LEARNING A NON-REDUNDANT COLLECTION OF CLASSIFIERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Supervised learning models constructed under the i.i.d. assumption have often been shown to exploit spurious or brittle predictive signals instead of more robust ones present in the training data. Inspired by Quality-Diversity algorithms, in this work we train a collection of classifiers to learn distinct solutions to a classification problem, with the goal of learning to exploit a variety of predictive signals present in the training data. We propose an information-theoretic measure of model diversity based on minimizing an estimate of conditional total correlation of final layer representations across models given the label. We consider datasets with synthetically injected spurious correlations and evaluate our framework's ability to rapidly adapt to a change in distribution that destroys the spurious correlation. We compare our method to a variety of baselines under this evaluation protocol, showing that it is competitive with other approaches while being more successful at isolating distinct signals. We also show that our model is competitive with Invariant Risk Minimization (IRM) under this evaluation protocol without requiring access to the environment information required by IRM to discriminate between spurious and robust signals.

1 INTRODUCTION

The Empirical Risk Minimization (ERM) principle (Vapnik, 2013), which underpins many machine learning models, is built on the assumption that training and testing samples are drawn i.i.d. from some hypothetical distribution. It has been demonstrated that certain violations of this assumption (in conjuction with potential misalignments between the formulation of the learning objectives and the underlying task of interest) lead to models that exploit spurious or brittle correlations in the training data. Examples include learning to exploit image backgrounds instead of objects in the foreground due to data biases (such as using grassy backgrounds to predict the presence of cows (Beery et al., 2018)), using textural as opposed to shape information to classify objects (Geirhos et al., 2018), and using signals not robust to small adversarial perturbations (Ilyas et al., 2019).

Implicit in work that attempts to address these phenomena is the assumption that more robust predictive signals are indeed present in the training data, even if for various reasons current models do not have the tendency to leverage them.

In this work, drawing inspiration from Quality-Diversity algorithms (Pugh et al., 2016) – which seek to construct a collection of high-performing, diverse solutions to a task – we aim to learn a collection of models, each incentivized to find a distinct, high-performing solution to a given supervised learning problem from a fixed training set. Informally, our motivation is that a sufficiently large collection of such distinct models would exploit robust signals present in the training data in addition to the brittle signals that current models tend to exploit. Thus, given the representations computed by such a collection, it may be possible to rapidly adapt to test-time shifts in distribution that destroy the predictive power of brittle features.

Addressing this problem hinges on defining and enforcing an appropriate measure of model diversity. To this end, we make the following contributions:

- We propose and motivate a novel measure of model diversity based on conditional total correlation (across models) of final layer representations given the label. Informally, this

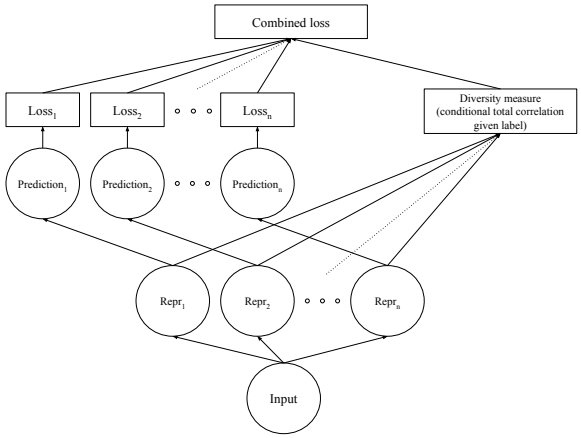

Figure 1: Graphical representation of our framework. We train multiple models to individually minimize supervised loss while simultaneously minimizing conditional total correlation of final layer representations given the label.

incentivizes models to each learn a non-redundant (in an information-theoretic sense) way to solve the given task.

- We estimate this measure using a proxy variational estimator computed using samples from the conditional joint and marginal distributions of final layer representations across models. We train a collection of models to be accordingly diverse, alternating between training a variational critic to maximize the variational estimator and minimizing a weighted sum of the classification losses across models and the variational estimator.
- We empirically validate this framework by training on datasets with synthetically injected spurious correlations. We demonstrate that our framework is able to learn a collection of representations which, under a linear fine-tuning protocol, are competitive with baselines in being able to rapidly adapt to a shift in distribution which destroys the spurious correlations. We show that our framework performs favourably relative to baselines in being able to isolate distinct predictive signals in different models in the collection, as a result being able to make accurate test-time predictions without fine-tuning. We also compare our approach to the Invariant Risk Minimization (Arjovsky et al., 2019) framework which leverages information from multiple training environments to identify signals robust to variations across environments. We show that our approach is able to exploit the robust signals without requiring the metadata needed by IRM to discriminate between spurious and robust signals.

2 TOTAL CORRELATION

Total correlation (Watanabe, 1960) is a multivariate extension of mutual information. The total correlation for n random variables $X_1, ..., X_n$ is given by:

$$TC(X_1, ..., X_n) := KL[p(X_1, ..., X_n) \| \prod_{i=1}^{n} p(X_i)] \tag{1}$$

For $n = 2$ this is equivalent to mutual information. We can interpret this as a measure of the amount of redundancy across the X_i's. A total correlation of zero corresponds to the X_i's being mutually independent.

Given a classification problem predicting label Y from X, we consider the conditional total correlation of a collection of vector-valued representations $h_1(X), ..., h_n(X)$ given Y for differentiable functions h_i , defined as:

$$TC(h_1(X), ..., h_n(X)|Y) := \mathbb{E}_Y[KL[p(h_1(X), ..., h_n(X)|Y) \| \prod_{i=1}^{n} p(h_i(X)|Y)]] \tag{2}$$

where X is drawn from the conditional distribution given Y.

3 METHOD

Consider a collection of differentiable representation function and linear classifier pairs $\{(h_i, c_i)\}_{i \in \{1,...,n\}}$ with parameters $\theta \in \mathbb{R}^k$, such that for a classification problem predicting Y from X, we compute logits $c_i \circ h_i(X)$.

We target the following optimization problem:

$$\begin{aligned} \min_{\theta} \quad & \sum_{i=1}^{n} \mathbb{E}_{X,Y}[l(c_i \circ h_i(X), Y)] \\ \text{subject to} \quad & TC(h_1(X), ..., h_n(X)|Y) < \epsilon \end{aligned} \tag{3}$$

for some $\epsilon > 0$, where l is some loss function.

That is, we seek to minimize the risk for each model while keeping the conditional total correlation of final layer representations given the label below some threshold.

We justify the choice of conditional total correlation in place of the unconditional version by appealing to the following argument: if we consider the case where the collection consists of just two models, total correlation is equivalent to mutual information. Minimizing mutual information between the two representations is equivalent to enforcing independence. However, it is not possible for the representations to be independent while requiring that both are sufficient to accurately predict the label. On the other hand, provided that the representations are not dependent due to some common cause, they will be conditionally independent given the label. Thus, minimizing the conditional total correlation should provide a more useful learning signal.

In practice, we attempt to solve the following unconstrained problem in place of (3):

$$\min_{\theta} \quad \sum_{i=1}^{n} \mathbb{E}_{X,Y}[l(c_i \circ h_i(X), Y)] + \beta TC(h_1(X), ..., h_n(X)|Y) \tag{4}$$

where $\beta > 0$ is a hyperparameter and we estimate the risk using the empirical risk (i.e. sample mean) .

3.1 TOTAL CORRELATION ESTIMATION

In initial experiments with the two-model case (i.e. where total correlation is equivalent to mutual information), the InfoNCE objective of Oord et al. (2018), shown to compute a lower bound of the mutual information (Poole et al., 2019), proved effective in computing useful gradient estimates for the latter term in (4). Consequently, we choose to construct a proxy estimator for the total correlation based on InfoNCE.

The InfoNCE objective for random variables X and Y is given by:

$$I_{\text{NCE}} := \frac{1}{K} \sum_{i=1}^{K} \log \frac{e^{f(x_i, y_i)}}{\frac{1}{K} \sum_{j=1}^{K} e^{f(x_i, y_j)}} \tag{5}$$

where (x_i, y_i) are sampled from the joint distribution of X and Y, K is the batch size, and f is a differentiable variational critic computing unnormalized scores. The parameters of f are optimized to maximize I_{NCE}.

By analogy, our proposed estimator $\widehat{TC}(X_1, ..., X_N)$ is given by:

$$\widehat{TC}(X_1, ..., X_N) := \frac{1}{K} \sum_{i=1}^{K} \log \frac{e^{f(x_{i1}, ..., x_{in})}}{\frac{1}{M} \sum_{j=1}^{M} e^{f(x_{\pi_{1,j}1}, ..., x_{\pi_{n,j}n})}} \tag{6}$$

where $(x_{i1}, ..., x_{in})$ are sampled from the joint distribution of $X_1, ..., X_n$, K is the batch size, $\pi_{k,j} \sim$ Uniform$(\{1, ...K\})$ is a random index of the batch dimension, M is a hyperparameter, and f is a differentiable variational critic computing unnormalized scores. The permuted $\pi_{k,j}$ indices are intended to yield $(x_{\pi_{1,j}1}, ..., x_{\pi_{n,j}n})$ samples drawn approximately from the marginals. Similarly we optimize the parameters of f to maximize $\widehat{TC}(X_1, ..., X_N)$.

We interpret (6) as a ratio of scores between samples from the joint distribution of the X_i's and samples approximately drawn from the marginals, such that a solution to the variational objective resulting in a large ratio corresponds to being able to easily discriminate between samples from the joint and marginals (i.e. a proxy for large total correlation) and conversely when the ratio is small.

We compute the conditional case as follows:

$$\widehat{TC}(X_1, ..., X_N|Y) := \mathbb{E}_Y \left[\frac{1}{K} \sum_{i=1}^{K} \log \frac{e^{f(x_{i1}, ..., x_{in})}}{\frac{1}{M} \sum_{j=1}^{M} e^{f(x_{\pi_{1,j}1}, ..., x_{\pi_{n,j}n})}} \right] \tag{7}$$

where $(x_{i1}, ..., x_{in})$ are sampled from the conditional joint distribution of $X_1, ..., X_n$ given Y.

3.2 Minimax training objective

We replace the latter term in (4) with our variational estimator to obtain:

$$\min_{\theta} \sum_{i=1}^{n} \mathbb{E}_{X,Y}[l(c_i \circ h_i(X), Y)] + \beta \max_{\phi} \widehat{TC}(h_1(X), ..., h_n(X)|Y) \tag{8}$$

where ϕ are the parameters of the variational critic f, and $\beta > 0$ is a hyperparameter.

In practice, we use gradient-based methods to approximately optimize (8) by alternating between optimizing ϕ to maximize the variational estimator (giving us a local total correlation estimate), and optimizing θ to minimize the sum of all terms (for fixed ϕ).

4 Experiments

The main hypothesis that we wish to test is that our method yields a collection of models that exploit distinct predictive signals in the training data, with an emphasis on the inclusion of robust predictive signals typically discarded in the presence of spurious/brittle correlations when training under the ERM assumptions. Our strategy for testing this hypothesis is to consider a training distribution with synthetically injected spurious correlations as well as a test distribution in which the spurious correlation has been destroyed. We then measure our collection of models's ability to adapt to the test distribution given small amounts of data relative to a number of baselines.

We consider the following baselines (implementation details can be found in Appendix A):

- Single ERM classifier
- Ensemble of ERM classifiers
- Collection of ERM classifiers trained to maximize pairwise cosine distances between final layer representations (for a given input)
- Importance-weighted ERM, with optimal weights (an idealized case requiring knowledge of training and test distributions and only included for comparison; optimal importance weights derived in Appendix B)

- IRM (needs access to additional environment metadata)

We require benchmarks in which we can either inject or verify the existence of spurious predictive signals. To that end, we consider variations on the Colored MNIST (C-MNIST) task of Arjovsky et al. (2019). This task consists of classifying MNIST digits into one of two classes (0-4 or 5-9). However, the true label is flipped with probability 0.25 and an additional binary signal (colour) is constructed from this corrupted label such that it is more strongly correlated with the corrupted label than the digit. As demonstrated in Arjovsky et al. (2019), a model trained with ERM will overwhelmingly tend to exploit the colour information. In Arjovsky et al. (2019), the goal was to demonstrate an ability to learn to exploit the digit information given the task of identifying a stable signal across environments with varying colour-label correlation. To test our framework, we assume no such access to multiple environments, and collapse the two training environments in the C-MNIST benchmark task of Arjovsky et al. (2019) into one training set.[1] The results for this task are collected in Table 1.

We also consider a modified dataset which introduces an obstacle to extracting mutually exclusive representations of digit and colour. Specifically, we use functions of a new binary 'common cause' variable to both rotate the digits and perturb the colour signal, such that minimizing total correlation will depend on an ability to discard the common information. This is to test our framework's ability to still extract the predictive signals in the presence of such an obstacle to minimizing conditional total correlation. We refer to this dataset as Rotated Colored MNIST (RC-MNIST). The results for this task are collected in Table 2.

We consider a third variation in which more than two predictive signals are present in the training data. Specifically, we construct another binary 'colour' signal from the corrupted label and add it to the input such that new signal can be exploited on the training set for a maximal accuracy of 0.75. We refer to this dataset as Two-Colored MNIST (TC-MNIST). The results for this task are collected in Table 3.

We similarly construct the same three benchmarks using the Fashion-MNIST dataset (Xiao et al., 2017).

Further implementation details can be found in Appendix A.

4.1 Evaluation

We evaluate our approach by constructing test distributions in which certain predictive signals present in the training data have been destroyed and training relatively simple models to adapt to the new distribution given the outputs computed from the frozen trained models. We consider the following adaptation models:

- **Best**: choose the single best performing model from the collection, and use its frozen predictions (0 parameters)
- **Linear**: train a logistic regression model (with L^2 regularization) on the concatenated frozen final layer representations from each model ($n \times D_{\text{repr}}$ parameters for n models and representation dimension D_{repr})

These adaptation models are trained on 500 observations (1% the size of training set) from the new distribution. We tune hyperparameters using a validation set drawn from the new test distribution and report results on a hidden test set corresponding to maximal validation scores achieved. We treat the number of models in our framework as a hyperparameter and present results for the best setting. We tune hyperparameters for IRM using the approach used to produce the results in Arjovsky et al. (2019). Results for each experimental condition are averaged over 10 runs, for which we report the standard deviation. For baselines requiring multiple models, we report results for 10 models as adding further models did not improve results.

[1] The two environments are constructed such that the colour signal is computed by flipping the corrupted label with probability 0.1 and 0.2 respectively. Thus, collapsing the two environments into one means that we can at best achieve an accuracy of 0.85 on the training set by exploiting the colour signal, versus the maximal accuracy of 0.75 by exploiting the digit.

Table 1: Test accuracies for the C-MNIST task. Test data is drawn from a distribution in which the colour signal has been destroyed.

Method	**Test accuracy (s.e.)** MNIST		**Test accuracy (s.e.)** Fashion-MNIST	
	Linear	Best	Linear	Best
Optimal	75.0	75.0	75.0	75.0
Optimal IWERM	70.3 (0.6)	70.7 (0.5)	70.2 (0.7)	70.5 (0.5)
IRM (requires env. info.)	69.7 (1.1)	68.6 (1.2)	69.0 (0.5)	68.4 (0.6)
Ours (conditional, 2 models)	69.8 (0.4)	69.8 (0.4)	70.0 (0.7)	70.2 (0.4)
ERM (single model)	56.8 (2.6)	50.3 (0.7)	61.4 (2.9)	51.4 (0.5)
ERM (10 models, ensemble)	67.4 (1.0)	51.0 (0.8)	67.6 (1.5)	50.5 (0.6)
Pairwise cosine (10 models)	70.5 (0.8)	56.4 (0.8)	69.5 (0.5)	54.4 (1.1)

Table 2: Test accuracies for the RC-MNIST task. Test data is drawn from a distribution in which the colour signal has been destroyed.

Method	**Test accuracy (s.e.)** MNIST		**Test accuracy (s.e.)** Fashion-MNIST	
	Linear	Best	Linear	Best
Optimal	75.0	75.0	75.0	75.0
Optimal IWERM	72.0 (0.6)	72.2 (0.5)	70.9 (0.6)	70.9 (0.4)
IRM (requires env. info.)	70.2 (0.9)	70.3 (0.5)	69.2 (0.4)	69.5 (0.5)
Ours (conditional, 2 models)	71.3 (0.7)	71.4 (0.8)	70.5 (0.5)	70.4 (0.7)
ERM (single model)	61.6 (7.1)	50.4 (0.4)	69.7 (1.1)	52.0 (1.6)
ERM (10 models, ensemble)	71.5 (0.5)	52.5 (1.5)	70.1 (0.4)	51.3 (0.9)
Pairwise cosine (10 models)	71.8 (0.4)	53.2 (0.8)	70.7 (0.4)	51.1 (0.6)

Table 3: Test accuracies for the TC-MNIST task. Test data is drawn from a distribution in which the two colour signals have been destroyed.

Method	**Test accuracy (s.e.)** MNIST		**Test accuracy (s.e.)** Fashion-MNIST	
	Linear	Best	Linear	Best
Optimal	75.0	75.0	75.0	75.0
Optimal IWERM	69.2 (0.8)	61.7 (2.0)	68.9 (2.1)	61.3 (1.9)
IRM (requires env. info.)	67.5 (1.3)	62.8 (0.8)	68.8 (0.7)	65.5 (0.7)
Ours (conditional, 3 models)	67.5 (0.7)	67.8 (0.8)	67.9 (0.8)	67.6 (2.4)
ERM (single model)	54.7 (3.7)	56.3 (0.5)	60.8 (2.8)	59.4 (0.6)
ERM (10 models, ensemble)	68.6 (1.6)	58.5 (1.1)	69.0 (0.9)	58.7 (1.0)
Pairwise cosine (10 models)	68.2 (1.1)	59.2 (0.7)	67.0 (1.2)	59.7 (0.6)

4.2 INTERPRETATION OF RESULTS

For the C-MNIST benchmark 1, our approach outperforms all baselines under the 'best' evaluation protocol (except for IRM and optimal IWERM, which require additional information to train and are only included for contrast). Results under this evaluation protocol indicate that the approach has successfully isolated the 'digit' signal into a single model to the extent that, given this model, no further fine-tuning is required to make accurate test time predictions. As expected, the single ERM classifier baseline fails across all experimental conditions. Perhaps surprisingly, the ERM baselines with multiple models as well as the 'pairwise cosine' baseline are competitive with our approach under the 'linear' protocol, even though no single frozen model from each respective collection can make accurate test-time predictions (as evidenced by the results under the 'best' protocol). This suggests that the superficial notions of model diversity used in these cases are sufficient, given a large enough number of models, to allow for each model to retain different, small amounts of information about the 'digit' signal such that, taken together, all the small sources of information are sufficient to make accurate test-time predictions. While we acknowledge that this may just be an artefact of the benchmarks considered in this work, we claim that exploring this phenomenon further in more diverse settings is an interesting direction for future work.

We observe a similar pattern of results for the RC-MNIST benchmarks 2. For our framework, this is an indication that the framework is robust to the obstacle to extracting conditionally independent predictive signals entailed in this benchmark.

A similar pattern of results emerges for the TC-MNIST benchmarks 3, with some exceptions: relative to the other benchmarks, performance drops for the optimally importance-weighted baseline. We have no explanation for this other than to defer to work pointing out potential issues with importance weighting when training overparameterized deep neural networks (Byrd & Lipton, 2019).

We note that our approach is competitive with IRM across all experimental conditions, despite having no access to the environment metadata needed by the former to identify invariant predictive signals.

5 RELATED WORK

Our work shares motivation with Quality-Diversity (QD) algorithms (Pugh et al., 2016) from the evolutionary computation literature, which also attempt to find diverse, high-performing solutions to a task. This is typically done by partitioning the search space using a predefined behaviour descriptor and finding high-performing solutions within each behavioural niche (Mouret & Clune, 2015; Cully & Demiris, 2017). We do not explicitly partition the search space but instead enforce diversity through an information theoretic objective. A central challenge in directly applying existing QD algorithms to the problems considered in this work is the difficulty in defining an appropriate behaviour descriptor. An interesting direction for future work would be to investigate the possibility of learning suitable behaviour descriptors from data, as explored in the context of control problems in Cully (2019).

Ensemble-based methods (Polikar, 2006; Rokach, 2010) train collections of models using various techniques (e.g. subsampling of observations and features (Breiman, 2001), boosting (Friedman, 2001), penalizing the correlation of model predictions (Liu & Yao, 1999)) to enforce some notion of model diversity such that aggregate predictions outperform single model predictions. While we have included ensemble baselines, none of the aforementioned approaches are appropriate for the setting considered in this paper.

A number of works (Kim & Mnih, 2018; Chen et al., 2018; Poole et al., 2019; Gao et al., 2019) use total correlation based objectives to enforce factorized representations in generative models. Similar to our approach, FactorVAE (Kim & Mnih, 2018) uses adversarial training to minimize an estimator of total correlation. In contrast to these approaches, we target total correlation for a collection of high-dimensional vector-valued representations while also leveraging label information. A possibility for future work is to impose probabilistic structure on the representations in our proposed method that may enable the use of the upper bound on total correlation presented in Poole et al. (2019), perhaps alleviating the need for adversarial training.

Independent component analysis (ICA) (Comon, 1994) addresses the problem of separating a signal into distinct components using an independence based criterion. This is done in an unsupervised setting and making different assumptions than we do. Brakel & Bengio (2017) use an adversarial training method similar to ours in the context of non-linear ICA.

Other works (Kwok & Adams, 2012; Mariet & Sra, 2015; Pang et al., 2019; Parker-Holder et al., 2020) use determinantal point processes (Kulesza & Taskar, 2012) to compute diversity measures. Exploring these approaches in the context of the problems tackled in this paper is an interesting direction for future work.

There is a large body of work attempting to account for test-time distribution shift (Quionero-Candela et al., 2009). The problems considered in this paper are an example of covariate shift, in which the marginal distribution of the label remains unchanged at test time. Approaches from the literature to address covariate shift are based on discarding input features, reweighting training observations, or domain adversarial training (Ganin et al., 2016). Approaches based on discarding input features are not readily applicable to our context considering high-dimensional unstructured inputs. We have included results for an optimally importance-weighted baseline and have refrained from exploring baselines which estimate the weights from data due to the challenges entailed in doing so in a deep learning context. We were unable to train a domain adversarial approach to produce good results under the evaluation protocols considered in this paper; we hypothesize that this may be due the scarcity of adaptation data under our evaluation protocols. We have deliberately refrained from considered standard image domain adaptation benchmarks as these lie outside of the intended scope of our method.

Some works from the GAN literature (Hoang et al., 2018; Ghosh et al., 2018) attempt to prevent mode collapse by training a collection of generators to produce easily discriminable samples. Experiments adapting such approaches to our context fail completely and we have excluded corresponding results. This is not surprising: in our setting it is easy for learned representations to be easily discriminable while still representing the same information content.

Other works drawing from ideas in the causal inference literature attempt to yield models that generalize better under certain violations of the ERM assumptions. Invariant Risk Minimization (Arjovsky et al., 2019) assumes access to data from multiple environments and attempts to find a representation such that the same classifier is simultaneously optimal in all environments (as do subsequent works addressing the same / similar problem (Ahuja et al., 2020; Krueger et al., 2020)). The problem of causal discovery (Pearl, 2009; Peters et al., 2017) consists in trying to identify underlying causal structure from data such that the tools of causal inference can then be used to compute predictions that are valid under certain shifts in distribution.

6 Conclusion

We have introduced a new method for training a collection of high-performing supervised learning models each incentivized to learn a distinct way to solve the task. We have empirically demonstrated how minimizing our proposed measure of model diversity in addition to the empirical risk suffices to yield a collection of models exploiting distinct predictive signals present in benchmark tasks.

Perhaps surprisingly, we find that simple notions of model diversity considered in certain baselines are sufficient to produce competitive results under a linear fine-tuning evaluation protocol. While this may just be an artefact of the benchmarks considered in this work, we claim that such a strategy (i.e. simple, scalable measures of model diversity coupled with linear fine-tuning given small amounts of new data) is an interesting avenue to explore in future work.

We have focused on evaluating our approach on datasets which we have sufficient control over to enable the testing of basic hypotheses about our proposed framework. To the best of our knowledge, there are currently no other more complex, natural datasets for us to evaluate our approach on. Furthermore, as in Arjovsky et al. (2019), we require access to evaluation data from a different distribution than the training data in order to tune hyperparameters. Rather than seeing this as a disadvantage of our approach, we see it as further evidence that confronting the problem of generalization in a data-driven way may call for the development of new kinds of datasets and benchmarks that are better suited to testing robustness to violations of the i.i.d. assumption on which the empirical risk minimization principle is founded.

REFERENCES

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

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

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

Philemon Brakel and Yoshua Bengio. Learning independent features with adversarial nets for non-linear ica. *arXiv preprint arXiv:1710.05050*, 2017.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, pp. 872–881, 2019.

Tian Qi Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pp. 2610–2620, 2018.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

Antoine Cully. Autonomous skill discovery with quality-diversity and unsupervised descriptors. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 81–89, 2019.

Antoine Cully and Yiannis Demiris. Quality and diversity optimization: A unifying modular framework. *IEEE Transactions on Evolutionary Computation*, 22(2):245–259, 2017.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

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

Shuyang Gao, Rob Brekelmans, Greg Ver Steeg, and Aram Galstyan. Auto-encoding total correlation explanation. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1157–1166, 2019.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Arnab Ghosh, Viveka Kulharia, Vinay P Namboodiri, Philip HS Torr, and Puneet K Dokania. Multi-agent diverse generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8513–8521, 2018.

Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Rémi Le Priol, Gabriel Huang, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative momentum for improved game dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1802–1811, 2019.

Quan Hoang, Tu Dinh Nguyen, Trung Le, and Dinh Phung. Mgan: Training generative adversarial nets with multiple generators. In *International Conference on Learning Representations*, 2018.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems 32*, pp. 125–136. Curran Associates, Inc., 2019. URL `http://papers.nips.cc/paper/8307-adversarial-examples-are-not-bugs-they-are-features.pdf`.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pp. 2649–2658, 2018.

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

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*, 2012.

James T Kwok and Ryan P Adams. Priors for diversity in generative latent variable models. In *Advances in Neural Information Processing Systems*, pp. 2996–3004, 2012.

Yong Liu and Xin Yao. Ensemble learning via negative correlation. *Neural networks*, 12(10):1399–1404, 1999.

Zelda Mariet and Suvrit Sra. Diversity networks: Neural network compression using determinantal point processes. *arXiv preprint arXiv:1511.05077*, 2015.

Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*, 2015.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning*, pp. 4970–4979, 2019.

Jack Parker-Holder, Aldo Pacchiano, Krzysztof Choromanski, and Stephen Roberts. Effective diversity in population-based reinforcement learning. *arXiv preprint arXiv:2002.00632*, 2020.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Robi Polikar. Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, 6 (3):21–45, 2006.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pp. 5171–5180, 2019.

Justin K Pugh, Lisa B Soros, and Kenneth O Stanley. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, 3:40, 2016.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

Lior Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39, 2010.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1):66–82, 1960.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

A IMPLEMENTATION DETAILS

In the notation of Section 3, we use multilayer perceptrons (MLP) as representation functions h_i with two hidden layers of size 128 and 64 respectively, 32 output units, and ReLu activations. We use the cross-entropy as a classification loss. Our single ERM classifier baseline is equivalent to a single model from the collection of models in our framework. The baselines calling for multiple models use multiple instantiations of this same model with different initial weights. The variational critic f is an MLP with 2 hidden layers of size 256 computing a scalar output from an input consisting of a concatentation of all representations computed by the h_i's. We use Leaky ReLus with a slope of 0.2 as activations in the critic as this appeared to speed up training in experiments. We use $\beta = 10$ throughout. For the 'pairwise cosine', we use a coefficient of 0.1 for the loss function term computing the average pairwise cosine distance between final layer representations for a given input.

We use RMSProp with a learning rate of 10^{-5} as an optimizer for the models in the collection as well as the critic. We note that not using a momentum-based optimizer appears to be crucial for stable training, following the trend of decreasing momentum values in the GAN literature (Gidel et al., 2019). We note that training appears to be further stabilized by normalizing representations computed by the h_i's to have unit L^2 norm prior to being input into the critic.

We use a batch size of 256 for training both the models and the critic, as well as $M = 64$ in Equations (6) and (7). To train the simple classifiers, we use 500 training and validation observations from the novel test distribution, the latter used to tune the hyperparameters (if any) of the newly trained classifiers. We report test accuracies computed for the remaining 9000 test observations. We train for 250 epochs and report test accuracies corresponding to maximal validation accuracies. We make the following exceptions for the TC-MNIST task: we train for 1000 epochs and alternate between 5 steps of training the critic and 1 step of training the collection of models. This would appear to be necessary for obtaining useful total correlation gradients given the relatively more challenging variational problem that needs to be solved in this case.

These hyperparameter settings were determined from manual tuning of our approach on the C-MNIST benchmark solely for the two-model case using the conditional TC estimator, and reused across all other experimental conditions. Minimal manual tuning of the regularization coefficient for the 'linear' evaluation protocol was subsequently performed for each experimental condition with a distinct number of models in the collection.

Hyperparameters for IRM were chosen using the tuning method used to produce the results in Arjovsky et al. (2019), owing to the similarity of the datasets considered. Hyperparameters were chosen using random search with 50 trials, sampling values as in `https://github.com/facebookresearch/InvariantRiskMinimization/blob/master/code/colored_mnist/reproduce_paper_results.sh`. The best model was chosen on the basis of the mean final test accuracy (over 10 runs) under each of our evaluation protocols.

A.1 DATA SYNTHESIS

C-MNIST

The Colored MNIST dataset is constructed as in Arjovsky et al. (2019).[2] MNIST images are subsampled to 14x14 size, images are assigned to one of two classes (depending on whether the MNIST label is 0-4 or 5-9), and this binary class label is corrupted by flipping with probability 0.25. Subsequently, a binary colour signal is computed from the corrupted label by flipping it with probability p_e, where e is an environment index. The colour signal is represented in the final dataset by appending a new channel to the grayscale image. The new channel is set to a copy of the grayscale image if the color signal is equal to 1, and zeroed out otherwise. The final dataset consists of these two-channel 'images' and the corrupted label. Two training environment with $p_{e_1} = 0.2$ and $p_{e_2} = 0.1$ are used to train the IRM baseline. Our models and the ERM baseline are trained by collapsing these two training environments into one, discarding the environment information.

[2]Based on code from: `https://github.com/facebookresearch/InvariantRiskMinimization`

RC-MNIST

There are a few key changes in synthesizing the Rotated Colored MNIST dataset relative to C-MNIST. Firstly, for each observation a new 'common cause' variable is sampled from $\text{Bernoulli}(0.5)$. Subsequently, the digits in the C-MNIST pipeline are rotated by 45 degrees if the common cause variable is equal to 1, and are not rotated otherwise. The color signal is represented differently than in C-MNIST. Instead of copying the grayscale image in a new channel, the new channel is set to a constant value of 1.0 across all spatial entries if the color signal is equal to 1, and is zeroed out otherwise. Finally, this color channel is shifted up by 0.1 if the common cause variable is equal to 1, and is shifted down by 0.1 otherwise to produce the final dataset.

TC-MNIST

Two-Colored MNIST is similar to RC-MNIST (but without the perturbations due to the common cause variable), with the addition of a new channel representing an additional color signal computed from the corrupted label. For each observation, this new color signal channel is perturbed by sampling a single value from $\text{Normal}(0, 0.1)$ and adding the value to all spatial entries in the channel.

For the purposes of training the IRM baseline, the new color signal is computed as for the original signal but using a new environment specific corruption parameter. Similarly, two environments are respectively created using environment parameters of 0.2 and 0.1 for the original color signal, and 0.3 and 0.2 for the new color signal. Both environments are collapsed into one for training our models and the ERM baseline.

B DERIVATION OF OPTIMAL IMPORTANCE WEIGHTS

Assuming that the supports of the training and test distributions overlap, it can be shown (Quionero-Candela et al., 2009) that the following importance weights for the risk under the training distribution produce an unbiased estimator of the risk under the test distribution:

$$w^*(X,Y) := \frac{p_{test}(X,Y)}{p_{train}(X,Y)} \tag{9}$$

where p_{train} and p_{test} denote training and test distributions respectively.

For the benchmarks considered in this paper, we have access to the complete generative model that produces the data and are thus able to compute these optimal weights in closed form (assuming the input X is computed from the raw digit image D and color signal C). For C-MNIST:

$$\begin{aligned}
w^*(X,Y) \quad &:= \frac{p_{test}(X,Y)}{p_{train}(X,Y)} \\
&= \frac{p_{test}(Y)p_{test}(X|Y)}{p_{train}(Y)p_{train}(X|Y)} \\
&= \frac{p_{test}(X|Y)}{p_{train}(X|Y)} \text{ (since marginal distributions of } Y \text{ are the same)} \\
&= \frac{p_{test}(C,D|Y)}{p_{train}(C,D|Y)} \\
&= \frac{p_{test}(C|Y)p_{test}(D|Y)}{p_{train}(C|Y)p_{train}(D|Y)} \text{ (since colour and digit are conditionally independent given } Y\text{)} \\
&= \frac{p_{test}(C|Y)}{p_{train}(C|Y)} \text{ (since conditional distributions of digit given label are the same)} \\
&= \frac{p_{test}(C)}{p_{train}(C|Y)} \text{ (since colour is independent of label at test time)} \\
&= \frac{0.5}{p_{train}(C|Y)} \text{ (under test distribution colour is Bernoulli(0.5))} \\
&= \frac{0.5}{0.85} \text{ if } C = Y \text{ else } \frac{0.5}{0.15}
\end{aligned} \tag{10}$$

That is, observations where colour and label do not coincide (i.e. contradicting the spurious correlation in the training distribution) are relatively upweighted.