Enhancing Unit-tests for Invariance Discovery

Piersilvio De Bartolomeis [1] **Antonio Orvieto** [1] **Giambattista Parascandolo** [2]

Abstract

Recently, Aubin et al. (2021) proposed a set of linear low-dimensional problems to precisely evaluate different types of out-of-distribution generalization. In this paper, we show that one of these problems can already be solved by established algorithms, simply by better hyper-parameter tuning. We then propose an enhanced version of the linear unit-tests. To the best of our hyper-parameter search and within the set of algorithms evaluated, AND-mask is the best performing algorithm on this new suite of tests. Our findings on synthetic data are further reinforced by experiments on an image classification task where we introduce spurious correlations.

1. Introduction

Over the last decade, deep learning has shown impressive performance in a variety of application domains, ranging from computer vision to natural language processing (Collobert & Weston, 2008; He et al., 2015). However, recent years have been marked by a multitude of examples showing that deep learning models are prone to exploiting spurious correlations (Beery et al., 2018). To address this issue, AND-mask (Parascandolo et al., 2020) and several other works (Arjovsky et al., 2019; Koyama & Yamaguchi, 2020; Khezeli et al., 2021) proposed to learn correlations that are invariant across multiple training distributions. The key idea behind these methods is that when different interventions are used to produce the training datasets, invariant correlations should reflect the fixed causal mechanism underlying the target variable. Aubin et al. (2021) noted, however, that these algorithms perform poorly across a catalog of simple low-dimensional linear problems. Since then, it has become common practice for a growing body of literature (Ahuja et al., 2021; Khezeli et al., 2021; Koyama & Yamaguchi, 2020; Du et al., 2021; Wang et al., 2022; Nguyen et al., 2022) to evaluate the performance of algorithms for invariance discovery using the Linear Unit-Tests proposed by Aubin et al. (2021). These unit-tests entail three classes of low-dimensional linear problems, each capturing a different structure for inducing spurious correlations. Here, we will focus on the two linear classification tasks:

- *Example 2* is a classification problem inspired by the cow vs. camel example (Beery et al., 2018) where spurious correlations are interpreted as background color.
- *Example 3* is based on a classification experiment in Parascandolo et al. (2020) where the spurious correlations provide a shortcut in minimizing the training error while the invariant classifier takes a more complex form.

In this paper we argue that, despite the conceptual appeal of these simple problems, evaluating an algorithm's performance on these tasks can be misleading. In Section 3, we show that *Example 2* is a trivial problem that can be solved via empirical risk minimization (ERM) (Vapnik, 1998), while in Section 4 we show that *Example 3* can become a much harder problem than intended. Towards resolving these issues, we propose an enhanced version of the Linear Unit-Tests in Section 5. We find that, to the best of our hyper-parameters search and within the set of algorithms evaluated, AND-mask (Parascandolo et al., 2020) is the only algorithm that effectively solves these new problems. To support our findings on synthetic data, we evaluate AND-mask on a more realistic image classification task in Section 6 and find that it significantly outperforms ERM when spurious correlations are introduced.

2. Preliminaries

We consider n_{env} environments; for each environment $e \in \mathcal{E} = \{E_j\}_{j=1}^{n_{\text{env}}}$ we denote by $\mathcal{D}^e = \{x_i^e, y_i^e\}_{i=1}^{n_e}$ the corresponding dataset contaning n_e samples. The input feature vector $x^e = (x^e_{\text{inv}}, x^e_{\text{spu}}) \in \mathbb{R}^d$ contains features $x^e_{\text{inv}} \in \mathbb{R}^{d_{\text{inv}}}$ that elicits invariant correlations as well as features $x^e_{\text{spu}} \in \mathbb{R}^{d_{\text{spu}}}$ that elicits spurious correlations. Our goal is to construct invariant predictors that estimate y^e by relying on x^e_{inv} and ignoring x^e_{spu}. We provide here a brief overview of the algorithms compared in this work:

- **Empirical Risk Minimization** (Vapnik, 1998) minimizes the error on the union of all the training splits.

[1] ETH Zürich [2] OpenAI (this work was done while GP was at ETH Zürich). Correspondence to: Piersilvio De Bartolomeis <piersilvio.debartolomeis@inf.ethz.ch>.

Published at the ICML 2022 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA. Copyright 2022 by the author(s).

- **Invariant Risk Minimization** (Arjovsky et al., 2019) finds a representation of the features such that there exists a classifier, on top of that representation, that is simultaneously optimal for all environments.
- **Inter-environmental Gradient Alignment** (Koyama & Yamaguchi, 2020) minimizes the error on the training splits while reducing the variance of the gradient of the loss per environment.
- **AND-mask** (Parascandolo et al., 2020) minimizes the error on the training splits by updating the model on those directions where the sign of the gradient of the loss is the same for most environments.
- **Information Bottleneck Invariant Risk Minimization** (Ahuja et al., 2021) finds the predictor with the least entropy among all the highly predictive invariant predictors.
- **Oracle** (Aubin et al., 2021) is a version of ERM where features x^e_{spu} in the train set are shuffled at random across examples, hence spurious features are trivial to ignore.

3. Example 2: Cows versus Camels

We consider *Example 2* from the linear unit-tests of Aubin et al. (2021) and prove that ERM, as well as all other algorithms considered in this paper, can solve this problem. We report the data generating process here:

$$\mu_{\text{cow}} = 1_{d_{\text{inv}}}, \quad \mu_{\text{camel}} = -\mu_{\text{cow}}, \quad \nu_{\text{animal}} = 10^{-2},$$
$$\mu_{\text{grass}} = 1_{d_{\text{spu}}}, \quad \mu_{\text{sand}} = -\mu_{\text{grass}}, \quad \nu_{\text{bg}} = 1,$$

where $1_m \in \mathbb{R}^m$ denotes a vector of ones. To construct the datasets $\mathcal{D}_e$ for every $e \in \mathcal{E}$ we sample:

$$j^e \sim \text{Categorical}\left(f^e s^e, b^e s^e f^e\left(1-s^e\right), b^e\left(1-s^e\right)\right),$$
$$x^e_{\text{inv}} \sim \begin{cases} \left(\mathcal{N}_{d_{\text{inv}}}(0,0.1) + \mu_{\text{cow}}\right) \cdot \nu_{\text{animal}} & \text{if } j^e \in \{1,2\} \\ \left(\mathcal{N}_{d_{\text{inv}}}(0,0.1) + \mu_{\text{camel}}\right) \cdot \nu_{\text{animal}} & \text{if } j^e \in \{3,4\} \end{cases},$$
$$x^e_{\text{spu}} \sim \begin{cases} \left(\mathcal{N}_{d_{\text{spu}}}(0,0.1) + \mu_{\text{grass}}\right) \cdot \nu_{\text{bg}} & \text{if } j^e \in \{1,4\} \\ \left(\mathcal{N}_{d_{\text{spu}}}(0,0.1) + \mu_{\text{sand}}\right) \cdot \nu_{\text{bg}} & \text{if } j^e \in \{2,3\} \end{cases},$$
$$y^e \leftarrow \begin{cases} 1 & \text{if } 1^\intercal_{d_{\text{inv}}} x^e_{\text{inv}} > 0 \\ 0 & \text{else} \end{cases},$$

where for the first three environments the background probabilities are $b^{e_0} = 0.95, b^{e_1} = 0.97$, $b^{e_2} = 0.99$ and the animal probabilities are $s^{e_0} = 0.3, s^{e_1} = 0.5, s^{e_2} = 0.7$. When the number of environments are greater than three, then $f^{e_j} \sim \text{Uniform}\,(0.9, 1)$, and $s^{e_j} \sim \text{Uniform}\,(0.3, 0.7)$.

Proposition 1. *(Informal) Consider the setting of Example 2. In this scenario, ERM, IRM, IB-IRM and AND-Mask solve the OOD generalization problem.*

The proof is a direct consequence of the results in (Ahuja et al., 2021) and can be found in the appendix.

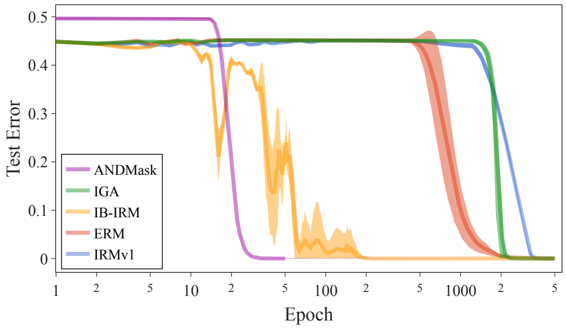


Figure 1. We plot mean and standard deviation over 5 runs for the average test error for all algorithms with $(d_{\text{inv}}, d_{\text{spu}}, n_{\text{env}}) = (5, 5, 3)$. All algorithms hyper-parameters are tuned to yield the best performance. Details in the appendix.

It should be noted that most of the experimental findings in the literature (Ahuja et al., 2021; Khezeli et al., 2021; Koyama & Yamaguchi, 2020; Du et al., 2021; Wang et al., 2022; Nguyen et al., 2022) fail to achieve satisfactory performance in this example. The issue lies with the hyperparameter search proposed in Aubin et al. (2021), in particular, the learning rates tested are too small for convergence. In light of Proposition 1, we argue that all the algorithms evaluated in the linear unit-tests should solve *Example 2*. We run experiments with higher learning rates and all the algorithms reach zero test error. We report these results in Figure 1, notably ANDMask converges significantly faster than all the other algorithms — even after tuning all hyperparameters (see Appendix A.1). Moreover, ERM achieves zero error, making the task trivial for the other algorithms. Overall, these results suggest that *Example 2*, in its current form, is not effective for evaluating invariance discovery algorithms. In Section 5, we address this issue and propose a more challenging version of the problem where an algorithm is ranked based on how much weight it puts on spurious features.

4. Example 3: Small Invariant Margin

We consider *Example 3* from the linear unit-tests of Aubin et al. (2021). This problem is a linear version of a classification experiment in Parascandolo et al. (2020) where the spurious correlations provide a shortcut in minimizing the training error while the invariant classifier takes a more complex form. Even though AND-mask was shown to solve the non-linear version of this problem in Parascandolo et al. (2020), it has been experimentally observed that it fails to solve its linear counter-part. In Section 4.2 we resolve this contradiction and provide an explanation for the failure of AND-mask observed in Aubin et al. (2021); Ahuja et al. (2021).

] Following Aubin et al. (2021), *Example 3* is generated as follows: $\gamma = 0.1 \cdot 1_{d_{\text{inv}}}$, and $\mu^e \sim \mathcal{N}_{d_{\text{spu}}}(0, 1)$, for all environments. To construct the datasets $\mathcal{D}_e$ for every $e \in \mathcal{E}$

we sample:

$$\mu^e \sim \mathcal{N}_{d_{\text{spu}}}(0,1),$$

$$y^e \sim \text{Bernoulli}\left(\frac{1}{2}\right),$$

$$x^e_{\text{inv}} \sim \begin{cases} \mathcal{N}_{d_{\text{inv}}}\left(+\gamma, 10^{-1}\right) & \text{if } y^e = 0 \\ \mathcal{N}_{d_{\text{inv}}}\left(-\gamma, 10^{-1}\right) & \text{if } y^e = 1 \end{cases},$$

$$x^e_{\text{spu}} \sim \begin{cases} \mathcal{N}_{d_{\text{spu}}}\left(+\mu^e, 10^{-1}\right) & \text{if } y^e = 0 \\ \mathcal{N}_{d_{\text{spu}}}\left(-\mu^e, 10^{-1}\right) & \text{if } y^e = 1 \end{cases}.$$

4.1. Can AND-mask solve Example 3?

We present a theoretical analysis of AND-mask in the setting of *Example 3*, proving that it can solve the OOD generalization problem when the number of environments is sufficiently large. For the sake of clarity, we consider the two dimensional case, i.e. $d_{\text{inv}} = d_{\text{spu}} = 1$. We assume that the data is balanced and that AND-mask is initialized with weights $w^{(0)} = (0,0)^\top$ — this choice can be interpreted as a prior belief that all features are spurious. Along the lines of Parascandolo et al. (2020) we define for every component $[m_\tau]_j$ of the mask m_τ, $[m_\tau]_j = 1\left[\tau \leqslant \frac{1}{n_{\text{env}}}\left|\sum_e \text{sign}\left([\nabla\mathcal{L}_e(w)]_j\right)\right|\right]$, where $\nabla\mathcal{L}_e$ is the average gradient within an environment. Our goal is to prove that AND-mask converges towards an *invariant* solution, that is a solution of the form:

$$w \in \mathcal{W}^* = \{(w_1, w_2)^\top \in \mathbb{R}^2 : w_2 = 0 \wedge w_1 < 0\}.$$

First, we prove in Theorem 4.1 that the component of the mask m_τ corresponding to the spurious feature is 0 with high probability when $w = w^{(0)}$.

Theorem 4.1. *Let $\mathcal{W} = \{(w_1, w_2)^\top \in \mathbb{R}^2 : w_2 = 0\}$ and $[m_\tau]_{spu}$ be the component of the mask corresponding to the spurious feature. If $w \in \mathcal{W}$ then:*

$$\mathbb{P}\left([m_\tau]_{spu} = 0\right) \geq 1 - \exp\left(-\frac{\tau^2 n_{env}}{2}\right). \quad (1)$$

It follows that, if n_{env} is big enough, $[w]_{\text{spu}}$ will likely never be updated during gradient descent, i.e

$$\forall k : w^{(k)} \in \mathcal{W} = \{(w_1, w_2)^\top \in \mathbb{R}^2 : w_2 = 0\}.$$

Next, we prove in Theorem 4.2 that the component of m_τ corresponding to the invariant feature converges almost surely to 1 for all incorrect solutions of the form:

$$w \in \widetilde{\mathcal{W}}^* = \{(w_1, w_2)^\top \in \mathbb{R}^2 : w_2 = 0 \wedge w_1 \geq 0\}.$$

Theorem 4.2. *Let $\widetilde{\mathcal{W}}^*$ be the set of incorrect solutions. If $w \in \widetilde{\mathcal{W}}^*$ then, as $n^e \to \infty$, for all $e \in \mathcal{E}$:*

$$[m_\tau]_{inv} \xrightarrow{a.s} 1 \quad (2)$$

Now if $w^{(k)} \in \mathcal{W}^*$ we are done. If instead $w^{(k)} \in \widetilde{\mathcal{W}}^*$ then from Thm. 4.1 and Thm. 4.2, we know that gradient descent will only update the invariant feature and eventually converge towards a correct solution [*]. We can conclude that with n_{env} sufficiently large and proper initialization, AND-mask will only rely on invariant features and converge towards the *invariant* solution of *Example 3*.

4.2. A Tough Problem

Our theoretical analysis suggests that AND-mask should correctly identify invariant signals in the setting of *Example 3* and yet experimental evidence is contradictory (Aubin et al., 2021; Ahuja et al., 2021). We argue that the problem of invariance discovery becomes much harder when constructing the dataset from a small number of environments.

Proposition 2. *Consider the setting of Example 3. We have that the sign of at least one component $[\mu_e]_j$ will be the same in all the environments with probability $p = 1 - (1 - 2^{1-n_{env}})^{d_{spu}}$.*

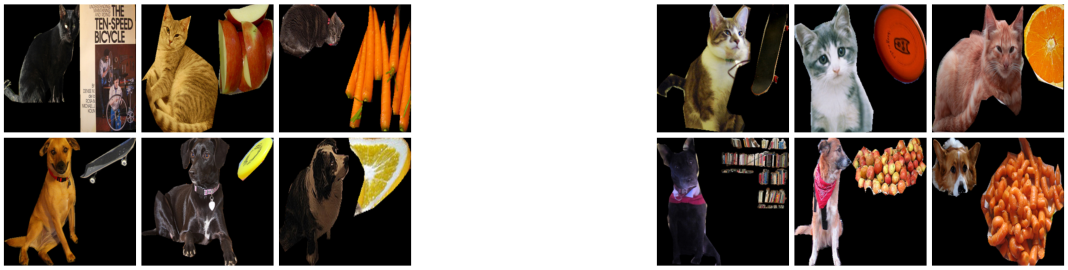
Note that, when $\text{sign}([\mu_e]_j)$ is constant over all the environments it becomes much hard to distinguish spurious signals from invariant signals. Now, if we consider the setting $(d_{\text{inv}}, d_{\text{spu}}, n_{\text{env}}) = (5, 5, 3)$ adopted in Ahuja et al. (2021); Aubin et al. (2021), with probability $p \approx 0.97$ at least one spurious dimension will have constant sign across environments, hence explained why no algorithm could solve the problem with these settings. In Section 5, we propose a modified version of *Example 3* where we force the sign of spurious signals to be inconsistent in at least one environment.

5. Enhancing Linear Unit-Tests

One way to make *Example* 2 more challenging would be to change the support of the spurious features distribution at test-time. In that case, both ERM and IRM fail to solve the OOD generalization problem as proved in Thm. 3 (Insufficiency) of Ahuja et al. (2021). However, changing the support of the spurious features does not provide a provable guarantee of robustess; instead, we propose to rank algorithms based on how much weight they place on spurious features. We denote by w_{spu} the components of the weight vector corresponding to the spurious features. Note that, if $\|w_{\text{spu}}\|_\infty > 0$ we can always find a distribution shift that would make the algorithm fail. We consider the task to be solved if $\|w_{\text{spu}}\|_\infty = 0$.

We run experiments in this new scenario and report the results in Figure 3. To the best of our hyper-parameters search, AND-mask is the only algorithm that solves this

[*] As the invariant component of the gradient for $w \in \widetilde{\mathcal{W}}^*$ is always positive (see Appendix A.3)



(a) Train dataset

(b) Test dataset

Figure 2. In the train dataset (a) Cat is associated with book, carrot and apple while Dog is associated with skateboard, frisbee and orange. In the test dataset (b) Cat is associated with skateboard, frisbee and orange while Dog is associated with book, carrot and apple.

example correctly, IB-IRM comes close but it could still fail for some significant distribution shift of the spurious features. The remaining algorithms fail to solve this task.

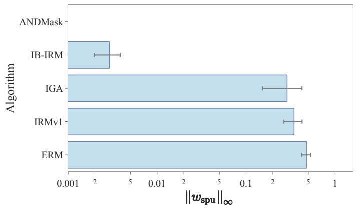


Figure 3. We plot mean and standard deviation over 5 runs of $\|w_{\text{spu}}\|_\infty$ for all algorithms for $(d_{\text{inv}}, d_{\text{spu}}, n_{\text{env}}) = (5, 5, 3)$. All algorithms hyper-parameters are tuned to yield the best performance. Details in the appendix.

As for *Example 3*, we propose a modified version of the problem to complement the more challenging one proposed by Aubin et al. (2021). In particular, to mitigate the issue raised in Prop. 2 we force the sign of the spurious mean in the last environment to be the opposite of the sign of the spurious mean in the first environment, i.e. $\text{sign}(\mu^{n_{\text{env}}}) = -\text{sign}(\mu^1)$. This simple modification introduces an asymmetry in the dataset which makes it possible for an algorithm to identify, in principle, the invariant features — in contrast to the data generation process outlined in Section 4.2. We run experiments for this modified version of *Example 3* and report the results in Table 1. We observe that, to the best of our hyper-parameters search, only AND-mask can effectively solve the OOD generalization problem.

Table 1. Average test errors for all algorithms for $(d_{\text{inv}}, d_{\text{spu}}, n_{\text{env}}) = (5, 5, 3)$. All algorithms hyper-parameters are tuned to yield the best performance. Details in the appendix.

METHOD	EXAMPLE3	SOLVED?
AND-MASK	0.01 ± 0.00	✓
ERM	0.38 ± 0.16	✗
IB-IRM	0.39 ± 0.05	✗
IGA	0.37 ± 0.18	✗
IRMV1	0.37 ± 0.18	✗
ORACLE	0.01 ± 0.00	✓

6. Image Classification Task

To further assess AND-mask's performance under more realistic conditions, we construct a binary classification problem for images with spurious correlations. The task is to correctly classify dog vs. cat; at training-time each class is associated with n spurious objects and at test-time these associations are reversed. In Figure 2 we show a few examples of the dataset for $n = 3$. The main challenge is learning a classifier that does not rely on spurious features, i.e. the associated objects, and thus can generalize well in the test environment. We extract supervised representations from the last layer of BiT-M-R152 (Kolesnikov et al., 2020). Then, we train a linear classifier on top of these representations using both AND-mask and ERM. We report the results of this experiment in Table 2. Overall, AND-mask consistently outperforms ERM — it is worthwhile to note how the gap in performance between AND-mask and ERM increases once spurious associations are introduced.

Table 2. Test accuracy for AND-mask and ERM when 0 and 3 spurious objects are associated with Dog and Cat.

METHOD	SPURIOUS OBJECTS	TEST ACCURACY
AND-MASK	0	**0.95**
ERM	0	0.95
AND-MASK	3	**0.80**
ERM	3	0.70

7. Conclusions

Out of distribution generalization is one of the most challenging issues in machine learning, and properly evaluating an algorithm's performance in this context is essential. This paper extends the original set of problems proposed by Aubin et al. (2021) with the purpose of providing a more fine-grained evaluation of invariance discovery algorithms. We hope researchers will continue to extend this suite of tests to learn more about the strengths and weaknesses of new algorithms in a transparent and standardized manner.

References

Ahuja, K., Caballero, E., Zhang, D., Bengio, Y., Mitliagkas, I., and Rish, I. Invariance principle meets information bottleneck for out-of-distribution generalization, 2021. URL https://arxiv.org/abs/2106.06607.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization, 2019. URL https://arxiv.org/abs/1907.02893.

Aubin, B., Słowik, A., Arjovsky, M., Bottou, L., and Lopez-Paz, D. Linear unit-tests for invariance discovery, 2021. URL https://arxiv.org/abs/2102.10867.

Beery, S., van Horn, G., and Perona, P. Recognition in terra incognita, 2018. URL https://arxiv.org/abs/1807.04975.

Collobert, R. and Weston, J. A unified architecture for natural language processing: deep neural networks with multitask learning. In *ICML '08*, 2008.

Du, X., Ramamoorthy, S., Duivesteijn, W., Tian, J., and Pechenizkiy, M. Beyond discriminant patterns: On the robustness of decision rule ensembles, 2021. URL https://arxiv.org/abs/2109.10432.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015.

Khezeli, K., Blaas, A., Soboczenski, F., Chia, N., and Kalantari, J. On invariance penalties for risk minimization, 2021. URL https://arxiv.org/abs/2106.09777.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. Big transfer (bit): General visual representation learning, 2020.

Koyama, M. and Yamaguchi, S. When is invariance useful in an out-of-distribution generalization problem ?, 2020. URL https://arxiv.org/abs/2008.01883.

Nguyen, T., Lyu, B., Ishwar, P., Scheutz, M., and Aeron, S. Conditional entropy minimization principle for learning domain invariant representation features, 2022. URL https://arxiv.org/abs/2201.10460.

Parascandolo, G., Neitz, A., Orvieto, A., Gresele, L., and Schölkopf, B. Learning explanations that are hard to vary, 2020. URL https://arxiv.org/abs/2009.00329.

Vapnik, V. Statistical learning theory. *Wiley*, 1998.

Wang, H., Si, H., Li, B., and Zhao, H. Provable domain generalization via invariant-feature subspace recovery, 2022. URL https://arxiv.org/abs/2201.12919.

A. Appendix

A.1. Experiments details

We used the published code of Arjovsky et al. (2019) to conduct all experiments. We follow the same protocol as prescribed in Aubin et al. (2021) for model selection, hyper-parameter selection, training, and evaluation. For all the examples, the models used are linear. The training loss is the binary cross-entropy for the classification setting. For AND-mask we use Adam optimizer while for all the other algorithms we use GD without any batching.

Hyper-parameter tuning for Figure 1 and Figure 3. We run 100 hyper-parameter queries for each model with 5 data seeds. AND-mask treats every data point as coming from its own environment and we initialize all the weights of the AND-mask network at 0 [†].

Below we report how the search over the hyper-parameter is performed.

METHOD	LEARNING RATE	WEIGHT DECAY	λ_{IRM}	λ_{IB}	τ	IGA PENALTY
ERM	$10^{\text{Uniform}(-2,-1)}$	$10^{\text{Uniform}(-6,-2)}$	✗	✗	✗	✗
IRMV1	$10^{\text{Uniform}(-2,-1)}$	$10^{\text{Uniform}(-6,-2)}$	$1-10^{\text{Uniform}(-3,-0.3)}$	✗	✗	✗
IB-IRM	$10^{\text{Uniform}(-2,-1)}$	$10^{\text{Uniform}(-6,-2)}$	$1-10^{\text{Uniform}(-3,-0.3)}$	$1-10^{\text{Uniform}(-2,0)}$	✗	✗
AND-MASK	$10^{\text{Uniform}(-2,-1)}$	$10^{\text{Uniform}(-6,-2)}$	✗	✗	$\text{Uniform}(0.8, 1.0)$	✗
IGA	$10^{\text{Uniform}(-2,-1)}$	$10^{\text{Uniform}(-6,-2)}$	✗	✗	✗	$10^{\text{Uniform}(1,5)}$

Hyper-parameter tuning for Table 1. We run 150 hyper-parameter queries and average over 3 data seeds. Below we report how the search over the hyper-parameter is performed.

METHOD	LEARNING RATE	WEIGHT DECAY	λ_{IRM}	λ_{IB}	τ	IGA PENALTY
ERM	$10^{\text{Uniform}(-3,-1)}$	$10^{\text{Uniform}(-6,-2)}$	✗	✗	✗	✗
IRMV1	$10^{\text{Uniform}(-3,-1)}$	$10^{\text{Uniform}(-6,-2)}$	$1-10^{\text{Uniform}(-3,-0.3)}$	✗	✗	✗
IB-IRM	$10^{\text{Uniform}(-3,-1)}$	$10^{\text{Uniform}(-6,-2)}$	$1-10^{\text{Uniform}(-3,-0.3)}$	$1-10^{\text{Uniform}(-2,0)}$	✗	✗
AND-MASK	$10^{\text{Uniform}(-3,-1)}$	$10^{\text{Uniform}(-6,-2)}$	✗	✗	$\text{Uniform}(0.4, 0.8)$	✗
IGA	$10^{\text{Uniform}(-3,-1)}$	$10^{\text{Uniform}(-6,-2)}$	✗	✗	✗	$10^{\text{Uniform}(1,5)}$

Image Classification Task. Note that there is no standard notion of environments here, which is why AND-mask treats every example as coming from its own environment. This assumption is not unreasonable, as every image in the dataset was literally collected in a different physical environment. We train both AND-mask and ERM using GD without any batching.

A.2. Example 2 Proofs

We provide here a proof for Proposition 1.

Proof. Example 2 follows the linear classification SEM (FIIF) with zero noise from Assumption 2 in Ahuja et al. (2021).The invariant features are strictly separable, bounded and satisfy support overlap; the spurious features are bounded and satisfy support overlap. It follows from Theorem 3 (Sufficiency) in Ahuja et al. (2021) that both ERM and IRM solve the OOD generalization problem. As ERM and IRM are edge cases respectively of ANDMask and IB-IRM, for $\tau = 0$ and $\gamma = 0$, it follows that they can both solve the OOD generalization problem as well. □

A.3. Example 3 Proofs

We consider datasets $\{\mathcal{D}^e\}_{e\in\mathcal{E}}$, and $\mathcal{D}^e = (x_i^e, y_i^e), i_e = 1, \ldots, n^e$. Here $x_i^e \in \mathcal{X} \subseteq \mathbb{R}^m$ is the vector containing the observed inputs, $w \in \mathbb{R}^m$ is the vector of weights and $y_i^e \in \{0, 1\}$ are the targets. The superscript $e \in \mathcal{E}$ can be

[†] This can be interpreted as a prior belief that all features are spurious. For a fair comparison, we also initialized the weights of all the other algorithms at 0. We repeated the sweep over hyper-parameters and found no evidence of an improvement in performance (e.g. speed of convergence or $||\, w_{\text{spu}} \,||_\infty$).

interpreted as an environment label. We define $\alpha_w \in \mathbb{R}^d$ as a (signed) measure of gradient agreement across environments $[\alpha_w]_i \coloneqq \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \operatorname{sign}([\nabla\mathcal{L}_e(w)]_i)$.

We report here the data generating process of Example 3. Let $\gamma = 0.1 \cdot 1_{d_{\text{inv}}}$, and $\mu^e \sim \mathcal{N}_{d_{\text{spu}}}(0,1)$, for all environments. To construct the datasets $\mathcal{D}_e$ for every $e \in \mathcal{E}$ we sample:

$$\begin{aligned}
y^e &\sim \text{Bernoulli}\left(\frac{1}{2}\right),\\
x^e_{\text{inv}} &\sim \begin{cases} \mathcal{N}_{d_{\text{inv}}}\left(+\gamma, 10^{-1}\right) & \text{if } y^e = 0\\ \mathcal{N}_{d_{\text{inv}}}\left(-\gamma, 10^{-1}\right) & \text{if } y^e = 1\end{cases}\\
x^e_{\text{spu}} &\sim \begin{cases} \mathcal{N}_{d_{\text{spu}}}\left(+\mu^e, 10^{-1}\right) & \text{if } y^e = 0\\ \mathcal{N}_{d_{\text{spu}}}\left(-\mu^e, 10^{-1}\right) & \text{if } y^e = 1\end{cases}
\end{aligned}$$

We analyse learning by minimizing an empirical loss of the form:

$$\mathcal{L}(w) \coloneqq -\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \overbrace{\frac{1}{n^e}\sum_{(x_i^e, y_i^e)\in\mathcal{D}^e} y_i^e \log\left(\sigma\left(w^\top x_i^e\right)\right) + (1-y_i^e)\log\left(1-\sigma\left(w^\top x_i^e\right)\right)}^{\coloneqq\mathcal{L}_e(w)}. \tag{3}$$

The following lemma characterizes the asymptotic behaviour of $\nabla\mathcal{L}_e$ as $n^e \to \infty$:

Lemma A.1. *Let $P_{(\theta_1,\theta_2)}$ be the distribution of $\mathcal{N}\left([\theta_1,\theta_2]^T, I\sigma^2\right)$. If $\{\mathcal{D}^e\}_{e\in\mathcal{E}}$ are sampled according to Example 3 then the following holds:*

$$\nabla\mathcal{L}_e(w) \xrightarrow{a.s.} \frac{1}{2}\left(\begin{bmatrix}\gamma\\ \mu_e\end{bmatrix} + \mathop{\mathbb{E}}_{x\sim P_{(\gamma,\mu_e)}}\left[\sigma(w^\top x)\cdot x\right] + \mathop{\mathbb{E}}_{x\sim P_{-(\gamma,\mu_e)}}\left[\sigma(w^\top x)\cdot x\right]\right) \tag{4}$$

Proof. It suffices to apply the law of large numbers, i.e. sums converge to expectations:

$$\begin{aligned}
\nabla\mathcal{L}_e(w) &= \frac{1}{n^e}\sum_i \left[\sigma(w^\top x_i^e) - y_i^e\right]\cdot x_i^e &(5)\\
&= \frac{1}{2}\left(\frac{2}{n^e}\sum_{i:y_i^e=0}\sigma(w^\top x_i^e)\cdot x_i^e + \frac{2}{n^e}\sum_{i:y_i^e=1}\sigma(w^\top x_i^e)\cdot x_i^e - \frac{2}{n^e}\sum_{i:y_i^e=1} x_i^e\right) &(6)\\
&\xrightarrow{a.s.} \frac{1}{2}\left(\begin{bmatrix}\gamma\\ \mu_e\end{bmatrix} + \mathop{\mathbb{E}}_{x\sim P_{(\gamma,\mu_e)}}\left[\sigma(w^\top x)\cdot x\right] + \mathop{\mathbb{E}}_{x\sim P_{-(\gamma,\mu_e)}}\left[\sigma(w^\top x)\cdot x\right]\right) &(7)
\end{aligned}$$

This concludes the proof. □

We now present the proof of Theorem 4.1.

Proof. We will denote with $[\nabla\mathcal{L}_e(w)]_{\text{spu}}$ the components of the gradient corresponding to the spurious features. Similarly $[x]_{\text{inv}}$ and $[x]_{\text{spu}}$ will denote the components of the feature vector correspoding to invariant and spurious features, respectively. From assumptions we have $w = (k, 0)^\top$ for some $k \in \mathbb{R}$. Applying Lemma A.1 we have that:

$$\begin{aligned}
[\nabla\mathcal{L}_e(w)]_{\text{spu}} &\xrightarrow{a.s.} \frac{1}{2}\left(\mu_e + \mathop{\mathbb{E}}_{x\sim P_{(\gamma,\mu_e)}}\left[\sigma(k\cdot[x]_{\text{inv}})\cdot[x]_{\text{spu}}\right] + \mathop{\mathbb{E}}_{x\sim P_{-(\gamma,\mu_e)}}\left[\sigma(k\cdot[x]_{\text{inv}})\cdot[x]_{\text{spu}}\right]\right) &(8)\\
&= \frac{1}{2}\left(\mu_e + \mathop{\mathbb{E}}_{x\sim P_\gamma}\left[\sigma(k\cdot x)\right]\cdot\mathop{\mathbb{E}}_{x\sim P_{\mu_e}}[x] + \mathop{\mathbb{E}}_{x\sim P_{-\gamma}}\left[\sigma(k\cdot x)\right]\cdot\mathop{\mathbb{E}}_{x\sim P_{-\mu_e}}[x]\right) &(9)\\
&= \frac{\mu_e}{2}\left(1 + \mathop{\mathbb{E}}_{x\sim P_\gamma}\left[\sigma(k\cdot x)\right] - \mathop{\mathbb{E}}_{x\sim P_{-\gamma}}\left[\sigma(k\cdot x)\right]\right). &(10)
\end{aligned}$$

Next, the fact that $\sigma(x) = 1 - \sigma(-x)$ and $x \sim P_{-\gamma} \implies -x \sim P_{\gamma}$, we get

$$
\begin{aligned}
[\nabla \mathcal{L}_e(w)]_{\text{spu}} &\xrightarrow{a.s.} \frac{\mu_e}{2}\left(\mathop{\mathbb{E}}_{x\sim P_\gamma}[\sigma(k\cdot x)] + \mathop{\mathbb{E}}_{x\sim P_{-\gamma}}[1-\sigma(k\cdot x)]\right) && (11)\\
&= \frac{\mu_e}{2}\left(\mathop{\mathbb{E}}_{x\sim P_\gamma}[\sigma(k\cdot x)] + \mathop{\mathbb{E}}_{x\sim P_{-\gamma}}[\sigma(-k\cdot x)]\right) && (12)\\
&= \frac{\mu_e}{2}\left(\mathop{\mathbb{E}}_{x\sim P_\gamma}[\sigma(k\cdot x)] + \mathop{\mathbb{E}}_{x\sim P_{\gamma}}[\sigma(k\cdot x)]\right) && (13)\\
&= \mu_e \cdot \mathop{\mathbb{E}}_{x\sim P_\gamma}[\sigma(k\cdot x)]\,. && (14)
\end{aligned}
$$

It follows that $\text{sign}\left([\nabla\mathcal{L}_e(w)]_{spu}\right)$ does not depend on w:

$$
\text{sign}\left([\nabla\mathcal{L}_e(w)]_{spu}\right) \xrightarrow{a.s.} \text{sign}\left(\mu_e \cdot \overbrace{\mathop{\mathbb{E}}_{x\sim P_\gamma}[\sigma(k\cdot x)]}^{>0}\right) = \text{sign}\left(\mu_e\right). \tag{15}
$$

Moreover, we have that for spurious signals α_w converges to:

$$
[\alpha_w]_{spu} \xrightarrow{a.s.} \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \text{sign}\left(\mu_e\right). \tag{16}
$$

Now, recall that $\mu_e \sim \mathcal{N}(0,1)$ are independent and normally distributed. Therefore $\text{sign}(\mu_1), \ldots, \text{sign}(\mu_{n_{\text{env}}})$ are independent bounded random variables with $\text{sign}(\mu_i) \in \{-1, 1\}$ for all i. We can bound this sum using Hoeffding's inequality:

$$
\mathbb{P}\left(|\,[\alpha_w]_{\text{spu}}\,| \geq \epsilon\right) \leq \exp\left(-\frac{\epsilon^2 n_{\text{env}}}{2}\right). \tag{17}
$$

and it follows that:

$$
\mathbb{P}\left([m_\tau]_{\text{spu}} = 0\right) = \mathbb{P}\left(|\,[\alpha_w]_{\text{spu}}\,| \geq \tau\right) \leq \exp\left(-\frac{\tau^2 n_{\text{env}}}{2}\right). \tag{18}
$$

□

Next, we provide a proof for Theorem 4.2.

Proof. First, we prove the statement for $w_0 = (0,0)^\top$. Applying Lemma A.1 we have that:

$$
\begin{aligned}
[\nabla\mathcal{L}_e(w_0)]_{\text{inv}} &\xrightarrow{a.s.} \frac{1}{2}\left(\gamma + \mathop{\mathbb{E}}_{x\sim P_{(\gamma,\mu_e)}}\left[\sigma(w_0^\top x)\cdot[x]_{\text{inv}}\right] + \mathop{\mathbb{E}}_{x\sim P_{-(\gamma,\mu_e)}}\left[\sigma(w_0^\top x)\cdot[x]_{\text{inv}}\right]\right) && (19)\\
&= \frac{1}{2}\left(\gamma + \frac{1}{2}\mathop{\mathbb{E}}_{x\sim P_\gamma}[x] + \frac{1}{2}\mathop{\mathbb{E}}_{x\sim P_{-\gamma}}[x]\right) && (20)\\
&= \frac{\gamma}{2}. && (21)
\end{aligned}
$$

Then it follows that:

$$
[\alpha_{w_0}]_{\text{inv}} = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \text{sign}([\nabla\mathcal{L}_e(w_0)]_{\text{inv}}) \xrightarrow{a.s.} \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \text{sign}\left(\frac{\gamma}{2}\right) = 1. \tag{22}
$$

Now, to prove that $[\alpha_w]_{\text{inv}} \xrightarrow{a.s.} 1$ for all $w \in \widetilde{\mathcal{W}}^* = \{(w_1, w_2)^\top \in \mathbb{R}^2 : w_2 = 0 \wedge w_1 \geq 0\}$ it suffices to show that $[\nabla\mathcal{L}_e(w_0)]_{\text{inv}} \leq [\nabla\mathcal{L}_e(w)]_{\text{inv}}$ for all $w \in \widetilde{\mathcal{W}}^*$. It is enough to prove that the first derivative of $\left[\nabla\mathcal{L}_e((w,0)^\top)\right]_{\text{inv}}$ is always positive:

$$
\begin{aligned}
\frac{d}{dw}\left[\nabla \mathcal{L}_e((w,0)^\top)\right]_{\text{inv}} &\xrightarrow{a.s.} \frac{1}{2}\frac{d}{dw}\left(\gamma + \mathop{\mathbb{E}}_{x\sim P_\gamma}\left[\sigma(w\cdot x)\cdot x\right] + \mathop{\mathbb{E}}_{x\sim P_{-\gamma}}\left[\sigma(w\cdot x)\cdot x\right]\right) && (23)\\
&= \frac{1}{2}\left(\mathop{\mathbb{E}}_{x\sim P_\gamma}\left[\frac{d}{dw}\sigma(w\cdot x)\cdot x\right] + \mathop{\mathbb{E}}_{x\sim P_{-\gamma}}\left[\frac{d}{dw}\sigma(w\cdot x)\cdot x\right]\right) && (24)\\
&= \frac{1}{2}\left(\mathop{\mathbb{E}}_{x\sim P_\gamma}\left[\sigma(w\cdot x)(1-\sigma(w\cdot x))\cdot x^2\right] + \mathop{\mathbb{E}}_{x\sim P_{-\gamma}}\left[\sigma(w\cdot x)(1-\sigma(w\cdot x))\cdot x^2\right]\right) && (25)\\
&\geq 0. && (26)
\end{aligned}
$$

Finally, to conclude the proof we have: $[m_\tau]_{\text{inv}} = 1\left[\tau \leq \mid [\alpha_w]_{\text{inv}} \mid\right] \xrightarrow{a.s.} 1.$ □

Finally, we provide the proof of Proposition 2.

Proof.

$$
\begin{aligned}
&P\{\exists j : \forall (e,e') \ : \text{sign}([\mu_e]_j) = \text{sign}([\mu_{e'}]_j)\} && (27)\\
&= 1 - P\{\forall j\ \exists (e,e') \ : \text{sign}([\mu_e]_j) \neq \text{sign}([\mu_{e'}]_j)\} && (28)\\
&= 1 - \prod_{j=1}^{d_{\text{spu}}} P\{\exists (e,e') \ : \text{sign}([\mu_e]_j) \neq \text{sign}([\mu_{e'}]_j)\} && (29)\\
&= 1 - \prod_{j=1}^{d_{\text{spu}}} 1 - P\{\forall (e,e') \ : \text{sign}([\mu_e]_j) = \text{sign}([\mu_{e'}]_j)\} && (30)\\
&= 1 - \prod_{j=1}^{d_{\text{spu}}} 1 - 2^{1-n_{\text{env}}} = 1 - (1 - 2^{1-n_{\text{env}}})^{d_{\text{spu}}}. && (31)
\end{aligned}
$$

□