Environment Agnostic Invariant Risk Minimization for Classification of Sequential Datasets

Praveen Venkateswaran
University of California Irvine
praveenv@uci.edu

Vinod Muthusamy
IBM Research
vmuthus@us.ibm.com

Vatche Isahagian
IBM Research
vatchei@ibm.com

Nalini Venkatasubramanian
University of California Irvine
nalini@ics.uci.edu

ABSTRACT

The generalization of predictive models that follow the standard risk minimization paradigm of machine learning can be hindered by the presence of spurious correlations in the data. Identifying invariant predictors while training on data from multiple environments can influence models to focus on features that have an invariant causal relationship with the target, while reducing the effect of spurious features. Such invariant risk minimization approaches heavily rely on clearly defined environments and data being perfectly segmented into these environments for training. However, in real-world settings, perfect segmentation is challenging to achieve and these *environment-aware* approaches prove to be sensitive to segmentation errors. In this work, we present an *environment-agnostic* approach to develop generalizable models for classification tasks in sequential datasets without needing prior knowledge of environments. We show that our approach results in models that can generalize to out-of-distribution data and are not influenced by spurious correlations. We evaluate our approach on real-world sequential datasets from various domains.

CCS CONCEPTS

• **Computing methodologies → Neural networks**; **Supervised learning by classification**.

KEYWORDS

out-of-distribution generalization, robust models, sequential prediction

ACM Reference Format:
Praveen Venkateswaran, Vinod Muthusamy, Vatche Isahagian, and Nalini Venkatasubramanian. 2021. Environment Agnostic Invariant Risk Minimization for Classification of Sequential Datasets. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3447548.3467324

1 INTRODUCTION

Machine learning models have been incorporated in multiple application domains. While the increased adoption has led to considerable success, there have been numerous examples of the brittleness of machine learning models in generalizing to out-of-distribution data. This is partly due to being influenced by spurious correlations and data biases that fail to hold outside training data distributions [34, 41]. A classic example was highlighted by Beery et al. [5] where a model, trained to classify images of cows in pastures and camels in the desert, failed when the backgrounds were switched because it was influenced by the spurious correlation (i.e., green pastures with cows and sandy deserts with camels) rather than relying on the invariant features (i.e., the cows and camels themselves).

There has been an increasing effort to improve the generalization of these models to out-of-distribution data using different approaches like meta-learning [6], adversarial learning [2], feature representations [45], among others.

Invariant Risk Minimization (IRM) is a framework recently proposed by Arjovsky et al. [4] that takes a different approach to the problem of model generalization. It assumes that the training data comes from multiple environments and that features whose distributions vary across the environments in the training data are likely to also vary between the training and test datasets and hence should be treated as spurious correlations. IRM identifies these spurious features and learns robust predictors by exploiting the varying degrees of spurious correlations present in the environments. Examples of environments can include images taken from different geographic regions, sensor readings from different types of sensors, or loans processed by different departments. The goal of IRM is to find a data representation such that the optimal classifier over this representation is identical or invariant over the training environments. There have been several extensions to the IRM framework. For instance Ahuja et al. [1] propose a game theoretic approach to IRM, while Krueger et al. [20] introduce the notion of risk extrapolation to encourage strict equality between training risks.

While these approaches have resulted in predictors that are effective in out-of-distribution generalization over a variety of datasets, they suffer from two inherent weaknesses: First, they rely on the assumption that the different training environments are known apriori. Second, they require perfect segmentation of the training data into these environments. In practice however, it can be challenging to identify the individual training environments, and there

KDD '21, August 14–18, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8332-5/21/08…$15.00
https://doi.org/10.1145/3447548.3467324

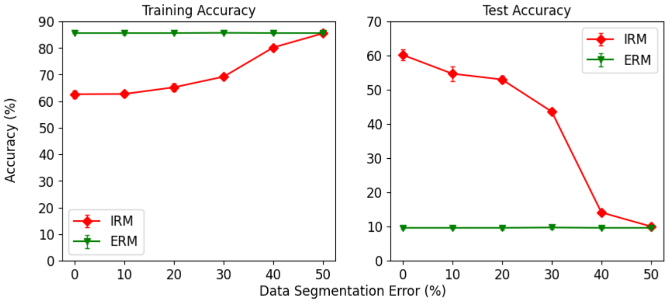


Figure 1: Comparing the sensitivity of Invariant Risk Minimization (IRM) to Empirical Risk Minimization (ERM) for imperfect segmentation of data into environments. The test accuracy of IRM degenerates to that of ERM when the training data cannot be segmented into environments.

can be errors in distinguishing data from different environments resulting in imperfect segmentation of data.

To demonstrate the sensitivity of IRM to imperfect segmentation, we use the Punctuated SST-2 dataset [9]. It consists of sentences and their binary sentiment labels divided into two training environments. A punctuation mark, either a '!' or '.', is introduced as a spurious feature with an 80% and 90% correlation with each of the binary sentiment labels in the two training environments respectively, and only has a 10% correlation in the test environment. Any model influenced by the punctuation feature rather than the sentence while predicting the sentiment, would do well during training but perform poorly at test time. To simulate imperfect segmentation of data into the training environments, we "incorrectly" assign a percentage of examples from the first environment to the second.

Figure 1 shows the resulting out-of-distribution accuracy on the test environment by the IRM model as compared to a standard Empirical Risk Minimization (ERM) model where the ERM model tries to minimize the average loss over all training examples. We observe that with perfect data segmentation (0% error), the IRM model is not influenced by the spurious feature correlation and achieves good generalization unlike the ERM model. However, as the segmentation error increases, its accuracy drops significantly. The IRM model becomes heavily influenced by the spurious punctuation feature, as evidenced by the high training accuracy and low out-of-distribution test accuracy. It converges to the accuracy obtained by ERM when there is no difference in the spurious correlations between the two segmented environments, thus achieving poor generalization.

The example above highlights the drawbacks of such *environment-aware* approaches. In this work, we address the drawbacks in the setting of classification tasks for sequential data. Sequential data is prevalent in many application domains including time-series analysis, natural language processing, click-stream analysis, and business process mining. The data consists of at least one sequential feature, and may also have other features such as metadata, customer information, etc, which can be spuriously correlated with the target variable. Motivated by this insight, we develop an *environment-agnostic* approach to training robust classifiers for sequential data which needs no prior information about environments nor any segmentation. Our approach exploits the structure of sequential data, and extends the IRM framework with a masking function that continually detects and gradually removes spurious features from the model during training, resulting in only the invariant features remaining.

Our contributions can be summarized as follows:

- We present a framework to develop an *Environment-Agnostic Sequential Predictor* (EASP) for classification tasks on sequential data, and formally prove the correctness of this framework.
- To ensure the generalization of EASP, we develop a masking function that exploits the structure of sequential data and variances in spurious correlations to identify invariant features.
- We compare our framework to IRM and ERM on a variety of sequential datasets from real-world domains, and demonstrate through extensive evaluations the significant advantage of EASP over those that require prior knowledge of the training environments.

2 BACKGROUND

Consider a multi-environment sequential dataset consisting of $\mathcal{E} = \{e_1, ..., e_n\}$ environments, each with a data distribution $\mathcal{D}^e$ on $X^e \times Y^e$, where X is the set of input features and Y is the target variable. The dataset contains at least one sequential feature $X^{seq} \subseteq X$, where $X^{seq} \in \mathbb{R}^{1\times d}$.

An invariant feature set X^I, is one where the target prediction probability is consistent across all environments, (i.e.) $p(Y|X_i \in X^I, \mathcal{E})$ is approximately constant. Conversely, the spurious feature set X^S consists of features whose prediction probabilities vary across environments due to the presence of data biases. It follows that $X^I \cup X^S = X$, and $X^I \cap X^S = \emptyset$, (i.e.) a feature cannot be both invariant and spurious.

We define a risk function $\mathcal{R}_e(\theta) : \mathbb{R}^n \longrightarrow \mathbb{R}$ which maps the model parameters θ to the expected loss on $\mathcal{D}^e$ for a given loss function ℓ:

$$\mathcal{R}_e(\theta) = \mathbb{E}_{(x^e,y^e)\sim\mathcal{D}^e}\ \ell(f_\theta(x^e), y^e) \tag{1}$$

where $x^e \in X^e$ and $y^e \in Y^e$, and $\mathcal{R}_i$ refers to the risk or expected loss on the i^{th} environment.

The standard Empirical Risk Minimization (ERM) approach tries to minimize the average loss over all training examples in an environment agnostic manner:

$$\mathcal{R}_{\text{ERM}}(\theta) = \mathbb{E}_{(x,y)\sim\cup_{e\in\mathcal{E}}\mathcal{D}^e}\ell(f_\theta(x), y) \tag{2}$$

While Empirical Risk Minimization has been shown to work well in practice for i.i.d. data [42], it can fail dramatically when test environments and distributions differ significantly from training environments [41].

Invariant Risk Minimization (IRM), proposed by Arjovsky et al. [4], searches for an invariant representation of inputs from different environments. The IRM principle states: *"An invariant representation $\Phi(X)$ is one such that the optimal linear predictor w is the same across all environments $e_i \in \mathcal{E}$"*. They show that finding the invariant predictor, $w \circ \Phi$, requires solving the following bi-level optimization problem:

$$\min_{\Phi, w} \sum_{e \in \mathcal{E}} \mathcal{R}_e(w^\top \Phi(X^e))$$
$$\text{s.t. } w \in \underset{\bar{w}}{\text{argmin}} \mathcal{R}_e(\bar{w}^\top \Phi(X^e)), \quad \forall e \in \mathcal{E}$$

However, since this optimization is highly intractable, particularly when Φ is non-linear, they propose a tractable variant (IRMv1) :

$$\min_{\Phi} \sum_{e} \mathcal{R}_e(\Phi(X^e)) + \lambda \|\nabla_w \mathcal{R}_e(w^\top \Phi(X^e)\|_2^2 \tag{3}$$

where the weights w are initialized to a vector of ones and $\lambda \in [0, \infty)$ is a regularizer that balances between predictive power within an environment (ERM), and the invariance of the predictor across environments.

One approach to determining whether the i^{th} feature X_i is spurious or invariant, is to measure the stability of its parameter weight w_i. Javed et al. [18] show that if X_i is an invariant feature, w_i converges to a fixed magnitude, (i.e.) $\mathbb{E}[Y|X_i] = c$ for some constant value c, across all training iterations. Whereas if $\mathbb{E}[Y|X_i]$ is changing, w_i would keep changing as well, and hence spurious features have parameter weights that exhibit high variance. This definition is equivalent to learning features whose correlations with the target variable are stable.

3 FORMULATION

We leverage the above intuition while developing EASP for sequential data. We first make the following assumption for classification tasks on sequential datasets:

Assumption 1. *A sequence classification task has a non-empty set of sequential features $X^{seq} \subseteq X$ that is predictive of the target variable and is hence invariant with respect to the target Y.*

Note that we do not assume the degree of invariance, and make no assumptions on whether other features are invariant or spurious, and hence the model can still be influenced by spurious correlations.

The assumption of the existence of an invariant feature set for prediction is common and similar to that of Peters et al. [31], but may not apply in all cases. We present empirical results in Section 4.8 showing that our approach still results in a generalized model (c.f. Table 7) when this assumption does not hold and the sequence feature is not predictive.

Based on the intuition outlined in Section 2 – spurious features have weights that exhibit high variance – we define a masking function $g(X)$ over X. The goal of the function is to measure the variances of the feature weights while training over mini-batches of data, and gradually remove spurious features while retaining invariant ones. Formally:

$$g(X_i) \rightarrow \begin{cases} X_i & \text{if } X_i \text{ is invariant} \\ 0 & \text{if } X_i \text{ is spurious} \end{cases}, \forall X_i \in X \tag{4}$$

where g is a monotonic function and the image of $g \in [0, 1]$. We note that the IRMv1 representation in equation (3) is equivalent to having the identity function $\mathbb{I}$ as a mask over X^e, i.e., $\Phi(\mathbb{I}(X^e))$.

We measure the variance of the weights of each feature $X_i \in X$ using a set of masks $\mathbf{M} = \{m_1, ..., m_k\}, m_i \in \mathbb{R}$, where k is the number of features in X. The masks are updated with two objectives: (a) Use the variances to emphasize invariant features and suppress spurious ones, and (b) Exploit the sequential structure and invariance of X^{seq} based on Assumption 1. During each training epoch, we first update the masks as:

$$m_i \leftarrow m_i + \mu(v(\mathbf{w})) - \alpha(v(w_i)), \ \forall m_i \in \mathbf{M} \tag{5}$$

where $\mu(v(\mathbf{w}))$ is the average variance observed over all features in X, $v(w_i)$ is the variance of the weights of feature X_i, and hyperparameter α is a scaling factor. Intuitively, the masks of invariant features gain in value over the training epochs since their variance $v(w_i)$ is very low. Masks of spurious features on the other hand become negative, since the variance of their weights, coupled with the scaling factor is larger than the average which is brought down by invariant features. We then achieve the second objective by updating the masks $\mathbf{M}^{seq}$ of X^{seq} as:

$$m_i^{seq} \leftarrow |m^{seq}|, \ \forall m_i^{seq} \in \mathbf{M}^{seq} \subseteq \mathbf{M} \tag{6}$$

where $|\cdot|$ is the absolute function, which exploits Assumption 1 and ensures the invariance of X^{seq}. The degree of invariance is still dependent on the magnitude of variance exhibited by the weights of X^{seq}. Since the values of $\mathbf{M}$ are unbounded, we scale the masks by using the sigmoid function σ. Since the sigmoid function is bounded between [0,1], $\sigma(\mathbf{M})$ satisfies equation (4). We then find an *environment agnostic sequential predictor* $\mathbf{Z}$ by solving:

$$\min_{\mathbf{Z}} \mathcal{R}(\mathbf{Z}) + \lambda \|\nabla_w \mathcal{R}(w^\top \mathbf{Z})\|_2^2 \text{ , s.t. } \mathbf{Z} = \sigma(\mathbf{M}) \odot X. \tag{7}$$

where $\odot$ denotes element-wise multiplication and $\sigma(\mathbf{M}) \in [0, 1]$. The penalty term here serves to balance the predictive power and invariance of the predictor over the entire training data as opposed to over each training environment. Our entire generalization approach is shown in Algorithm 1.

In order to prove that the masking function and environment agnostic predictor $\mathbf{Z}$ in equation (7) results in a generalized model for out-of-distribution data, we formulate it as a minimax problem in Theorem 1 and show that our solution minimizes the risk or loss using invariant features, even under the most adverse test environment.

Theorem 1. *Given a training environment e_{tr}, and a test environment e_{test}, the set of invariant features X^I is the saddle point of the following minimax problem*

$$X^I = \min_{Z} \max_{X^I, X^S} \mathcal{L}_{test}(Z; X^I, X^S) \ \text{ , where } \ Z = \sigma(M) \odot X \text{ ,}$$

where $\mathcal{L}_{test}$ is the cross-entropy loss in the test environment, and X^I, X^S denote the set of invariant and spurious features respectively such that $X^I \cup X^S = X$, and $X^I \cap X^S = \emptyset$ (i.e.) they are disjoint.

Proof. For every $\mathbf{Z}$, we can partition it into invariant variables $\mathbf{Z}^I$ and non-invariant variables $\mathbf{Z}^S$ as:

$$\mathbf{Z}^I = \sigma(\mathbf{M}) \odot X^I \text{ ,} \quad \mathbf{Z}^S = \sigma(\mathbf{M}) \odot X^S. \tag{8}$$

Consider a test distribution or environment where the set of spurious features X^{S*} are not predictive of the output Y, and only the invariant features X^I are predictive of Y, (i.e.)

$$p(Y|\mathbf{Z}, e_{test}) = p(Y|\mathbf{Z}^I, e_{test}), \quad p(Y|\mathbf{Z}, e_{tr}) = p(Y|\mathbf{Z}^I, e_{tr}) \tag{9}$$

Algorithm 1 Environment Agnostic Sequential Predictor

```
Require: Distribution over inputs $X$ and targets $Y$;
Require: $s$: Total learning steps. $f_\theta$: Function to learn; $X^{seq}$: Sequence features
Require: $s^{init}$: Warm up steps. $\mathcal{L}$: Cross-Entropy loss
Require: $\gamma$: Learning rate; $\alpha$: Scaling factor;
Require: $\theta = (w_1, ..., w_n)$: Classifier weights; $\mathbf{M} = (m_1, ..., m_n)$: Mask weights;
Initialize $m_j = -1, \ \forall j : 1 \rightarrow n$
Initialize mean $\mathbf{u} = (u_1, ..., u_n) = 0$
Initialize variance $\mathbf{v} = (v_1, ..., v_n) = 0$
for $i = 1 \rightarrow s$ do
    Sample batch $\mathbf{x}, \mathbf{y}$                          ▷ Environment Agnostic
    $\ell = \mathcal{L}(f(\sigma(\mathbf{M}) \times \mathbf{X}), y)$  ▷ Prediction error from equation 7
    $l_1 = ||\theta||_1$                                ▷ $l_1$ penalty loss
    $l_{irm} = IrmPenalty(\mathbf{x}, \mathbf{y})$               ▷ Penalty from equation 7
    $l_{final} = \ell + l_1 + l_{irm}$                      ▷ Total loss
    $\theta = \theta - \gamma \nabla \theta l_{final}$
    $\mathbf{u}_{old} = \mathbf{u}$
    $\mathbf{u} = \beta\theta + (1 - \beta)\mathbf{u}_{old}$                    ▷ Mean estimate
    $\mathbf{v} = \delta\mathbf{v} + (1 - \delta)(\theta - \mathbf{u}_{old})^2$           ▷ Variance Estimate
    if $i > s^{init}$ then
        $m_j \mathrel{+}= \mu(v(\mathbf{w})) - \alpha(v(w_j)), \ \forall m_j \in \mathbf{M}$   ▷ Update mask
        $m_j^{seq} = |m^{seq}|, \ \forall m_j^{seq} \in \mathbf{M}^{seq} \subseteq \mathbf{M}$        ▷ Assumption 1
    end if
end for
```

Therefore,

$$\begin{aligned}\mathcal{L}_{test}(\mathbf{Z}\ ; X^I, X^{S*}) &= H(p(Y|\mathbf{Z}, e_{test}); p(Y|\mathbf{Z}, e_{tr})) \\ &\overset{(i)}{=} H(p(Y|\mathbf{Z}^I, e_{test}); p(Y|\mathbf{Z}^I, e_{tr})) \\ &\overset{(ii)}{=} H(p(Y|\sigma(\mathbf{M}) \odot X^I, e_{test}); p(Y|\sigma(\mathbf{M}) \odot X^I, e_{tr})) \\ &\overset{(iii)}{=} H(p(Y|X^I, e_{test}); p(Y|X^I, e_{tr})) \\ &= \mathcal{L}_{test}(X^I\ ; X^I, X^{S*})\end{aligned} \tag{10}$$

where $H(\cdot)$ is the cross-entropy loss function. Step (i) is obtained from applying equation (9). Step (ii) is obtained by applying equation (8), and step (iii) is due to the property of the masks from equation (4).

Recall that X^{S*} was assumed to be non-predictive of Y. However, in most cases, the spurious feature X^S would have some predictive power over Y in the training environment. Hence, from the definition of spurious features X^S, their biased influence on the model performance during training will lead to an increased loss in the worst case test environment:

$$\max_{X^S} \mathcal{L}_{test}(\mathbf{Z}; X^I, X^S) \geq \mathcal{L}_{test}(\mathbf{Z}; X^I, X^{S*}) \tag{11}$$

Recall X^I denotes the set of invariant features, thus $p(Y|X^I, e_{test})$ does not depend on X^S. Therefore,

$$\max_{X^S} \mathcal{L}_{test}(X^I; X^I, X^S) = \mathcal{L}_{test}(X^I; X^I, X^{S*}) \tag{12}$$

By combining equations (10), (11), and (12), we have:

$$\max_{X^S} \mathcal{L}_{test}(\mathbf{Z}; X^I, X^S) \geq \max_{X^S} \mathcal{L}_{test}(X^I; X^I, X^S) \tag{13}$$

The above formulation holds for all X^I. Hence, taking the maximum over X^I in equation (13) preserves the inequality,

$$\max_{X^I, X^S} \mathcal{L}_{test}(\mathbf{Z}; X^I, X^S) \geq \max_{X^I, X^S} \mathcal{L}_{test}(X^I; X^I, X^S)$$

which in turn implies,

$$X^I = \min_{\mathbf{Z}} \max_{X^I, X^S} \mathcal{L}_{test}(\mathbf{Z}; X^I, X^S)$$

□

4 EXPERIMENTS

In this section, we present an extensive evaluation of our approach on five different sequential datasets spanning multiple application domains: natural language processing (NLP), temporal sequences, and business process mining (Table 1). Similar to prior work [1, 4, 9, 20], we augment these benchmark datasets with spurious features. We assess the quality of generalization as the classification accuracy obtained on the test environment, disjoint from the set of training environments. We also present results from an ablation study of the masking function as well as additional experiments that highlight the robustness of our approach.

Dataset	Train Seq.	Test Seq.	Classes	#Spurious
SST-2	67,350	873	2	1
AG News	120,000	7600	4	1
HAR	7352	2947	6	2
BPIC 2018	306,615	63,692	14	1
BPIC 2019	56,736	26,499	12	2

Table 1: Summary of Datasets

4.1 Benchmarks

We compare our Environment Agnostic Sequential Predictor (EASP) to four benchmark approaches:

- Empirical Risk Minimization (ERM): Standard classifier that minimizes the average loss over the entire training data.
- Invariant Empirical Risk Minimization (Inv-ERM): ERM trained on data without any spurious features. This approach reflects the setting where spurious correlations are not present.
- Invariant Risk Minimization (IRMv1): Environment-aware predictor proposed by Arjovsky et al. [4]. While Ahuja et al. [1], Krueger et al. [20] have built on this and achieved similar or slightly better results, they are also environment-aware approaches. Hence, we use the original IRMv1 approach as a representative for environment-aware approaches.
- IRMv1 with 5% segmentation error (IRM5%): Measure of the performance of IRMv1 when there is a 5% error in correctly segmenting the training data into different environments.

4.2 Implementation Details

We implement our EASP approach and the ERM approach using the PyTorch library [28]. We implement the Invariant Risk Minimization approach of Arjovsky et al. [4] based on their publicly available code[1]. We preprocess the datasets used in the paper similar to Choe et al. [9][2] and describe them in detail below.

Hyper-parameter	Range
Learning rate	$[10^{-5}, 10^{-1}]$
Steps	$[101, 501]$
Regularization weight	$[10^{-6}, 10^{-2}]$
Penalty	$[10, 10^{4}]$
Scaling factor	$[10^{-2}, 10]$

Table 2: Hyper-parameter ranges tried for all comparison approaches

For each dataset, we tried both MLP and LSTM classifiers of different architectures and selected the model that performed the best. In each dataset, we used the same model architecture for all five approaches (ERM, Inv-ERM, EASP, IRMv1, IRM5%) to ensure a fair comparison. All experiments were done on a 6-core i7 CPU with 32GB memory. We used the cross-entropy loss for classification during training and the Adam optimizer. We searched for hyper-parameters based on the values in Table 2, and chose the configurations with the best performance for each of the approaches. The results reported are the average over 10 runs. Appendix A provides additional implementation details.

4.3 Natural Language Processing

Text classification models learn from sequences of text (sentences, paragraphs, documents, etc.) and assign them into categories. These sequences can be augmented by other features, such as the source of text and length of the sequence, which could be spurious. Several papers [16, 26] have shown how the presence of spurious correlations in text can cause state-of-the-art NLP models to make mistakes and fail to generalize to out-of-distribution data.

4.3.1 Punctuated SST-2: We modify the Stanford Sentiment Treebank (SST-2) [37], a benchmark dataset for binary sentiment analysis in a similar manner as Choe et al. [9]. The dataset consists of 67350 texts and their associated sentiment for training and 873 for testing. We split the training set into two balanced subsets (environments) and treat the test data as a third OOD environment and corrupt each label with a probability $\eta_e = 0.25$. We add a spurious feature X_s by pairing each sentiment with a punctuation mark (. or !) as follows:

$$p(X_s = .\ |Y = 0, e_i) = p(X_s = !\ |Y = 1, e_i) = \alpha_i$$
$$p(X_s = !\ |Y = 0, e_i) = p(X_s = .\ |Y = 1, e_i) = 1 - \alpha_i$$

Here e_i refers to each environment and we set $\alpha_0 = 0.8, \alpha_1 = 0.9$, and $\alpha_{\text{OOD}} = 0.1$. The addition of the punctuation mark as a separate feature, and not as part of the sentence as in Choe et al. [9], allows us to isolate the sentence sequence embeddings to mask, while still preserving the structure and meaning of the sentence.

Table 3 shows the mean accuracies obtained by the different approaches. The standard ERM based model is highly reliant on the spurious correlations and achieves poor generalization, as evidenced by the low test (9.6%) but high training (85.6%) accuracy. Inv-ERM which represents the accuracy when trained on data without spurious correlations achieves 62.7%. Our EASP approach ignores the spurious correlation and achieves 61.2% accuracy, which is similar to IRMv1 (60.2%) while importantly remaining environment-agnostic unlike IRMv1. This difference grows larger when there are data segmentation errors as reflected by IRM5% which achieves 57.5% test accuracy.

4.3.2 Punctuated AG News: To evaluate our approach for multi-class NLP predictions, we use the AG News dataset [47], which consists of a corpus of news articles and their titles that have been classified into four categories - (1) World, (2) Sports, (3) Business, and (4) Sci/Tech. There are 30,000 training and 1,900 test examples for each class. Similar to the Punctuated SST-2 dataset, we split the training data into two balanced environments and treat the test data as a third OOD environment and add a spurious feature X_s that pairs each news category with a punctuation mark in the set $\mathcal{P} = \{.\ !\ @\ \wedge\}$ as follows:

$$p(X_s = .\ |Y = 0, e_i) = p(X_s = !\ |Y = 1, e_i) = \alpha_i$$
$$p(X_s = @\ |Y = 2, e_i) = p(X_s = \wedge\ |Y = 3, e_i) = \alpha_i$$
$$p(X_s = \text{rand}(\mathcal{P} \setminus \mathcal{P}_j)\ |Y = y_j, e_i) = 1 - \alpha_i$$

where $\mathcal{P}_j$ is the j^{th} punctuation mark and $\text{rand}(\mathcal{P} \setminus \mathcal{P}_j)$ represents a random punctuation mark selected among the remainder. We set $\alpha_0 = 0.8$, $\alpha_1 = 0.9$, and $\alpha_{\text{OOD}} = 0.1$. From Table 3 we observe that ERM again relies on the spurious correlation, achieving 63.4% test accuracy. EASP achieves an OOD accuracy of 80.8%, close to Inv-ERM which achieves 83.8%. It also outperforms IRMv1 and IRM5% which achieve 76.9% and 75.0% accuracy on the OOD test dataset, respectively.

Algorithm	SST-2		AG News	
	Train	Test	Train	Test
ERM	85.6 ± 0.1	9.6 ±0.3	92.3 ± 0.1	63.4 ±0.5
IRMv1	62.8 ±1.8	60.2 ±2.1	87.8 ±0.2	76.9 ±0.3
IRM5%	62.4 ±1.1	57.5 ±2.5	88.2 ±0.4	75.0 ±1.2
EASP(ours)	63.2 ±0.2	**61.2 ± 1.1**	80.9 ±0.1	**80.8 ± 0.2**
Inv-ERM	63.4 ±0.1	62.7 ±1.1	84.0 ±0.1	83.8 ±0.3

Table 3: Punctuated SST-2 and Punctuated AG News train and test accuracy comparison.

4.4 Temporal Sequences

Sequential predictions of time-series data have a lot of applications in different domains, including financial predictions, customer behaviour based on their clickstreams, predictions using sensor data sequences, etc. Moe and Fader [27] have shown that features like average browsing time and number of clicks in clickstream data can have spurious correlations with the likelihood of purchase. Temporal sequences collected from distributed and heterogeneous sensor sources can also contain spurious correlations due to location, weather, etc [21].

4.4.1 Colored HAR:. The Human Activity Recgonition (HAR) dataset [3] consists of smartphone accelerometer and gyroscope readings

[1] https://github.com/facebookresearch/InvariantRiskMinimization
[2] https://github.com/kakaobrain/irm-empirical-study

Algorithm	HAR	
	Train	Test
ERM	98.1 ±0.3	73.2 ±2.2
IRMv1	97.6 ±1.4	86.9 ±1.0
IRM5%	97.9 ±1.2	78.9 ±1.9
EASP(ours)	94.2 ±1.0	**87.5 ± 1.6**
Inv-ERM	94.3 ±0.4	89.3 ±0.8

Table 4: Colored Human Activity Recognition (HAR) train and test accuracy comparison

corresponding to six activities (walking, standing, sitting, etc.) performed by participants. The data consists of sequences of 128 timesteps of sensor readings which correspond to a particular activity. We split the training examples (7352 sequences) into two balanced environments and treat the test examples (2947 sequences) as a third OOD environment. To evaluate our approach in settings where more than one spurious feature X_s are present, we "color" each activity with a unique value (X_s^1) similar to the Colored MNIST dataset [4], and also assign a unique sensor type for each activity (X_s^2) as follows:

$$
\begin{aligned}
&p(X_s^1 = c_j | Y = y_j, e_i) = p(X_s^2 = s_j | Y = y_j, e_i) = \alpha_i \\
&p(X_s^1 = c_{(j+1) \bmod |j|} \ |Y = y_j, e_i) = 1 - \alpha_i \\
&p(X_s^2 = s_{(j+1) \bmod |j|} \ |Y = y_j, e_i) = 1 - \alpha_i
\end{aligned}
$$

Here, we assign a unique color c_j and sensor type s_j to each activity y_j with probability $p = \alpha_i$, and assign another with probability $p = 1 - \alpha_i$ for each environment e_i, where $c_j, s_j \in [0, 1]$. We set $\alpha_0 = 0.8, \alpha_1 = 0.9$, and $\alpha_{\text{OOD}} = 0.1$. The relatively high accuracies shown in Table 4 reflects the inherent predictive power of the invariant sensor features in identifying each activity, hence diminishing the impact of the spurious features. However, ERM is still influenced by the spurious correlation achieving 73.2% test accuracy, compared to 89.3% achieved without spurious features by Inv-ERM. Our EASP approach continues to result in a generalized model with 87.5% OOD test accuracy, similar to IRMv1 with 86.9%. However, there is a significant drop in accuracy of IRM even with small data segmentation errors, where IRM5% only achieves 78.9% test accuracy, further highlighting the advantage of an environment-agnostic approach like EASP.

4.5 Business Process Traces

Business processes form an integral part of many enterprise operations including loan applications, insurance claims, hospital records management, etc. Process traces consist of sequences of events corresponding to activities occurring in each process (e.g. credit score check, patient discharge, etc.), and each trace can result in differing sequences based on the process features and the business logic of the process variant it belongs to. The goal is to accurately predict the next event in trace sequences of varying lengths, where spurious correlations can exist in the business process features, e.g., patient gender and discharge rate [7, 10]. We use two real-world event logs from enterprises provided as part of the Business Process Intelligence Challenge (BPIC) series.

4.5.1 Augmented BPIC 2018: The dataset[3] consists of payment applications from German farmers to the European Agricultural Guarantee Fund collected over a period of three years. Each application is processed by one of four departments, and to evaluate our approach in a multi-environment setting, we consider each department as an environment and use three environments for training and the fourth as OOD test data. We consider two business process variants and augment the process trace sequences with a spurious feature X_s denoting the area of the farm as a continuous value in $[500, 10000]$ hectares. We cluster the area into low (L) and high (H) values and spuriously correlate them to the two business process variants ($\mathcal{A}, \mathcal{B}$) as follows:

$$
\begin{aligned}
&p(X_s \in H \ |Y \in \mathcal{A}, e_i) = p(X_s \in L \ |Y \in \mathcal{B}, e_i) = \alpha_i \\
&p(X_s \in L \ |Y \in \mathcal{A}, e_i) = p(X_s \in H \ |Y \in \mathcal{B}, e_i) = 1 - \alpha_i
\end{aligned}
$$

We set $\alpha_0 = 0.85, \alpha_1 = 0.9, \alpha_2 = 0.95$, and $\alpha_{\text{OOD}} = 0.1$, and evaluate on trace sequence lengths ranging from 5 to 10. Figure 2a, shows that our EASP model – unlike ERM – is not influenced by the spurious correlations and performs comparably with Inv-ERM. Furthermore our approach outperforms IRMv1 over all sequence lengths by 4% on average and IRM5% by 7%.

4.5.2 Augmented BPIC 2019: The dataset[4] consists of purchase orders from a company in the Netherlands. Each order consists of multiple items and goes through several vendors reflecting their payments, invoices, and receipts. We consider two process variants and split the training data (56736 sequences) into two balanced environments and consider the test data (26499 sequences) as a third OOD environment. We augment the process sequences with two spurious features (X_s^1, X_s^2) : item type ($\mathcal{I} = \{\mathcal{I}_1, \mathcal{I}_2\}$) and item valuation ($L \in [\$500, \$5000], H \in [\$50000, \$100000]$), whose values are spuriously correlated with the two business process variants ($\mathcal{A}, \mathcal{B}$) as follows:

$$
\begin{aligned}
&p(X_s^1 = \mathcal{I}_1 \ |Y \in \mathcal{A}, e_i) = p(X_s^1 = \mathcal{I}_2 \ |Y \in \mathcal{B}, e_i) = \alpha_i \\
&p(X_s^1 = \mathcal{I}_2 \ |Y \in \mathcal{A}, e_i) = p(X_s^1 = \mathcal{I}_1 |Y \in \mathcal{B}, e_i) = 1 - \alpha_i \\
&p(X_s^2 \in H \ |Y \in \mathcal{A}, e_i) = p(X_s^2 \in L \ |Y \in \mathcal{B}, e_i) = \alpha_i \\
&p(X_s^2 \in L \ |Y \in \mathcal{A}, e_i) = p(X_s^2 \in H \ |Y \in \mathcal{B}, e_i) = 1 - \alpha_i
\end{aligned}
$$

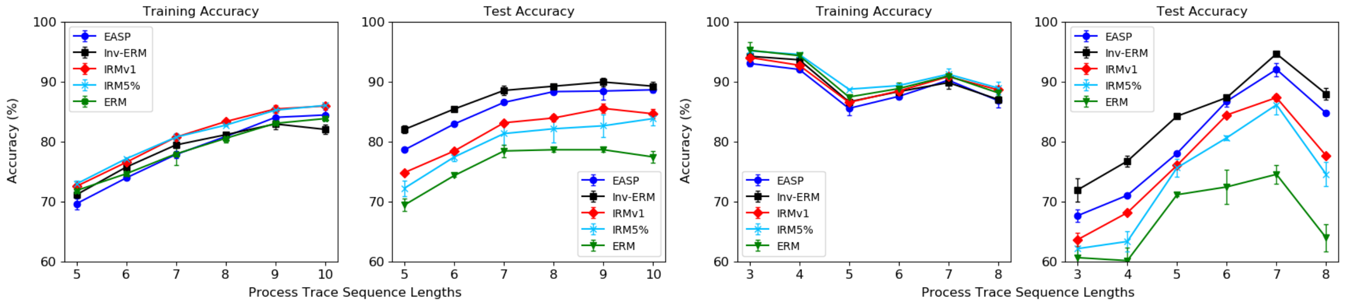
Here we set $\alpha_0 = 0.8, \alpha_1 = 0.9$, and $\alpha_{\text{OOD}} = 0.1$ and evaluate the models on trace sequence lengths ranging from 3 to 8. Figure 2b shows that ERM is heavily influenced by the spurious correlations, while EASP achieves good generalization and outperforms IRMv1 by 5% and IRM5% by 8% on the OOD test set on average over all sequence lengths.

4.6 Ablation Study

We perform an ablation experiment to measure the contribution of our masking function and each of its components on the performance of EASP as shown in Table 5. For the business process datasets, we set the trace sequence length to 7. We first remove the masking function $\sigma(\mathbf{M})$, essentially reducing the EASP formulation in equation (7) to an environment-agnostic version of IRMv1. This is reflected by the accuracy values which are similar to those of

[3] https://data.4tu.nl/articles/BPI_Challenge_2018/12688355

[4] https://data.4tu.nl/articles/BPI_Challenge_2019/12715853



(a) Augmented BPIC 2018

(b) Augmented BPIC 2019

Figure 2: Business Process Traces event classification, train and test accuracy comparison

Algorithm	Test Accuracy (%)				
	SST-2	AG News	HAR	BPIC'18	BPIC'19
EASP	61.2	80.8	87.5	88.4	92.0
$-\ \sigma(\mathbf{M})$	9.6	60.5	72.9	76.1	71.5
$-\ \lvert\mathbf{M}^{seq}\rvert$	52.1	24.9	18.2	19.8	35.1
− scaling (α)	9.6	53.6	74.7	84.0	72.1

Table 5: Ablation Study

Dataset	Test Accuracy After $\lvert M\rvert$			
	All	Spurious	Sequence	None
	($\lvert\mathbf{M}\rvert$)	($\lvert\mathbf{M}^S\rvert$)	($\lvert\mathbf{M}^{seq}\rvert$)	($\lvert\mathbf{M}^{\phi}\rvert$)
SST-2	9.6	9.6	61.2	52.1
AG News	55.9	9.9	80.8	24.9
HAR	74.6	17.5	87.5	18.2
BPIC'18	82.0	19.8	86.5	19.8
BPIC'19	76.0	39.6	92.0	35.1

Table 6: Impact of selecting right features for |M|

ERM, thus demonstrating that the masking function is the reason for an environment-agnostic predictor. Second, we do not leverage Assumption 1 of setting an invariant $|\mathbf{M}^{seq}|$ and observe that there is a significant drop in accuracy in most of the datasets. This can be attributed to the masking function being unable to accurately measure the variance of the sequence as whole. Third, we remove the scaling factor from the mask updation, and see that it also has an impact on all datasets because the degree of spuriousness may not get fully captured, and hence the model can get influenced by spurious features.

4.7 Selecting the Right Features for |M|

While we have shown the effectiveness of updating the mask of the sequence feature with $|\mathbf{M}^{seq}|$, we perform an experiment to measure the performance if this was applied to the masks of other features in the data. For each dataset, we apply the absolute function to all masks ($|\mathbf{M}|$), only those of the spurious features ($|\mathbf{M}^S|$), the sequence feature ($|\mathbf{M}^{seq}|$), and if it was not applied at all ($|\mathbf{M}^{\phi}|$) which was shown in the ablation study.

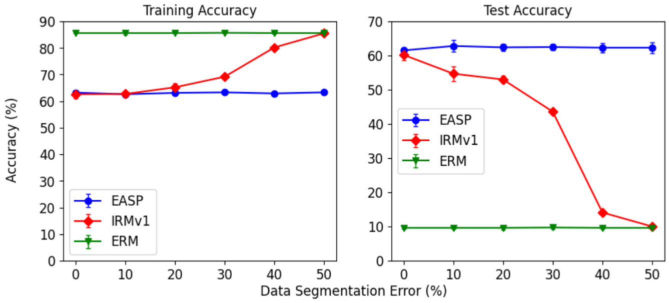


Figure 3: Sensitivity of EASP, IRMv1, and ERM to imperfect segmentation of data for Punctuated SST-2

From Table 6, we see that selecting the right masks to update with the absolute function has an impact on the OOD test accuracy. Across all datasets, updating spurious feature masks with $|\mathbf{M}^S|$ results in very low accuracies since the masking function considers them to now have some predictive power. Along the same lines, doing this for all features ($|\mathbf{M}|$), does not perform as well as $|\mathbf{M}^{seq}|$, since the model still gets influenced to some degree by the spurious features. Hence, using Assumption 1 allows us to exploit the sequential data structure to result in a model that generalizes well.

4.8 Robustness of EASP

4.8.1 *Imperfect segmentation of environments:* For the same motivating example in Figure 1, we measure the performance of EASP when the data are imperfectly segmented into the different training environments for the Punctuated SST-2 dataset. As before, we set $\alpha_0 = 0.8, \alpha_1 = 0.9$, and $\alpha_{\text{OOD}} = 0.1$. Figure 3, shows that our EASP approach is robust to spurious correlations irrespective of the degree of imperfect data segmentation, unlike IRMv1 and ERM.

4.8.2 *Sequence feature is not predictive:* We evaluate the performance of EASP when Assumption 1 (i.e. there exists a sequential feature that is predictive of the target variable) does not hold. In our Punctuated SST-2 dataset, we set the invariant correlation of the text with respect to the sentiment to be 0.5 by changing the probability of label switching (η_e). We again set $\alpha_0 = 0.8, \alpha_1 = 0.9$, and $\alpha_{\text{OOD}} = 0.1$. From Table 7 we see that the OOD test accuracy of

Algorithm	Train Acc.	Test Acc.
ERM	85.4 ±0.3	9.6 ±0.0
IRMv1	52.1 ±0.5	42.8 ±2.0
IRM5%	52.8 ±0.4	40.3 ±1.7
EASP (ours)	51.1 ±1.1	**52.4 ± 2.3**
Inv-ERM	50.0 ±0.1	54.5 ±0.0

Table 7: Sequence is not predictive

Algorithm	Train Acc.	Test Acc.
ERM	70.3 ±0.3	31.4 ±0.3
IRMv1	61.5 ±0.5	**61.0 ± 1.2**
IRM5%	62.1 ±0.4	58.1 ±1.5
EASP (ours)	63.1 ±0.2	**61.0 ± 1.1**
Inv-ERM	63.4 ±0.1	62.7 ±1.1

Table 8: Invariant correlation stronger than spurious

Algorithm	Number of Environments			
	2	4	6	8
ERM	9.6	9.7	9.2	9.4
IRMv1	60.5	60.4	61.0	60.0
IRM5%	58.1	56.1	57.3	58.6
EASP (ours)	**61.2**	**61.0**	**61.0**	**61.1**
Inv-ERM	62.7	61.6	62.4	62.2

Table 9: Varying Number of Environments

EASP, IRMv1, IRM5% and Inv-ERM all drop close to random chance (50%) reflecting the randomness now associated with the invariant correlation. This also shows that EASP outperforms IRMv1 and IRM5% and still continues to be robust to the spurious correlation unlike ERM.

4.8.3 Invariant correlation is stronger than spurious: We compare the performance of the algorithms for cases where the spurious correlation exists in the dataset but the strength of the invariant correlation is stronger. For the Punctuated SST-2 dataset, we set the invariant correlation to be 0.75 by changing η_e and set the spurious correlations to be $\alpha_0 = 0.6$ and $\alpha_1 = 0.7$. The OOD spurious correlation is kept as $\alpha_{\text{OOD}} = 0.1$. From Table 8, we see that while the accuracy of ERM improves now that the influence of the spurious features are reduced, it still fails to completely ignore the spurious correlation. Both EASP and IRMv1 perform similarly and return generalized models.

4.8.4 Varying Number of Training Environments. We measure the robustness of EASP to varying numbers of training environments in the Punctuated SST-2 dataset. We set the maximum and minimum values of α_i as 0.7 and 0.9 respectively and spread out the environments evenly. Table 9 shows that EASP continues to return a generalized model, and the variance in EASP's performance is similar to IRMv1, IRM5% and Inv-ERM even when the number of environments increases.

5 RELATED WORK

There are various approaches to improving out-of-distribution generalization of deep learning models. Causal model discovery [17, 29] aims to find an underlying causal graph to obtain an invariant feature set that is a causal predictor of the target. Invariant Risk Minimization [4] is an optimization based improvement that allows searching over transformations in a continuous space.

Data augmentation techniques are also popular and aim to make the model more robust by training using instances obtained from neighbouring domains hallucinated from the training domains, and thus make the network ready for these neighbouring domains. Shankar et al. [36] augment data using instances perturbed along directions of domain change and use a second classifier to capture this. Volpi et al. [44] apply this to single domain data, while Carlucci et al. [8] apply augmentation to images during training by simultaneously solving an auxiliary unsupervised jigsaw puzzle alongside.

Decomposition based approaches represent the parameters of the network as the sum of a common parameter and domain-specific parameters during training [11]. Khosla et al. [19] applied decomposition to domain generalization by retaining only the common parameter for inference. Li et al. [23] extended this work to CNNs where each layer of the network was decomposed into common and specific low-rank components. Piratla et al. [32] recently proposed a more efficient approach that decomposes only the last layer, imposes loss on both the common and domain-specific parameters, and constrains the two parts to be orthogonal.

Another approach is to pose the domain generalization problem as a meta-learning task, whereby we update parameters using meta-train loss but simultaneously minimizing meta-test loss [22]. Prior work on meta-learning has been studied either in the context of few-shot supervised learning methods which adapt using small amounts of labeled data from the new domain [13, 33, 35], distribution shifts in only test domains [12, 46], or only considering label shifts [25, 39].

Other approaches include adversarially learning representations that are invariant with respect to domain-specific features using perturbations [2, 41] as well as domain erasure methods which estimate features that have the same distribution across different domains using techniques like data-reconstruction, projection, MMD, etc [14, 15, 24]. While out-of-distribution generalization has primarily been done for image and text classification, there have also been interesting applications to visual question answering [40], business process predictions [43] and medical diagnosis using human annotated spurious features [38].

6 CONCLUSION

We present an environment-agnostic approach to identify invariant features and improve the generalization of deep learning models for classification of sequential datasets. Our approach overcomes the inherent drawback of invariant risk minimization based methods which rely on prior knowledge of the different environments or sources of spurious correlations while training. We develop a masking function over the input features that continually detects and gradually removes spurious features from the model during training, resulting in only the invariant features remaining. We also prove that the family of masking functions satisfying these conditions will minimize loss even under the most adverse test distributions. We show that our approach results in models that can generalize to out-of-distribution data and perform competitively on a range of sequential datasets without the need for prior environment knowledge and perfect data segmentation. We believe that this work motivates the strength of environment-agnostic approaches for generalization and as part of future work, we intend to formally identify the domain of masking functions that can be used in this setting, determine the effectiveness of other models that satisfy the minimax optimization such as GANs, as well as extend this approach to other prediction tasks.

REFERENCES

[1] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. 2020. Invariant risk minimization games. *arXiv preprint arXiv:2002.04692* (2020).

[2] Isabela Albuquerque, João Monteiro, Tiago H Falk, and Ioannis Mitliagkas. 2019. Adversarial target-invariant representation learning for domain generalization. *arXiv preprint arXiv:1911.00804* (2019).

[3] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. 2013. A public domain dataset for human activity recognition using smartphones.. In *Esann*, Vol. 3. 3.

[4] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).

[5] Sara Beery, Grant Van Horn, and Pietro Perona. 2018. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 456–473.

[6] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. 2019. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912* (2019).

[7] Alfredo Bolt, Massimiliano de Leoni, and Wil MP van der Aalst. 2018. Process variant comparison: using event logs to detect differences in behavior and business rules. *Information Systems* 74 (2018), 53–66.

[8] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. 2019. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2229–2238.

[9] Yo Joong Choe, Jiyeon Ham, and Kyubyong Park. 2020. An Empirical Study of Invariant Risk Minimization. *arXiv preprint arXiv:2004.05007* (2020).

[10] Carsten Cordes, Thomas Vogelgesang, and Hans-Jürgen Appelrath. 2014. A generic approach for calculating and visualizing differences between process models in multidimensional process mining. In *International Conference on Business Process Management.* Springer, 383–394.

[11] Hal Daumé III. 2009. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815* (2009).

[12] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. 2019. Domain generalization via model-agnostic learning of semantic features. In *Advances in Neural Information Processing Systems.* 6450–6461.

[13] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400* (2017).

[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17, 1 (2016), 2096–2030.

[15] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. 2015. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision.* 2551–2559.

[16] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324* (2018).

[17] Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. 2018. Invariant causal prediction for nonlinear models. *Journal of Causal Inference* 6, 2 (2018).

[18] Khurram Javed, Martha White, and Yoshua Bengio. 2020. Learning Causal Models Online. *arXiv preprint arXiv:2006.07461* (2020).

[19] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. 2012. Undoing the damage of dataset bias. In *European Conference on Computer Vision.* Springer, 158–171.

[20] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Remi Le Priol, and Aaron Courville. 2020. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688* (2020).

[21] Manish Kumar, Devendra P Garg, and Randy A Zachery. 2007. A method for judicious fusion of inconsistent multiple sensor data. *IEEE Sensors Journal* 7, 5 (2007), 723–733.

[22] D Li, Y Yang, Yi-Zhe Song, and TM Hospedales. 2018. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018).* AAAI press, 3490–3497.

[23] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision.* 5542–5550.

[24] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. 2018. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 5400–5409.

[25] Zachary C Lipton, Yu-Xiang Wang, and Alex Smola. 2018. Detecting and correcting for label shift with black box predictors. *arXiv preprint arXiv:1802.03916* (2018).

[26] R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007* (2019).

[27] Wendy W Moe and Peter S Fader. 2004. Capturing evolving visit behavior in clickstream data. *Journal of Interactive Marketing* 18, 1 (2004), 5–19.

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems.* 8026–8037.

[29] J Pearl. 2009. Causality:models, reasoning, and inference. 2nd edn Cambridge University Press. *New York* (2009).

[30] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).* 1532–1543.

[31] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2015. Causal inference using invariant prediction: identification and confidence intervals. *arXiv preprint arXiv:1501.01332* (2015).

[32] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. 2020. Efficient Domain Generalization via Common-Specific Low-Rank Decomposition. *arXiv preprint arXiv:2003.12815* (2020).

[33] Sachin Ravi and Hugo Larochelle. 2016. Optimization as a model for few-shot learning. (2016).

[34] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? *arXiv preprint arXiv:1902.10811* (2019).

[35] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. Meta-learning with memory-augmented neural networks. In *International conference on machine learning.* 1842–1850.

[36] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. 2018. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745* (2018).

[37] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing.* 1631–1642.

[38] Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. 2020. Robustness to Spurious Correlations via Human Annotations. *arXiv preprint arXiv:2007.06661* (2020).

[39] Milan Sulc and Jiri Matas. 2019. Improving cnn classifiers by estimating test-time priors. In *Proceedings of ICCV Workshops.* 0–0.

[40] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. 2020. Unshuffling data for improved generalization. *arXiv preprint arXiv:2002.11894* (2020).

[41] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 7167–7176.

[42] Vladimir Vapnik. 1992. Principles of risk minimization for learning theory. In *Advances in neural information processing systems.* 831–838.

[43] Praveen Venkateswaran, Vinod Muthusamy, Vatche Isahagian, and Nalini Venkatasubramanian. 2021. Robust and Generalizable Predictive Models for Business Processes. In *International Conference on Business Process Management.* Springer, To appear.

[44] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. 2018. Generalizing to unseen domains via adversarial data augmentation. In *Advances in neural information processing systems.* 5334–5344.

[45] Haohan Wang, Zexue He, Zachary C Lipton, and Eric P Xing. 2019. Learning robust representations by projecting superficial statistics out. *arXiv preprint arXiv:1903.06256* (2019).

[46] Marvin Zhang, Henrik Marklund, Abhishek Gupta, Sergey Levine, and Chelsea Finn. 2020. Adaptive Risk Minimization: A Meta-Learning Approach for Tackling Group Shift. *arXiv preprint arXiv:2007.02931* (2020).

[47] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems.* 649–657.

A APPENDIX

Our appendix contains implementation details of the experiments in the paper as described in Section 4.

A.1 Experiment Implementation Details

For both the Punctuated SST-2 and AG News datasets, we used a 3-layer (50-50-2) MLP with ReLU activations, based on the average of the 50-dimensional GloVe (glove.6B.50d) [30] embedding word vectors as inputs. These are pre-trained word vector embeddings that have been trained on Wikipedia articles. We achieve similar results with the 300-dimensional word vectors as well (glove.6B.300d). For Punctuated SST-2 we use 501 steps and penalty weight of 7.5k where the penalty is used after 100 steps. ERM and Inv-ERM use a learning rate of 0.01 and regularizer weight of 0.0005. EASP uses a learning rate of 0.05, scaling factor of 10, and regularizer weight 0.0001 with a warm up of 50 steps, while IRMv1 and IRM5% use a learning rate of 0.001 and regularizer weight of 0.0001. For AG News we use 401 steps and penalty weight of 7.5k where the penalty is used after 100 steps. We set the learning rate to 0.01 and regularizer weight to 0.0001 for all the methods and the scaling factor to 10.

For the Colored HAR, Augmented BPIC 2018, and Augmented BPIC 2019 datasets, we use an LSTM with one layer and hidden size of 100, followed by a linear layer of size 100 with ReLU activation.
For the Colored HAR dataset we use a batch size of 64 and use 401 steps with a penalty weight of 10 which is used after 5 steps. The learning rate is set to 0.01 for all methods. We don't use a regularizer for ERM and Inv-ERM, set it to 0.000001 for EASP with scaling factor of 10 and 0.000007 for IRMv1 and IRM5%. For the Augmented BPIC 2018 dataset we use a batch size of 20% of training and 451 steps with a penalty weight of 10 used after 5 steps. The learning rate is set to 0.005 for all models with a regularizer weight of 0.000001 for EASP with scaling factor 10 and 0.000007 for IRMv1 and IRM5%. For the Augmented BPIC 2019 dataset we use a batch size of 10% of training and 451 steps with a penalty weight of 10 after 5 steps. The learning rate is set to 0.005 for ERM and Inv-ERM, 0.05 for EASP with a scaling factor of 10 and 0.009 for IRMv1 and IRM5%. Regularizer weights are same as BPIC 2018 dataset.