	Acc.	AFR_P	PA
SD	99.3	99.3	**99.4**	**88.4**	**88.4**	83.1	53.9	53.9	**78.5**	57.9	57.9	**80.5**	63.1	63.1	**77.0**	**38.8**	**38.8**	35.4
ID	99.5	99.5	99.5	59.2	59.2	**88.8**	**62.5**	**62.5**	60.4	71.7	71.7	**76.8**	70.9	70.9	**71.5**	35.2	35.2	**36.4**
1F-MD	99.5	99.5	99.5	88.8	88.8	88.8	54.4	54.4	54.4	76.8	76.8	76.8	71.5	71.5	71.5	36.4	36.4	36.4
5F-MD	99.2	99.2	**99.4**	**86.6**	**86.6**	85.1	71.1	71.1	**74.8**	77.3	77.3	**83.7**	73.7	73.7	**81.9**	**49.4**	**49.4**	48.6
OOD	**99.3**	99.2	99.2	91.8	**95.8**	84.4	59.0	63.0	**83.0**	77.6	79.0	**88.0**	73.8	73.6	**84.8**	34.4	40.4	**47.3**

Table: Test performance under increasing levels of shift for models selected through different configurations of validation datasets. PA, AFR_P and accuracy are used as early stopping criteria for model selection in the hue factor experiment.

Model selection – Experiment 2

Experimental setup. Both position and hue are considered as learning factors. The setting for ID model selection (i.e. domain adaptation) requires that validation datasets contain the same configuration of factors than the training datasets. Experiments will be performed for ZGO, ZSO and single, double and triple CGO.

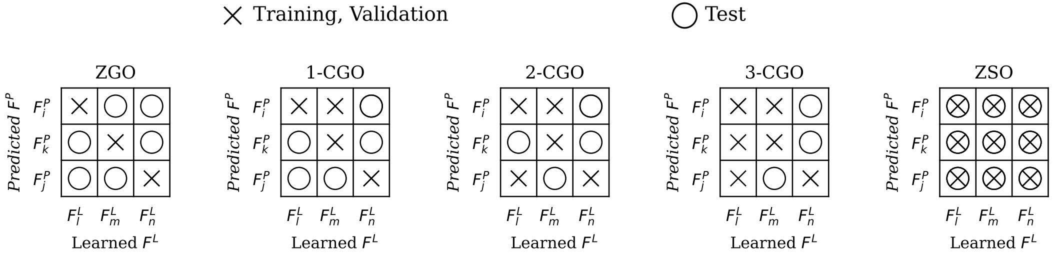


Figure: Representation of the co-occurrence pattern in between learning factors F^L and predicted factors F^P for the ZGO, CGO and ZSO settings that will be considered in this experiment.

Model selection – Experiment 2

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

Table: Test performance under increasing levels of shift for models selected through different configurations of factor co-occurrence for the hue learning factor experiment. Specifically, the performance of models selected through validation accuracy (Acc) and the difference between accuracy-based and PA-based selection (ΔAcc) is reported.

Domain adaptation in WILDS

WILDS – Dataset 1

Experiment 1. The **waterbirds** dataset was considered. The classes `waterbird` and `landbird` are influenced by a spurious correlation with the background.

	Training	Validation	Test
Ratio water / land	~ 2.86	~ 1	~ 1

	Average Acc.			**Worst-case Acc.**		
	Acc.	AFR$_P$	PA	Acc.	AFR$_P$	PA
ERM	**78.52**	73.72	**78.52**	**67.11**	58.90	**67.11**
IRM	**90.16**	89.70	**90.16**	**89.21**	88.67	**89.21**

Table: Average and worst-case test accuracy for the `waterbirds` [9] dataset.

WILDS – Dataset 2

Experiment 2. The **celebA** dataset was considered. The classification task involves the prediction of hair color from images of American celebrities. The subpopulation shift arises from the spurious correlation with the gender.

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

	Average Acc.			**Worst-case Acc.**		
	Acc.	AFR_P	PA	Acc.	AFR_P	PA
ERM	97.80	**98.27**	**98.27**	97.76	**97.77**	**97.77**
IRM	98.12	**98.41**	**98.41**	**98.06**	98.05	98.05

Table: Average and worst-case test accuracy for the `celebA` [9] dataset.

WILDS – Dataset 3

Experiment 3. The **camelyon17** dataset was considered. The classification task involves the identification of tumor tissue in lymph node patches sampled from different hospitals. The out-of-distribution setting originates from the differences in the samples taken from different hospitals.

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

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

Table: Test accuracy for the `camelyon17` [9] dataset.

Conclusions

Conclusions

Blibliography

[1] Sravanti Addepalli, Samyak Jain, Gaurang Sriramanan, and R. Venkatesh Babu. Scaling adversarial training to large perturbation bounds. In *European Conference on Computer Vision*, 2022. URL https://api.semanticscholar.org/CorpusID:252968354.

[2] Martin Arjovsky, LÃ©on Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization. URL http://arxiv.org/abs/1907.02893.

[3] Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, volume 80, pages 274–283. PMLR, 2018.

[4] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: A standardized adversarial robustness benchmark. URL http://arxiv.org/abs/2010.09670.

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

[6] 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. URL http://arxiv.org/abs/2108.05779.

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

[8] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. URL http://arxiv.org/abs/1412.6980.

[9] 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. URL http://arxiv.org/abs/2012.07421.

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

[11] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and

Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. URL http://arxiv.org/abs/1706.06083.

[12] Maura Pintor, Fabio Roli, Wieland Brendel, and Battista Biggio. Fast Minimum-norm Adversarial Attacks through Adaptive Norm Constraints. URL http://arxiv.org/abs/2102.12827.

[13] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training, 2023. URL https://arxiv.org/abs/2302.04638.

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

[15] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving Out-of-Distribution Robustness via Selective Augmentation. URL http://arxiv.org/abs/2201.00299.

[16] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. URL http://arxiv.org/abs/1905.04899.

[17] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016. ISBN 1-901725-59-6. doi: 10.5244/C.30.87. URL https://dx.doi.org/10.5244/C.30.87.

[18] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond Empirical Risk Minimization. URL http://arxiv.org/abs/1710.09412.