Identifying Invariant and Sparse Predictors in High-dimensional Data

Amin Mansouri [*1] **Sean Spinney** [*1] **Amin Memarian** [*1] **Patricia Conrod** [2] **Irina Rish** [1]

Abstract

In many practical applications (e.g., medical imaging), we are often concerned not only with the accuracy, but also with the interpretability of a predictive model trained on high-dimensional data, such as its ability to identify a sparse subset of invariant (robust) predictors, generalizing across a wide variety of datasets affected by spurious noise (e.g., key factors related to a disease, across different patients and hospitals). Towards this goal, we explore here a combination of the sparsity-inducing l_1 and l_2 regularization with the recently proposed approach for robust predictive modeling, Invariant Learning Consistency. We investigate the ability of the combined approach to identify robust sparse predictors, and demonstrate promising results on several datasets, including synthetic data, the MNIST benchmark, and a functional MRI dataset. Our approach tends to improve the robustness of sparse models in practically all cases, albeit with varying degrees of success and under certain conditions.

1. Introduction

In recent years, there has been a surge of interest in the deep learning community towards better understanding which aspects of deep network models allow them to better generalize to test data drawn from distributions different from the one on which these models were trained - the problem commonly referred to as out-of-distribution (OoD) generalization. A specific focus is on methods and models that aim to disentangle the true *causal* predictors from the possibly many spuriously correlated ones, often present in high-dimensional data. (Arjovsky, 2020; S Chandra Mouli, 2021; Nagarajan et al., 2021; Ahuja et al., 2010; Arjovsky et al., 2020).

One of the recently proposed methods, Invariant Learning Consistency (Parascandolo et al., 2020), aims to improve the model robustness by choosing gradient steps that are consistent across different data points, and are more likely to arrive at solutions (models) consistent across different data distributions, thus achieving better OoD generalization. Herein, we build upon this approach, applying it to the problem of learning robust sparse predictive models which are *structurally consistent* across a range of data distributions, in terms of high overlap across the sparse feature subsets selected.

1.1. Objectives

We propose to combine sparse regularization (e.g., Lasso (Tibshirani, 2011), Elastic Net (Zou & Hastie, 2005)) for graph structure selection when learning predictive models and *Invariant Learning Consistency* (ILC,(Parascandolo et al., 2020)), a recently proposed approach for imposing higher levels of consistency/stability of the parameters in a neural network (or other predictive models) and shown to improve out-of-distribution generalization and consistently outperform the popular IRM approach (Arjovsky et al., 2020) for invariant/robust representation learning.

How do we learn a structure? We know that $l1$ regularization is conducive to getting a *sparse* set of weights. We apply sparse regularization in order to select the most relevant links in a feedforward neural network. But how does this relate to causation? Any neural network could be equivalently considered as a *deterministic* Bayes net. If we impose sparsity by having a large $l1$ regularizer, then we get a subset of nodes in this graphical model but this could not be conceived as a *causal graph*. However, $l1$ regularization won't necessarily return the *same* sparse representation for each round of training. In other words, imposing sparsity alone won't probably yield a true causal model, rather, each time we get a different set of sparse selectors. This is the advantage of ILC: by forcing agreement between gradient updates, we suspect it will result in *consistent and causal sparse representations*.

[*] Equal contribution [1] Université de Montréal, DIRO, Mila, Montreal, Quebec, Canada [2] Université de Montréal, Department of Psychiatry, Montreal, Quebec, Canada. Correspondence to: Amin Mansouri, Sean Spinney, Amin Memarian <amin.mansouri@mila.quebec, sean.spinney@mila.quebec, amin.memarian@mila.quebec>.

Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning., Copyright 2021 by the author(s).

2. Methodology

2.1. Invariant Learning Consistency

Consider a collection of datasets $\{\mathcal{D}^e\}_{e\in\mathcal{E}}$ with $|\mathcal{E}| = d$ and $\mathcal{D}^e = \{(x_i^e, y_i^e)|\, i = 1, \ldots, n_e\}$. The subscript $e \in \mathcal{E}$ representing the different environment from which the data points stem from. With these environment subset of the training data, we can define the loss for one environment:

$$\mathcal{L}_e(\theta) := \frac{1}{|\mathcal{D}^e|} \sum_{(x_i^e, y_i^e)\in\mathcal{D}^e} \ell(f(x_i^e;\theta), y_i^e); \qquad (1)$$

Using the new loss with respect to the environment definition, we can define the new ILC-regularized loss function proposed in (Parascandolo et al., 2020):

$$\mathcal{L}^{\text{ILC}}(\theta) := \mathbb{E}_{\theta\sim p(\theta)}\left[\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\mathcal{L}_e(\theta)\right] - \lambda\cdot\text{ILC}(\theta) \quad (2)$$

Here the regularization factor $\text{ILC}(\theta)$ reflects the consistency score of a given gradient across all environment. This means that if the neighborhood of the minima is not consistent across environments, it is probably not a consistent minima and it doesn't relate to an invariant mechanism. Notice that λ adjusts our desire of predictive power (ERM (Vapnik, 1992)) and invariance (ILC), if $\lambda = 0$ we get back to the classic gradient descent (gradients are averaged and agreement among them will not be considered.) and $\lambda > 0$ means we care about finding invariant mechanism.

AND-Mask On top of the new ILC-regularized loss, the paper introduces the *AND-mask*. The regularization term needs not to be explicitly added to the loss function. We can update the model by evaluating if the gradient directions (sign) are consistent across all environments and take the optimization step in those directions. If a component of the gradient has a majority of the same sign above a certain threshold $\tau \in [0, 1]$ the component is left as is, if not, then the component is zeroed-out. Mathematically, the mask m_τ for a given component j is given by:

$$[m_\tau]_j = \mathbb{1}\left[\tau d \leq \left|\sum_e \text{sign}([\nabla\mathcal{L}_e]_j)\right|\right] \qquad (3)$$

where $d = |\mathcal{E}|$ is the number of environments in the batch. Finally we have the final definition of the masked-ILC-regularized gradient:

$$\nabla_\theta\mathcal{L}^{\text{m-ILC}}(\theta) = m_\tau \odot \nabla_\theta\mathcal{L}(\theta) \qquad (4)$$

Note that the notion of *environments* is very general and could be interpreted differently in various settings. In (Parascandolo et al., 2020) they treat every single sample as its own environment, and we follow the same convention.

2.2. Evaluation

In order to assess the quality of the improvement in recovering consistent estimators, we use the following metrics and motivate their relevance in evaluating the ILC algorithm: test accuracy, and an overlap score for the MNIST experiment, which measures the consistency of learned sparse representations across environments. Together we can compare the predictive power (test accuracy) with robustness.

3. Experiments

3.1. Synthetic Dataset

Using a procedure inspired by the original paper (Parascandolo et al., 2020), we simulate a binary classification dataset with weak invariant predictors across samples (environments). Let $X \in \mathbf{R}^{N\times D}$ and $y \in \{0, 1\}^N$ be a Bernoulli random variable, where N is the number of samples and D the number of predictors. We assume the data is generated according to the following:

$$\epsilon_n \sim N(0, \sigma^2), \forall n = 1, ..., N$$

$$y_n = \begin{cases} 1 & \beta_R X_{n,\Gamma_R} + \beta_S X_{n,\Gamma_S} + \beta_I X_{n,\Gamma_I} + \epsilon_n > 0 \\ 0 & 0 \end{cases}$$

where we consider that $\beta = \{\beta_R; \beta_S; \beta_I : \beta_S \gg \beta_I\}$ are the coefficients defining the relationship between X and y. Γ_R refers to completely random valued columns (features), Γ_S refers to columns that contain spurious features but are *strong* (have high values), and Γ_I refers to columns that are invariant, *but* are weak (have lower values). $\beta_R, \beta_S, \beta_I$ are the coefficients determining the values for each of these column sets. That is, for every n we have that $\beta_{I,n} \neq 0$ reflecting the fact that the invariant features of X_n have non-zero coefficients across environments by definition. The set $R = \{k \in D; k \notin \{\Gamma_S, \Gamma_I\}\}$ represents the set of all the random predictors in X (no association to y; *i.e.* β_R is random noise).

Note that here $\Gamma_R := \{1, 2, 3, 4\}$ and $\Gamma_I := \{5\}$, and there is no added noise to X in (1). The error is drawn from a normal distribution for simplicity (*i.e.* probit). An example of such a dataset is given in the appendix.

3.2. MNIST

We divide the training set of MNIST into two environments, and keep the test set for evaluating the predictive power of learned sparse representations from training environments. For this experiment, the environments are created by splitting the training set into two evenly shuffled subsets with images and labels corresponding to the 10 digit classification problem. Then we train a *sparse multiclass logistic regression* on the two environments on a range of values for the tuple $(l1, l2, \tau)$, where $l1, l2$ denote the coefficient for $l1, l2$ regularization terms, and τ denotes the agreement threshold among gradients (see section 2.1). So our model would be an Elastic Net regularized network with gradients being computed according to the AND-mask. Evaluation metrics were presented in 2.2. Overlap score is calculated by normalizing the number of weights that are nonzero and have the same index (after rounding to 0.001) in the two sparse representations (obtained from each split), by the number of non-zero weights.

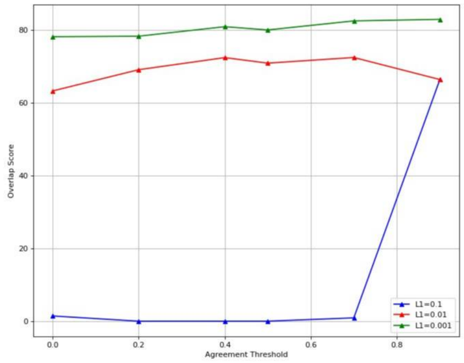


Figure 1. ILC is applied after regularization. Each color corresponds to a value of l1 coefficient and fixed l2, and shows the behavior of sparse representation's consistency (overlap score) averaged over all digits.

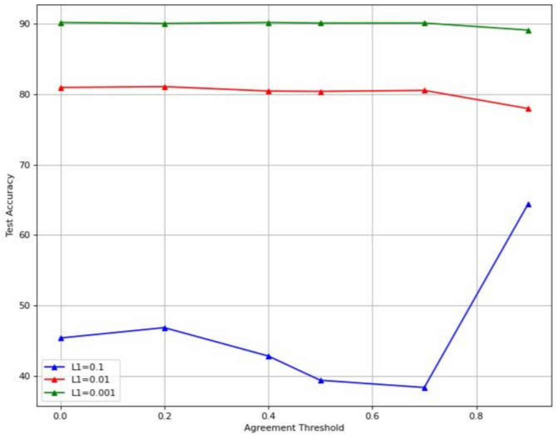


Figure 2. ILC is applied after regularization. Each color shows the trend of test accuracy for varying $l1$ values, with fixed $l2$, over increasing τ.

3.3. fMRI

The dataset is composed of 38,700 3D MRI brain images, each of dimension 53x63x52, taken during a task where subjects must either go or stop following a stimulus that is viewed on a screen directly in front of them. Each subject is scanned three times over a period of five years. The objective of this experiment is to train a model that can predict the event which is a combination of what the subject sees and the action taken, while being invariant to the inherit noise involved in this high dimensional data (53x63x52 = 173,628 variables). We propose to use a fully connected single hidden layer network with both $l1$ and $l2$ regularization such that the recovered sparse model represents the causal graph of brain activation for these events. We compare this model with the same setup, but with ILC applied to enforce invariance across different sampled scans. Data from subjects and the first two timepoints are shuffled together and we test generalization by evaluating the model on the third timepoint from the dataset.

4. Results

4.1. Synthetic

For the sake of brevity, we review the main results and mention interesting auxiliary findings observed through experimentation. First, we note that agreement threshold has a significant impact on the test set performance of ILC, Figure 3. There is considerable improvement in out-of-distribution accuracy when using ILC (agreement $\in$ 0.2,0.4), and we note the sharp rate of increase in test accuracy in the first epochs following the onset of ILC.

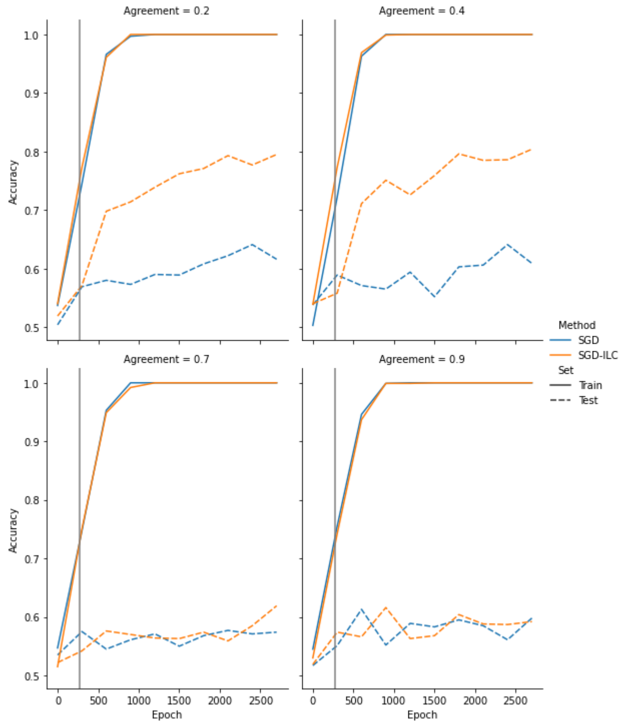


Figure 3. Grey vertical line indicates when ILC is turned on.

Setting the agreement threshold to 0.4, we run additional experiments (Appendix A.2) in order to compare performance under different experimental setups (high-dimensionality

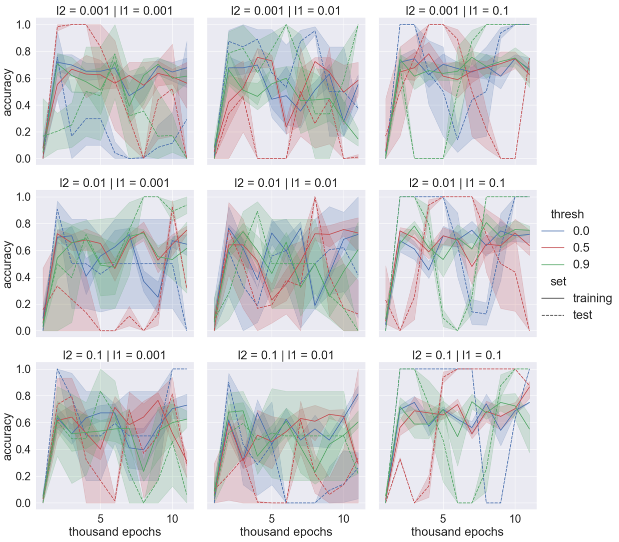


Figure 4. ILC was turned on half way through training. Note that an agreement threshold of 0 means only Adam optimization is used.

and large number of environments, high-dimensionality and small number of environments). The conditions when ILC outperforms traditional regularized Adam is when the number of environments is greater than or equal to the dimensionality of the feature space. For the details of hyperparameters refer to appendix 1.

4.2. MNIST

Results for MNIST using sparse logistic regression are shown in Figures 1-2. Table 2 in the appendix contains the hyperparameters that we have trained our environments on. It should be noted that the lack of distributional shift in MNIST is suboptimal for using invariant representations (see appendix fig. 6 for the measurement of distribution shift.). Thus OoD performance in this setting is not much meaningful and this experiment only serves to showcase the effect of ILC on robustness, not OoD generalization. With that in mind, our findings are as follows:

The overlap score on its own is not very interesting. It becomes valuable when it enables robust prediction (good accuracy). In figures 1,2 for large l1 coefficient and increasing agreement threshold, we see that more consistent sparse representations have been achieved at the expense of accuracy, i.e. very sparse representations yet with *good* predictive power on the test set. This suggests that we have achieved sparse and invariant predictors in the case of MNIST that has very small distribution shift (**??**. Thus we are much looking forward to promising improvements in cases where meaningful distribution shifts occur. As an initial step, we observed such improvements with our synthetic data (See 4.1).

Late start of ILC: We observed that applying this kind of gradient alignment (AND-masking) should not be from the very first epochs. The reason is that masking all gradients in the initial stages where no meaningful direction has been traversed would result in learning nothing and getting stuck in a plateau far from the minima, so we start masking gradients close to the middle of the training.

4.3. fMRI

The results for the fMRI experiment can be found in Figure 4. There is no obvious set of hyperparameters which achieves good classification scores across the five classes, and we note that the volatile nature of the accuracy is due to the imbalanced distribution of classes across batches, which was not controlled for in this case (see Table 3). Considering that we find poor results even with just ElasticNet without ILC (thresh=0), we should not expect ILC to perform better since without any learned useful features we cannot enforce invariance. The complexity of brain fMRI data is very high (the number of voxels is on the order of 100000), and we have approached this challenging problem with only a single fully connected layer network (with the motivation of finding a deterministic causal Bayes net). These results do not rule out the possibility for ILC to improve our understanding of the sparse predictors for brain activities, but exploring more complex architectures may yield better results and is the focus of future work.

5. Conclusion

In this work, we extended the original ILC experiments to a larger set of synthetic experiments, outlining failure and success modes. Beyond this, we applied the method to the popular MNIST dataset and a much more difficult task of predicting events using voxel activations for an fMRI task. More specifically, we showed that for the case of MNIST ILC had a meaningful impact in finding sparse predictors that also contribute to better OoD performance. The synthetic experiments showed that when the number of environments is large compared to the feature space, ILC outperforms ElasticNet. Finally, we found that ILC did not perform significantly better for fMRI by a noticeable margin, however there is a number of considerations which may improve these results. For example, given the insights gleaned from the synthetic dataset, using more environments may benefit ILC especially given the very high dimensionality involved in fMRI. As such, we conclude that ILC should not be used blindly with the expectation of improving OoD performance and sparse variable selection for all datasets, but these experiments suggests an important relationship between the number of environments and the dimensionality of the feature space, as well as the strength of $l1$ regularization.

References

Ahuja, K., Wang, J., Dhurandhar, A., Shanmugam, K., and Varshney, K. R. Empirical or Invariant Risk Minimization? A Sample Complexity Perspective. pp. 1–10, 2010.

Arjovsky, M. *Out of Distribution Generalization in Machine Learning*. PhD thesis, 2020.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization, 2020.

Li, C.-S. R., Huang, C., Constable, R. T., and Sinha, R. Imaging response inhibition in a stop-signal task: neural correlates independent of signal monitoring and post-response processing. *Journal of Neuroscience*, 26(1): 186–192, 2006.

Nagarajan, V., Andreassen, A., and Neyshabur, B. Understanding The Failure Modes of Out-of-Distribution Generalization. pp. 1–25, 2021.

Parascandolo, G., Neitz, A., Orvieto, A., Gresele, L., and Schölkopf, B. Learning explanations that are hard to vary, 2020.

S Chandra Mouli, B. R. Neural Network Extrapolations With G-Invariances From A Single Environment. pp. 1–20, 2021.

Tibshirani, R. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282, 2011.

Vapnik, V. Principles of risk minimization for learning theory. In Moody, J., Hanson, S., and Lippmann, R. P. (eds.), *Advances in Neural Information Processing Systems*, volume 4, pp. 831–838. Morgan-Kaufmann, 1992. URL https://proceedings.neurips.cc/paper/1991/file/ff4d5fbbafdf976cfdc032e3bde78de5-Paper.pdf.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

A. Synthetic task

A.1. Example of synthetic dataset

Below is an example of X where $N = 4$, $D = 5$:

$$X_{4,5} = \begin{bmatrix} 3 & 0 & 0 & 0 & 0.3 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0.3 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}; y_4 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \tag{5}$$

A.2. Hyperparameters for training sparse logistic regression on the synthetic dataset

	Elastic Net	Elastic Net + ILC
log L1 Regularization	1e-4	1e-4
log L2 Regularization	1e-4	1e-4
Agreement Threshold	0	[0.2,0.4,0.7,0.9]

Table 1. Hyperparameters used for training. Optimizer in all cases is Adam, and parameters for optimizers are default: lr=0.001, b1=0.9, b2=0.999. For each experiment, the number of true causal factors was set to 5.

A.3. Additional results:

Effect of late starting ILC 5.

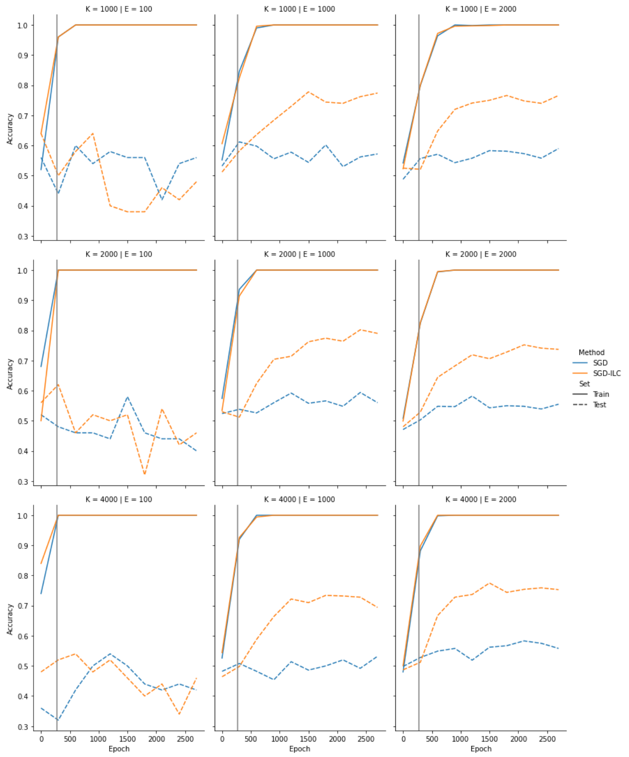


Figure 5. Grey vertical line indicates when ILC is turned on. ILC performs best when the number of predictors is close to or greater than the number of environments.

B. MNIST task

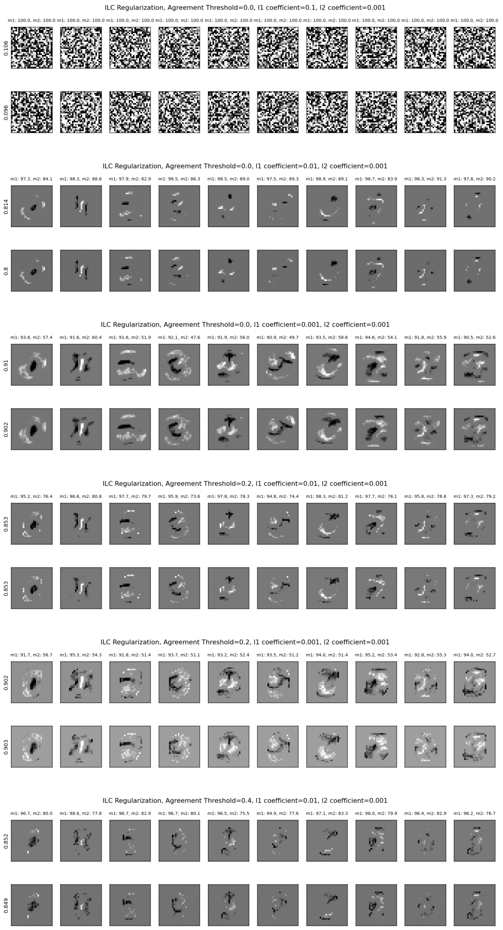
B.1. Hyperparameters for training sparse logistic regression on the two environments on MNIST

	Elastic Net	Elastic Net + ILC
log L1 Regularization	[-1,-2,-3]	[-1,-2,-3]
log L2 Regularization	[-3,-4,-5]	[-3,-4,-5]
Agreement Threshold	0	[0.2,0.4,0.5,0.7,0.9]

Table 2. Hyperparameters used for training, overall there are 54 settings. Optimizer in all cases is Adam, and parameters for optimizers are default: lr=0.001, b1=0.9, b2=0.999, see below for samples of the sparse representations learned.

B.2. Sparse representations achieved for each digit in several sets of hyperparameters

Each row corresponds to an environment or MNIST train split.



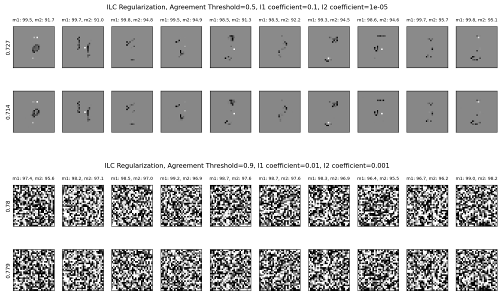


Figure 6 shows that entropy of the predicted class on test set is higher than training set in each environment. However, there is no significant distribution shift across environments.

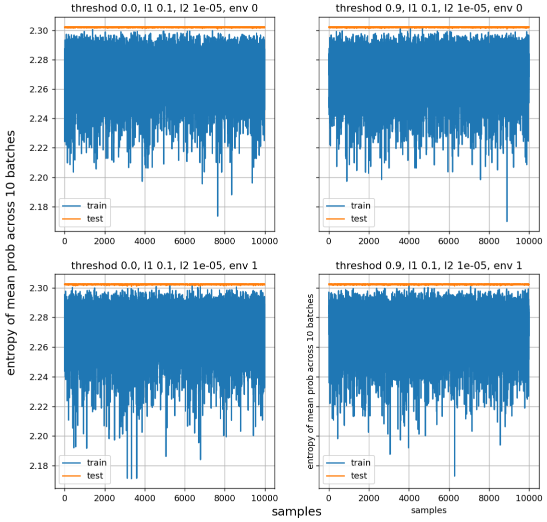


Figure 6. Entropy of of the predicted class for each sample in the test and train batches for two environments after 100 epochs with ILC turned on after epoch 50. The probabilities used for computation were averaged over all batches of size 10000.

C. fMRI task

The task analyzed in this work is the stop-signal task and refer the reader to previous work which details the same methodology used for our dataset (Li et al., 2006). The scans were gathered from three different time points for each subject. There are five distinct events which are used for the classification task and they are: go-success, go-toolate, go-wrong, stop-failure, and stop-success. The data was normalized using z-normalization. The data of subjects from the first two time points were concatenated and shuffled for the training set. The data of the last time point was used for test set. Each data sample was considered an environment in our experiments. There are other curation of environments that could yield more meaningful results, e.g., considering data from each time point as an environment.

C.1. Distribution of classes

	Training Set	Test Set
go-success	4619	2559
go-toolate	293	68
go-wrong	363	73
stop-failure	851	447
stop-success	864	453

Table 3. The number of samples used in training and test set belonging to each class. As we can see the classes are imbalanced.