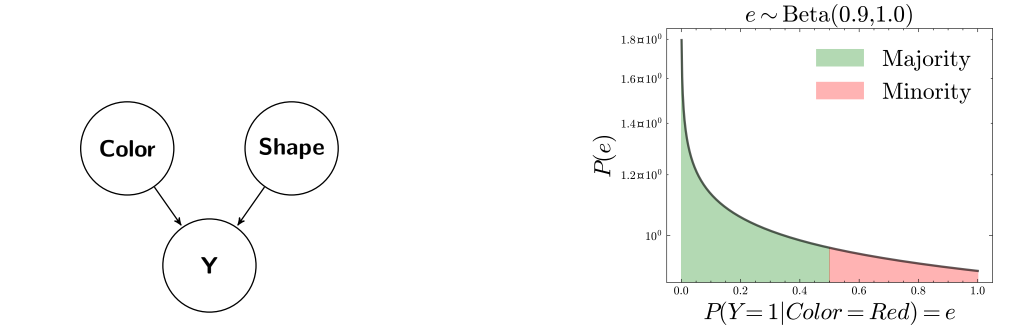

(a) DAG of features and target in CMNIST

(b) Long tail distribution of train environments

Figure 4: In Figure 4a we describe the features that affect the target. The mechanism by which color affects target changes across environments. However, shape has a stable mechanism across environments. In Figure 4b we consider a long tail distribution of environments from which we sample training environments. This is often realistic that many subpopulations are underrepresented in training data, eg low resource languages for translation tasks.

Experiments on Modified CMNIST

C.1. Dataset Setup

We conduct a large-scale experiment using an extension of the CMNIST dataset (Arjovski, 2021). The CMNIST comprises data from the MNIST dataset modified to the task of binary classification. For the standard task in CMNIST, the digits (0-4) and (5-9) have to be classified into two labels 0 and 1. Another feature as color is introduced in the training domain where digits are colored red or green such that the color is predictive of the true label e.g. domain 0.3 i.e. $P(Y = 1 \,|\, \text{color} = \text{red}) = 0.3$ and $P(Y = 0 \,|\, \text{color} = \text{red}) = 0.7$. Whereas for domain 0.9 it would mean $P(Y = 1 \,|\, \text{color} = \text{red}) = 0.9$ and $P(Y = 0 \,|\, \text{color} = \text{red}) = 0.1$. That is the mechanism by which color influences the label changes across domains. However, shape has a stable mechanism of prediction across domains i.e. $P(Y = 0 \,|\, \text{shape} \in \{0, 1, \ldots, 4\}) = 0.75$ and $P(Y = 1 \,|\, \text{shape} \in \{5, 6, \ldots, 9\}) = 0.75$.

C.2. Experimental Setup and Baselines

We consider a scenario where we sample environments from a long-tail distribution at training time to model data collection in the real world, such as low-resource languages. We sample 10 training environments from a Beta(0.9,1) distribution exactly $\{0.0, 0.01, 0.05, 0.07, 0.09, 0.12, 0.14, 0.58, 0.7, 0.99\}$. However, we do not assume IID distribution on environments, i.e. at test time we evaluate all the environments $\{0.0, 0.1, \ldots, 0.9, 1.0\}$. Each environment is assumed to be influenced by both color and shape where the mechanism of color's influence changes but shape affects the target stably. This forces all the precise learners with a fixed hypothesis, i.e., **PL**-f to learn the invariant risk minimizer across domains that rely only on shape as a predictor to generalize to minority domains. We compare performance to baselines (precise learners with fixed hypothesis **PL**-f) based on different assumptions like ERM (average-case risk), GrpDRO (Sagawa et al., 2020), V-REx (Krueger et al., 2021) (worst-case risk) and IRM (Arjovski, 2021), IGA (Koyama and Yamaguchi, 2020) (Invariant Predictors), EQRM (Eastwood et al., 2022a) (probable domain generalizer) and SD (Pezeshki et al., 2021) which avoids implicit regularization from Gradient starvation by decoupling features. We also consider Inf-Task which is a baseline for comparing how an Imprecise Learner (**IL**) performs against precise learners with an augmented hypothesis (**PL**-$\bar{h}$). Based on the initialization setup for CMNIST described by Eastwood et al. (2022a), all baseline methods perform poorly without ERM pretraining. Therefore, to ensure a fair comparison, we consider the ERM pretraining for **PL**-f learners for the initial 400 steps out of a 600-step training. All other hyper-parameters remain consistent with the established setup. For the learners with augmented hypotheses, it does not make sense to initialize with ERM because it may predispose the imprecise learner towards specific outcomes. Therefore, we assess the best-case performance across all learners across types of initialization. To implement the augmented hypothesis, we append FILM layers (Perez et al., 2018) to MLP architecture used in Eastwood et al. (2022a).

Table 2: Maximal regret and test accuracy across all CMNIST test environments.**Bold** denotes the hypothetical best invariant and Bayes classifier performance. Highlighted **Green** denotes the best performance amongst all algorithms for each domain and best regret. Bayes classifier is defined w.r.t the IID learner trained for a particular environment

Objective	**Algorithm**	Test Environments based on $P(Y = 1 \mid \text{color} = \text{red}) = e$											**Regret**
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	
Average Case	ERM	**96.1**	87.1	78.0	72.1	65.8	59.2	51.8	47.1	39.9	33.6	28.3	72.7
Worse Case	GrpDRO	54.1	55.6	58.1	595	61.5	64.5	66.3	69.1	70.5	73.9	**75.5**	46.9
	SD	52.1	54.1	56.6	58.6	59.7	63.7	65.8	67.0	68.5	70.3	73.3	47.9
Invariance	IRM	72.0	72.0	72.0	**72.0**	**72.1**	**69.7**	69.3	69.9	69.2	69.7	67.7	32.3
	IGA	71.8	72.0	72.0	72.1	69.8	65.2	62.4	60.5	57.2	57.7	50.3	49.7
	EQRM ($\lambda \to 1$)	67.8	67.7	68.3	68.8	70.5	69.1	70.3	**72.0**	**72.1**	71.4	72.1	32.2
	VREx	72.7	71.3	71.8	71.4	71.7	69.5	69.5	70.2	69.5	**71.6**	68.5	31.5
	Oracle						**73.5**						**27.9**
PL-$\bar{h}$	Inf-Task	96.0	86.3	78.6	68.0	62.1	61.3	63.2	65.0	66.6	68.4	68.3	31.7
IL (Ours)	IRO	95.8	87.2	78.8	68.9	69.4	69.5	**70.8**	70.1	70.0	70.4	70.3	**29.7**
Bayes Classifier	ERM (IID)	**100.0**	**90.0**	**80.0**	**75.0**	**75.0**	**75.0**	**75.0**	**75.0**	**80.0**	**90.0**	**100.0**	

C.3. Imprecise Learner can learn relevant features in context

In Table 2 we compare **IL** to other methods, showing that **IL** can learn relevant features in context. This also allows us to guide users on how to select appropriate λ. Suppose the user expects data at test time to come from the majority environments of their training. In that case, they can be less risk averse and use $\lambda = 0$ whereas if the user is unsure and anticipates test environments to look like unlike training, i.e. more minority environments they can choose $\lambda \to 1$. This is also reflected in the performance of **IL** such that for the majority domains $e \in \{0.0, \dots, 0.4\}$ it performs similar to average case learner and for relatively less seen i.e. minority domains $e \in \{0.5, \dots, 1.0\}$ it performs similar to the invariant learner.

D. Limitations of Imprecise Learner

D.1. Computational Complexity

The additional computation costs result from solving (9) compared to solving for a single notion of generalization which grows by the $O(m)$ where m is the number of estimates needed. Since the convergence rate for Monte Carlo estimates is $O(\frac{1}{\sqrt{m}})$ the quality of estimates of the gradient improves slowly w.r.t. the number of samples. The generalization to the user's choice of risk λ_{op} with high probability is also given by $O(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}})$ in Proposition 4.2, where n is the number of data samples from each environment. In practice, there is room to obtain a better approximation of (9) with possibly quasi-Monte Carlo sampling methods.

D.2. Challenges in Specifying User Preferences

One of the main challenges in the Imprecise Learning (**IL**) framework is to specify user preference in terms of risk level i.e. a choice of λ_{op} in practical scenarios where a user may want to generalize to a domain. We recommend a naive strategy that a user can be more risk averse if they wish to generalize to domains from minority environments whereas for generalizing to a domain from majority environments users can be more risk-seeking.

D.3. Generalization with no access to minority environments

In the context of the standard CMNIST setup where the learner has access to no minority environments, CVaR as a risk measure does not allow to generalize beyond the credal set which can be constructed from the convex combination of majority environments alone. For standard CMNIST setup training envs are $\{0.1, 0.2\}$ and test env is $\{0.9\}$. This means that the mechanism by which color affects the target is anti-correlated at test time, such situations can arise in adversarial settings. Since for $\lambda \to 1$, CVaR only minimizes the higher risks in a profile to achieve invariance it cannot recover the invariant mechanism without access to at least one environment from a subgroup. However, we argue that by using additional assumptions i.e. a different risk measure Imprecise learners can still learn to generalize to novel unseen domains outside of the credal set. We can extend the risk measure to enforce invariance by using VREx as an additional regularizer.

$$\rho_\lambda[\boldsymbol{\mathcal{R}}] := CVaR_\lambda[\boldsymbol{\mathcal{R}}] + \lambda Variance(\boldsymbol{\mathcal{R}}) \tag{15}$$

In Table 3, we observe that **IL** for $\lambda = 1$ obtains poor performance on a novel test domain however with an additional risk measure it obtains a closer performance to ERM on grayscale (Oracle) and outperforms several baselines. Note that with random initialization **IL**+VREx significantly outperforms other baselines.

Table 3: CMNIST Test Accuracy. Training Environments are $\{0.1, 0.2\}$ & Test Environment $\{0.9\}$

Objective	**Algorithm**	**Initialization**		
		Rand.	ERM	Best Case
	ERM	27.9 ± 1.5	27.9 ± 1.5	27.9 ± 1.5
	IRM	52.5 ± 2.4	69.7 ± 0.9	69.7 ± 0.9
	GrpDRO	27.3 ± 0.9	29.0 ± 1.1	29.0 ± 1.1
PL-f	SD	49.4 ± 1.5	70.3 ± 0.6	70.3 ± 0.6
	IGA	50.7 ± 1.4	57.7 ± 3.3	57.7 ± 3.3
	V-REx	55.2 ± 4.0	**71.6 ± 0.5**	**71.6 ± 0.5**
	EQRM	53.4 ± 1.7	71.4 ± 0.4	71.4 ± 0.4
IL	IRO	28.4 ± 0.7	27.4 ± 0.1	28.4 ± 0.7
PL-$\bar{h}$+VREX	Inf-Task	68.4 ± 0.1	64.6 ± 0.0	68.4 ± 0.1
IL+VREX	IRO	**71.4 ± 0.2**	65.4 ± 0.1	71.4 ± 0.2
Invariant Pred.	Oracle		**73.5 ± 0.2**	