Domain Generalisation via Imprecise Learning

Anurag Singh [1] **Siu Lun Chau** [1] **Shahine Bouabid** [2] **Krikamol Muandet** [1]

Abstract

Out-of-distribution (OOD) generalisation is challenging because it involves not only learning from empirical data, but also deciding among various notions of generalisation, e.g., optimising the average-case risk, worst-case risk, or interpolations thereof. While this choice should in principle be made by the model operator like medical doctors, this information might not always be available at training time. The institutional separation between machine learners and model operators leads to arbitrary commitments to specific generalisation strategies by machine learners due to these deployment uncertainties. We introduce the Imprecise Domain Generalisation framework to mitigate this, featuring an imprecise risk optimisation that allows learners to stay imprecise by optimising against a continuous spectrum of generalisation strategies during training, and a model framework that allows operators to specify their generalisation preference at deployment. Supported by both theoretical and empirical evidence, our work showcases the benefits of integrating imprecision into domain generalisation.

1. Introduction

The capability to generalise knowledge, a hallmark of both biological and artificial intelligence (AI), has seen remarkable progress in recent years. Developments in general-purpose learning algorithms (Vapnik, 1991; Hofmann et al., 2008; LeCun et al., 2015; Goodfellow et al., 2016), model architectures (Krizhevsky et al., 2012; Cohen and Welling, 2016; Vaswani et al., 2017), and training infrastructures (Ratner et al., 2019) have given rise to AI systems such as generative models (GenAI) and large language models (LLM) that surpass human-level generalisation capabilities in specific domains.

Despite notable achievements, these systems may catastrophically fail when operated on out-of-domain (OOD) data because theoretical guarantees for their generalisation hinge on the assumption of independent and identically distributed (IID) training and deployment data, with empirical risk minimisation (ERM) being the dominant learning algorithm (Vapnik, 1991; 1995). Emerging challenges like distribution shifts (Quionero-Candela et al., 2009; Beery et al., 2018; 2020; Koh et al., 2021), adversarial attacks (Szegedy et al., 2013; Goodfellow et al., 2014), and strategic manipulations (Hardt et al., 2016; Perdomo et al., 2020; Vo et al., 2023) have prompted researchers to question the validity of algorithms developed under this assumption. This gap has fueled interest in OOD generalisation, prompting the exploration of novel learning algorithms and resulting in rapid developments in domain adaptation (Wilson and Cook, 2020; Zhao et al., 2022), domain generalisation (Wang et al., 2021b; Zhou et al., 2023; Shen et al., 2021), and test-time adaptation (Sun et al., 2020; Wang et al., 2021a; Chen et al., 2023a), among others.

In IID generalisation, where test loss aligns with training loss, the learner's goal of minimising the training loss aligns with the operator's expectation of small test loss. Bounded data uncertainty, within finite data, enables the learner to assess model generalisation during deployment. Historically, the IID assumption is accompanied by another critical, but often overlooked assumption: the overlap between the learner and the operator, who employs the model in real-world contexts. Conversely, OOD generalisation still lacks a precise definition, leading to additional ambiguity termed "generalisation uncertainty". Unlike data uncertainty, generalisation uncertainty arises from a lack of knowledge about deployment environments, whether due to natural shifts (across hospitals, experimental conditions, and time) or artificial ones (adversarial attacks, strategic manipulation), and cannot be mitigated by additional data collection.

Prior research has addressed generalisation uncertainty independently by introducing various concepts of OOD generalisation including worst-case generalisation (Arjovsky et al., 2019; Ben-Tal et al., 2009; Sagawa et al., 2020; Krueger et al., 2021), average-case generalisation (Blan-

Anurag Singh is part of the Graduate School of Computer Science at Saarland University, Saarbrücken, Germany. [1]CISPA Helmholtz Center for Information Security, Saarbrücken, Germany [2]Department of Statistics, University of Oxford, UK. Correspondence to: Anurag Singh <anurag.singh@cispa.de>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

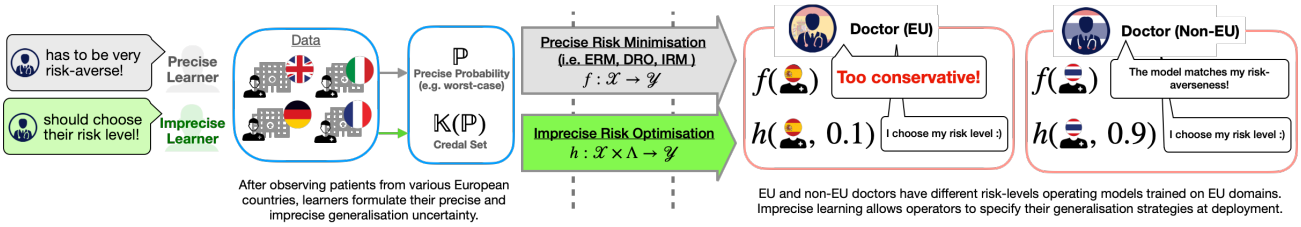


Figure 1: An illustration of our proposed *imprecise learning* framework. We allow learners to stay imprecise to avoid over-commit in light of generalisation uncertainty. Instead, we defer this choice of precise generalisation to the operator.

chard et al., 2011; 2021; Muandet et al., 2013; Zhang et al., 2021), and their interpolations (Eastwood et al., 2022a). Learning algorithms like distributional robust optimisation (DRO) (Rahimian and Mehrotra, 2022), invariant risk minimisation (IRM) (Arjovsky et al., 2019), and quantile risk minimisation (QRM) (Eastwood et al., 2022a) have been tailored for these OOD generalisation notions. This line of research relaxes the IID assumption, but still assumes alignment between the learner's objective and the operator's goal to tackle generalisation uncertainty. Due to the need for precise concept of generalisation in these scenarios, we collectively term them "precise generalisation".

Precise generalisation hinges on the assumption that the learner's objective aligns with the operator's goal, necessitating close collaboration during model development. However, this approach presents two primary drawbacks. Firstly, institutional separation between the learner (e.g., machine learning engineers) and model operators (e.g., doctors) can make collaboration costly, time-consuming, or even impractical. Secondly, tailoring the model to a specific operator may restrict its deployment usability, as the operator's beliefs can change or conflicts may emerge when the model is operated by a different individual. Consider an example depicted in Figure 1. Using data obtained from hospitals across Europe, an engineer is developing a machine learning model that will be embedded into a medical software that will be used by the doctors. Here, the engineer confronts uncertainty regarding where the model will ultimately be deployed—it could be within Europe (IID) or outside it (OOD). The engineer might anticipate the doctor's generalisation strategy during the model's training phase. For instance, if the doctor is perceived to be risk-averse, the engineer might prioritise training a model robust to worst-case scenarios. However, ideally, it should be the doctor, often equipped with domain-specific expertise, who decides the generalisation strategy, drawing upon their in-depth knowledge of the field, at deployment time. Customising models effectively to the clinical settings where they operate can significantly impact healthcare outcomes (Beede et al., 2020).

In this work, we extend the relaxation of the IID assumption further by loosening the requirement for overlap between the learner and the operator. Since there is no need of specific concept of generalisation at training time, we term this scenario "imprecise generalisation" (see Figure 1). We operationalise imprecise learning in the context of domain generalisation (Blanchard et al., 2011; 2021; Muandet et al., 2013), aiming to answer the question: *How to take knowledge acquired from an arbitrary number of related domains and apply it to previously unseen domains?* This concept comprises two main components: (1) An optimisation process enabling learners to remain imprecise during learning, thus not committing to a specific generalisation notion during training, and (2) a model framework allowing operators to define their preferred generalisation strategy at deployment. We delve into the formulation and existing work on OOD generalisation in Section 2. Our primary contribution, the framework of Imprecise Domain Generalisation, is detailed in Section 3, along with its optimisation strategy, termed Imprecise Risk Optimisation (IRO), in Section 4. Experimental results are presented in Section 5, and we conclude our paper in Section 6.

All proofs are in the appendix and we open-source our code at `https://github.com/muandet-lab/dgil`.

2. Preliminaries

Consider $\mathcal{X} \subseteq \mathbb{R}^d$ as our instance space and $\mathcal{Y}$ as our target space, where $\mathcal{Y} \subseteq \mathbb{R}$ is used for regression and $\mathcal{Y} = 1, \dots, C$ for C-class classification. In supervised learning, the process of learning a function mapping from $\mathcal{X}$ to $\mathcal{Y}$ involves the learner specifying their inductive biases. These inductive biases include: (1) selecting a hypothesis class $\mathcal{H}$, consisting of functions $f : \mathcal{X} \to \mathcal{Y}$, (2) defining a suitable loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ based on the problem, (3) assuming the presence of a joint probability distribution $\mathbb{P}$ over the variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ from which the data are sampled. Most critical to our work are (4) the assumptions regarding the deployment environment where the model f is expected to generalise.

2.1. Precise Learning

In the following, we briefly review various generalisation assumptions commonly adopted in the literature and unify

them under the setting of *precise learning*.

IID assumption. Perhaps the most fundamental generalisation assumption in supervised learning is that the training and deployment environments are independent and identically distributed (IID). Under this assumption, a model that performs well in training is expected to generalize effectively in deployment. This concept is formalized by finding the function $f \in \mathcal{H}$ that minimizes the population risk for $\mathbb{P}$, known as the Bayes optimal model:

$$\mathcal{R}(f) \triangleq \mathbb{E}_{\mathbb{P}}[Z_f] = \mathbb{E}_{(X,Y)\sim\mathbb{P}}\big[\ell(f(X), Y)\big]. \quad (1)$$

For simplicity, we have denoted $Z \triangleq (X, Y)$, $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$, and $Z_f \triangleq \ell(f(X), Y)$ as the random loss associated with $f \in \mathcal{H}$. In practice, since the true distribution $\mathbb{P}$ is unknown, we focus on minimizing an empirical estimate of this risk based on IID samples $(x_i, y_i)_{i=1}^n$ from $\mathbb{P}$, expressed as:

$$\widehat{\mathcal{R}}(f) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i) + \eta\|f\|_{\mathcal{H}}^2, \quad f \in \mathcal{H}, \quad (2)$$

where the second term is a regularization term to prevent overfitting, following the empirical risk minimization (ERM) principle (Vapnik, 1991; 1995). This scenario introduces **data uncertainty**, stemming from the finite nature of data when approximating Bayes optimal models. The IID assumption also enjoys favourable guarantees, e.g., as the sample size n increases, the uniform convergence of $\widehat{\mathcal{R}}(\cdot)$ over $\mathcal{H}$ ensures that the gap between the empirical and the population risk becomes negligible with high probability; see, e.g., Vapnik (1998); Cucker and Smale (2002).

Beyond IID assumptions. The IID assumption is often not viable in real-world scenarios due to various factors such as distribution shifts (Quionero-Candela et al., 2009; Beery et al., 2018; 2020; Koh et al., 2021), sub-population shifts (Santurkar et al., 2021; Yang et al., 2023), adversarial attacks (Szegedy et al., 2013; Goodfellow et al., 2014), strategic manipulation (Hardt et al., 2016; Perdomo et al., 2020; Vo et al., 2023), and time shifts (Gagnon-Audet et al., 2022). In response to these challenges, learners must consider **generalisation uncertainty** when designing their learning algorithm. This uncertainty is typically represented by a credal set $\mathbb{K}(Z)$ (Walley, 1991), a closed set of potential probability distributions that reflect the learner's ignorance, or partial knowledge about the deployment environments.

For example, in distributionally robust optimisation (Rahimian and Mehrotra, 2022), the credal set comprises distributions within an ϵ distance from the empirical distribution, and the goal is to optimise f for the worst-case empirical risk within it. Another approach, involves learning across multiple domains $\mathbb{P}_1, \ldots, \mathbb{P}_d$, and assumes the deployment distribution lies within their convex hull (Mansour et al., 2012; Krueger et al., 2021; Föll et al., 2023). In invariant causal prediction (Peters et al., 2016; Heinze-Deml et al., 2018), hypothetical interventional distributions associated with a structural causal model (SCM) constitute the credal set. Learning algorithms here aim to optimize for worst-case empirical risk (Arjovsky et al., 2019; Ben-Tal et al., 2009; Sagawa et al., 2020; Krueger et al., 2021), average-case empirical risk (Blanchard et al., 2011; 2021; Muandet et al., 2013; Zhang et al., 2021), and interpolations thereof (Eastwood et al., 2022a). The choice of risk corresponds to selecting a particular distribution within the credal set, such as the centroid of the convex hull referring to the average case. Notably, the credal set in the IID case reduces to a single distribution, $\mathbb{K}(Z) = \{\mathbb{P}\}$.

2.2. Previous Work

Limitation of precise learning. A majority of previous work in both IID and OOD generalisation falls into the precise learning setting. A fundamental requirement is for the learner to commit to a specific notion of generalisation. This involves precisely selecting a particular distribution $\mathbb{P} \in \mathbb{K}(Z)$ during training and performing statistical learning to develop the model f. Although widely used, this might not always be optimal in modern machine learning settings, especially when there is a clear institutional separation between those who build and those who operate the model (cf. Section 3). This separation presents two significant challenges. First, it assumes that the learners either fully understand the specific generalisation needs of the operators, or that the operators have comprehensive access to the datasets and a thorough understanding of statistical inference, effectively making them the learners. Second, the choice of generalisation strategy is inherently subjective, involving normative decisions by the operators. For instance, a risk-averse operator might lean towards a worst-case empirical risk optimiser, while an operator with in-depth knowledge of the deployment environment might prefer an average-case empirical risk optimiser.

Domain generalisation strategies. The core of domain generalisation is the invariance principle (Muandet et al., 2013; Arjovsky, 2019), which asserts that certain properties remain constant across different environments and thus are expected to generalise to unseen settings. This principle is reflected in approaches focusing on feature representation (Muandet et al., 2013; Ghifary et al., 2015; Arjovsky et al., 2019), causal mechanism (Peters et al., 2016; Rojas-Carulla et al., 2018; Heinze-Deml and Meinshausen, 2021), and risk functional (Krueger et al., 2021), all aimed at identifying and leveraging these invariant properties. While necessary, this principle faces two challenges: it abstracts away the inherent heterogeneity across environments (Heckman, 2001), which might give rise to non-invariant yet generalisable properties (Eastwood et al., 2023; Nastl and Hardt, 2024). Furthermore, identifying and utilising invariant properties

faces practical difficulties due to their need for large sample sizes (Rosenfeld et al., 2021; Kamath et al., 2021).

Addressing these challenges, recent work suggests combining domain-invariant with domain-specific properties (Liu et al., 2021a; Mahajan et al., 2021). While these approaches have been shown to improve in-domain generalisation performance, how domain-specific properties affect OOD generalisation in unseen environments remains unclear. To overcome this, it is popular to utilise various forms of test-time adaptation via auxiliary tasks (Sun et al., 2020; Wang et al., 2021a; Chen et al., 2023a). However, Liu et al. (2021b) has shown that this strategy can improve the pre-trained model only when the auxiliary loss aligns with the main loss. This suggests that a certain degree of precision in aligning losses is essential for effective domain generalisation.

Generalisation uncertainty representation. As opposed to the credal set $\mathbb{K}(Z)$, some authors have instead adopted a second-order probability (aka meta distributions) as a belief over the "true" or "ideal" probabilities $\mathbb{P}(Z)$ (Blanchard et al., 2011; 2021; Muandet et al., 2013; Eastwood et al., 2022a). However, Walley (1991, Sec. 5.10) pointed out that if probability distributions entail behavioural dispositions, it is necessary that the credal set must collapse to a singleton to avoid incoherent behaviour. Paradoxically, this implies that assuming the existence of meta-distributions is equivalent to making the IID assumption in the first place, emphasising that one must clearly differentiate generalisation uncertainty from data uncertainty.

Learning under imprecision. Machine learning inherently grapples with imprecision due to its inductive nature. One common approach to mitigate this is to create precision at various stages of model development. Techniques like data up/downsampling (He and Garcia, 2009) and fusion (Chau et al., 2021a;b) address issues of granularity and missing data by drawing information from a precise empirical distribution. During algorithm selection, approaches like the Bayesian paradigm, ensembling, and AutoML (He et al., 2021) are used to handle potential model misspecification by selecting a precise model from a set of alternatives. Furthermore, model deployment requires a precise definition of generalisation, such as optimising for average-case or worst-case risks, to be determined before training.

When the introduced precision is not warranted, imprecise probabilists advocate for learning *along with* imprecision. For instance, Walley's Imprecise Dirichlet Model effectively handles incomplete and missing data (Utkin et al., 2021). Dempster-Shafer Theory (Shafer, 1992) enables the fusion of multiple information sources, considering all available evidence. Credal learners, including credal decision trees (Abellan and Masegosa, 2010), credal networks (Cozman, 2000), and imprecise Bayesian neural networks (Caprio et al., 2023), propagate imprecision to prediction, resulting in models that capture the full range of possible outcomes. Central to these methods is the concept of a set of permissible solutions. This approach leads to indeterminate yet credible models, particularly in domains where uncertainty is prevalent. Our research aligns with this line of work, focusing on developing domain generalisation strategies that acknowledge and adapt to imprecision. By embracing imprecision, we aim to create models that offer a range of permissible solutions, empowering model operators to make informed choices at test time. The use of a credal set to model epistemic uncertainty has been concurrently explored by Caprio et al. (2024) to derive generalization bounds under credal uncertainty.

3. Imprecise Domain Generalisation

In this work, we advocate for an *imprecise learning*, where learners do not commit to any particular $\mathbb{P} \in \mathbb{K}(Z)$ at training time, but express their uncertainty through a credal set $\mathbb{K}(Z)$, where we discuss our choice in Section 3.2. We operationalise this idea in the context of domain generalisation (DG) problems. To this end, consider data coming from d distinct domains, each with its own distribution $\mathbb{P}_1, \ldots, \mathbb{P}_d$, and corresponding risk profiles $(\mathcal{R}_1, \ldots, \mathcal{R}_d) \triangleq \boldsymbol{\mathcal{R}}$. The learner's objective is to select an optimal hypothesis from $\mathcal{H}$ considering both the risk profiles and $\mathbb{K}(Z)$, based on a certain optimality criterion defined below. While we mainly focus on multi-domain environments, this framework is also relevant and adaptable to single-domain scenarios (see Appendix C for further discussion).

Credal set and partial preference. A crucial distinction between precise and imprecise learning lies in their approach to learner's preferences (Chau et al., 2022a;b). Precise learners commit to a specific distribution $\mathbb{P} \in \mathbb{K}(Z)$ during training, creating a complete[1] and transitive preference order $\succeq$ based on empirical risk in $\mathcal{H}$. That is, for any $f, g \in \mathcal{H}$, $f \succeq g$ if and only if $\widehat{\mathcal{R}}(f) \leq \widehat{\mathcal{R}}(g)$. Conversely, imprecise learning based on the credal set $\mathbb{K}(Z)$ results in a *partial* order over $\mathcal{H}$ (Giron and Rios, 1980; Walley, 1991):

Lemma 3.1. *The binary relation $\succeq$ represented by $\mathbb{K}(Z)$ is such that for $f, g \in \mathcal{H}$, $f \succeq g$, if and only if $\mathbb{E}_{\mathbb{P}}[Z_f] \leq \mathbb{E}_{\mathbb{P}}[Z_g]$ for every $\mathbb{P} \in \mathbb{K}(Z)$.*

This leads to an *incomplete* preference ordering. Lemma 3.1 highlights the challenge of learning with imprecision, implying that unless the learners are willing to exert their judgement over the distributions in $\mathbb{K}(Z)$, as was previously done in precise learning, it is no longer possible to unanimously identify the "best" hypothesis in $\mathcal{H}$ from the observed data alone. In the following, we describe how the learners can implement imprecise learning at training time such that the operators can make prediction efficiently at test time.

[1] For every $f, g \in \mathcal{H}$, either $f \succeq g$, $g \succeq f$, or both hold.

3.1. Aggregation Functions and Optimality Criteria

To facilitate learning, we need a certain notion of optimality taking into account $\mathbb{K}(Z)$. We formalise this by considering an *aggregated learning algorithm* $\mathfrak{P} : \boldsymbol{\mathcal{R}} \mapsto h^*$ that takes a risk profile and returns a hypothesis $h^* \in \mathcal{H}$. In particular, we focus on a specific type of aggregation function called an aggregated risk minimizers:

$$\mathfrak{P} : \boldsymbol{\mathcal{R}} \mapsto \arg\min_{\theta\in\Theta} \rho_\lambda[\boldsymbol{\mathcal{R}}](h_\theta), \quad \lambda \in \Lambda, \tag{3}$$

where $\rho_\lambda : \mathcal{L}_2^d(\mathcal{H}) \to \mathcal{L}_2(\mathcal{H})$ is a risk aggregation function indexed by $\lambda \in \Lambda$, which yields the non-negative real-valued statistical functional $\rho_\lambda[\boldsymbol{\mathcal{R}}] : \mathcal{H} \to \mathbb{R}_+$. Here, we assume that our model class $\mathcal{H}$ is parametrized by a parameter space $\Theta \subseteq \mathbb{R}^p$, e.g., a weight vector in a neural network. We call Λ a choice space which arises exclusively due to the imprecision of the learning problem (cf. Lemma 3.1) and serves merely as an index set. In practice, we only have access to the empirical risks $(\widehat{\mathcal{R}}_1, \ldots, \widehat{\mathcal{R}}_d) \triangleq \widehat{\boldsymbol{\mathcal{R}}}$ which we can substitute directly into (3). In Section 3.2, we consider Conditional Value-at-Risk (CVaR) as a concrete example of the risk aggregator ρ_λ. Our formulation (3) is not only pertinent to financial risk measures but has also gained traction for creating interpretable, risk-aware machine learning algorithms (Williamson and Menon, 2019).

For each $\lambda \in \Lambda$, we denote the Bayes optimal models by $h_\lambda^* \in \arg\min_{\theta\in\Theta} \rho_\lambda[\boldsymbol{\mathcal{R}}](h_\theta)$ and the associated parameter by θ_λ^*. Unfortunately, for continuous choice space, it is unrealistic for the learner to find the Bayes optimal models in $\mathcal{H}$ simultaneously for all $\lambda \in \Lambda$. For this reason, we generalise the notion of Pareto optimality (Pareto, 1897) from multi-objective optimisation to its continuous counterpart and propose an alternative optimality criterion with respect to all $\lambda \in \Lambda$: C-Pareto optimality.

Definition 3.2 (C-Pareto optimality)**.** The hypothesis h_θ dominates $h_{\theta'}$, denoted by $h_\theta \triangleright h_{\theta'}$, if $\rho_\lambda[\boldsymbol{\mathcal{R}}](h_\theta) \leq \rho_\lambda[\boldsymbol{\mathcal{R}}](h_{\theta'})$ for all $\lambda \in \Lambda$ and $\rho_{\tilde{\lambda}}[\boldsymbol{\mathcal{R}}](h_\theta) < \rho_{\tilde{\lambda}}[\boldsymbol{\mathcal{R}}](h_{\theta'})$ for some $\tilde{\lambda} \in \Lambda$. Then, h_θ is C-Pareto optimal if there exists no $h_{\theta'}$ such that $h_{\theta'} \triangleright h_\theta$.

When λ takes values on a finite set Λ with m elements, i.e., $\Lambda = \{\lambda_1, \ldots, \lambda_m\}$, Definition 3.2 coincides with the Pareto optimality in standard multi-objective optimization (MOO); see, e.g., Sener and Koltun (2018); Lin et al. (2019); Zhang and Golovin (2020); Ma et al. (2020) and references therein. Chen et al. (2023b) have recently studies trade-offs between ERM and existing OOD objectives using MOO.

It is not hard to show that, like $\succeq$ introduced in Lemma 3.1, $\triangleright$ can be incomplete and that any Bayes optimal models h_λ^* are also C-Pareto optimal. Intuitively, instead of obtaining the Bayes optimal model for all $\lambda \in \Lambda$, the learner can at best find the non-dominating models, i.e., the models upon which an improvement is only possible at a cost of deterioration of another non-dominating model.

Next, we introduce the notion of C-Pareto stationary used to check if a model is C-Pareto optimal.

Definition 3.3 (C-Pareto stationary)**.** Suppose $\rho_\lambda[\boldsymbol{\mathcal{R}}](h_\theta)$ is a smooth function of h_θ and define the local gradient at $h_\diamond$ as $v_\lambda := \nabla\rho_\lambda[\boldsymbol{\mathcal{R}}](h_\diamond)$. The point $h_\diamond$ is called C-Pareto stationary if and only if there exists a probability density q such that $\int v_\lambda \, dq(\lambda) = 0$.

3.2. Conditional Value-at-Risk (CVaR)

In theory, all aggregation functions $\rho_\lambda[\boldsymbol{\mathcal{R}}]$ can be expressed as a type of weighted average of $\boldsymbol{\mathcal{R}}$, as detailed in Proposition B.1. A high level of generality could be achieved by formulating $\mathbb{K}(Z)$ as the convex hull of $\mathbb{P}_1, \ldots, \mathbb{P}_d$. This corresponds to treating the choice parameter $\lambda \in \mathbb{R}^d$ as all possible averaging weight, thus defining $\rho_\lambda[\boldsymbol{\mathcal{R}}] = \lambda^\top\boldsymbol{\mathcal{R}}$. However, this approach has its serious drawbacks, since λ might be difficult for the operators to interpret, potentially leading to *irrational* decisions. For instance, operators may inappropriately assign more weight to domains that are easier to train, resulting in atypical "risk-seeking" behaviour.

To select an appropriate aggregation function (equivalent to formulating an appropriate credal set) that is both interpretable and aligned with typical behaviour such as risk aversion, we opt for ρ_λ from the class of risk measures. This corresponds to formulating credal set as distributions that are mixtures of $\mathbb{P}_1, \ldots \mathbb{P}_d$ with weights determined by the aggregation function. Notably, we choose the Conditional Value-at-Risk (CVaR):

Definition 3.4 (Conditional Value-at-Risk (Rockafellar and Uryasev, 2002))**.** Let $\boldsymbol{\mathcal{R}} = (\mathcal{R}_1, \ldots, \mathcal{R}_d)$ represent our risk profile, and $F_{\boldsymbol{\mathcal{R}}}(r) = \frac{1}{d}\sum_{i=1}^d \mathbb{I}[\mathcal{R}_i \leq r]$ as the cumulative distribution function (CDF) for $\boldsymbol{\mathcal{R}}$. Define $r_\lambda = \min_r\{r \mid F_{\boldsymbol{\mathcal{R}}}(r) \geq \lambda\}$ as the λ-level quantile. Then, the Conditional Value-at-Risk for $\boldsymbol{\mathcal{R}}$ at level λ is given by:

$$\sum_{i=1}^d \left(\eta_\lambda \mathbb{I}[\mathcal{R}_i = r_\lambda] + \frac{(1-\eta_\lambda)\mathbb{I}[\mathcal{R}_i \geq r_\lambda]}{\sum_{i=1}^d \mathbb{I}[\mathcal{R}_i \geq r_\lambda]}\right)\mathcal{R}_i \tag{4}$$

where $\eta_\lambda = (F_{\mathcal{R}}(r_\lambda) - \lambda)(1-\lambda)^{-1}$, indicating the discontinuity level of the CDF at λ.

CVaR effectively enables operators to express their level of risk aversion through λ, which in turn influences the selection of riskier domains for optimization. Additionally, this approach provides a means to transition smoothly between two prevalent notions of generalisation (Robey et al., 2022; Eastwood et al., 2022a; Li et al., 2023), namely optimising average risks ($\lambda = 0$) and worst-case risks ($\lambda = 1$). Furthermore, CVaR belongs to a class of coherent risk measures, which possess desirable properties (Artzner et al., 1999) and

have been studied in the robust optimisation literature; see, e.g., Ben-Tal et al. (2010).

3.3. Augmented Hypothesis

To further institutionalize the separation of statistical decision-making, i.e., choosing appropriate notion of generalisation (performed by the operator) from statistical learning (performed by the learner), we propose to shift the problem view to an imprecise setting where the learner does not assume a priori which $\lambda \in \Lambda$ is relevant to the operator, but instead designs a model that allows the operator to choose their own λ at deployment time.[2]

To this end, we extend the hypothesis space to an augmented hypothesis space $\mathcal{H}_\Lambda$ of functions of both x and λ, and propose to learn an *augmented* hypothesis $\bar{h}_\xi : \mathcal{X} \times \Lambda \to \mathcal{Y}$ parametrized by a parameter $\xi \in \Xi \subseteq \mathbb{R}^q$. In contrast with $h_\theta \in \mathcal{H}$ which is fixed across $\lambda \in \Lambda$, an augmented hypothesis $\bar{h}_\xi \in \mathcal{H}_\Lambda$ describes a range of possible hypothesis $\bar{h}_\xi(\cdot, \lambda)$ for each $\lambda \in \Lambda$, such that the user can choose the one that best fits their needs. By abuse of notation, we consider $\rho_\lambda[\boldsymbol{\mathcal{R}}](\bar{h}_\xi(\cdot, \lambda))$ as a point-wise aggregated risk for the augmented hypothesis $\bar{h}_\xi \in \mathcal{H}_\Lambda$. The subtle difference here is that we evaluate the objective at $\bar{h}_\xi(\cdot, \lambda)$ for the same λ used in ρ_λ. While the idea of augmented hypothesis with loss-conditional learning has previously been considered (Brault et al., 2019; Dosovitskiy and Djolonga, 2020), existing work still fall into the setting of precise learning, as we describe subsequently in Section 4.

The function $h^* : (x, \lambda) \mapsto h^*_\lambda(x)$ that maps onto a Bayes optimal model for each $\lambda \in \Lambda$ is an example of such augmented hypothesis. However, we may again prefer to consider a more amenable optimality criterion that seeks optimality jointly across Λ. To this end, we extend Definition 3.2 to an augmented hypothesis.

Definition 3.5 (C-Pareto optimal augmented hypothesis)**.** The augmented hypothesis $\bar{h}_\xi$ dominates $\bar{h}_{\xi'}$, denoted $\bar{h}_\xi \triangleright \bar{h}_{\xi'}$, if $\rho_\lambda[\boldsymbol{\mathcal{R}}]\big(\bar{h}_\xi(\cdot, \lambda)\big) \leq \rho_\lambda[\boldsymbol{\mathcal{R}}]\big(\bar{h}_{\xi'}(\cdot, \lambda)\big)$ for all $\lambda \in \Lambda$ and $\rho_{\tilde{\lambda}}[\boldsymbol{\mathcal{R}}]\big(\bar{h}_\xi(\cdot, \tilde{\lambda})\big) < \rho_{\tilde{\lambda}}[\boldsymbol{\mathcal{R}}]\big(\bar{h}_{\xi'}(\cdot, \tilde{\lambda})\big)$ for some $\tilde{\lambda} \in \Lambda$. Then $\bar{h}_\xi$ is C-Pareto optimal if there exists no $\bar{h}_{\xi'}$ such that $\bar{h}_{\xi'} \triangleright \bar{h}_\xi$.

We can again verify that a function $h^* : (x, \lambda) \mapsto h^*_\lambda(x)$ that maps onto a Bayes optimal model for each λ is in fact C-Pareto optimal. The following result shows that, under the assumption of existence of a Bayes optimal model, C-Pareto optimality is in fact equivalent to Bayes optimality.

Proposition 3.6. *Suppose there exists $h^* \in \mathcal{H}_\Lambda$ such that $h^*(\cdot, \lambda)$ is Bayes optimal for all $\lambda \in \Lambda$. Then an augmented hypothesis $g^* \in \mathcal{H}_\Lambda$ is C-Pareto optimal if and only if $g^*(\cdot, \lambda)$ is a Bayes optimal model for all $\lambda \in \Lambda$.*

Proposition 3.6 illustrates that all C-Pareto optimal augmented hypotheses can simultaneously learn all the Bayes optimal models. While this provides a strong guarantee, finding a C-Pareto optimal solution may still in practice be challenging and, when possible, one will prefer optimising against a scalar objective.

Let $\Delta(\Lambda)$ be the space of probability density functions over Λ. In our imprecise learning setting, a learner can scalarise the objective by choosing a distribution $Q \in \Delta(\Lambda)$, and taking an expectation over all objectives. This substitutes the learning problem over all of Λ with the scalarised objective

$$J_Q(\bar{h}_\xi) = \mathbb{E}_{\lambda \sim Q}\left[\rho_\lambda[\boldsymbol{\mathcal{R}}](\bar{h}_\xi(\cdot, \lambda))\right], \tag{5}$$

where the choice of distribution Q corresponds to a choice of scalarisation from the learner. The following proposition shows that all choices of Q lead to Bayes optimal models on their support.

Proposition 3.7. *Let $Q \in \Delta(\Lambda)$. If $g^* \in \mathcal{H}_\Lambda$ solves the scalarised optimisation problem, i.e., $g^* \in \arg\min_{g \in \mathcal{H}_\Lambda} J_Q(g)$, then $g^*(\cdot, \lambda)$ is a Bayes optimal model for all $\lambda \in \Lambda$ such that $Q(\lambda) > 0$.*

A similar result has previously been shown in Dosovitskiy and Djolonga (2020, Proposition 1) under the continuity and infinite model capacity assumptions. This result implies in particular that for any choice of distribution Q with full support, the scalarised objective can in theory yields a Bayes optimal model for every $\lambda \in \Lambda$.

4. Imprecise Risk Optimisation

Unfortunately, Proposition 3.7 does not inform specific choices of Q for the learner, leaving them in a state of ignorance. Under this scenario, the most popular narrative in the literature is to leave the choice of Q to the operators or to adopt non-informative priors such as Jeffreys prior and uniform priors (Brault et al., 2019; Dosovitskiy and Djolonga, 2020). However, both approaches would defeat the purpose of this work as they render the learning problem precise again (see the discussions in Section 2). In particular, it has been argued that complete or partial ignorance cannot be fully represented by a single precise probability (Walley, 1991, Sec. 5.5). For example, uniform distribution is not an appropriate way of representing ignorance because it coincides with a precise judgement of uniform belief.

C-Pareto improvement. To overcome this challenge, we adopt the concept of *C-Pareto improvement* which allows us to develope a learning algorithm that respects not only the limitation of evidence and resource, but also the complete ignorance of the learner. Specifically, we focus on the

[2] While we focus primarily on the learning aspect and assume throughout that the operator knows how to specify λ, we acknowledge the challenge of eliciting operators' preferences at test time; see Appendix F.2 for further discussion on test-time elicitation.

gradient-based method:

$$\xi_t \leftarrow \xi_{t-1} - \eta \cdot \nabla_\xi J_{Q_t}(\bar{h}_\xi), \quad Q_t \in \Delta(\Lambda). \tag{6}$$

We say that the update (6) makes a C-Pareto improvement if $\bar{h}_{\xi_t}$ dominates $\bar{h}_{\xi_{t-1}}$ according to the aggregation $\rho_\lambda[\boldsymbol{\mathcal{R}}]$. The central concept involves the adaptive selection of Q_t at each step, ensuring that the parameter update remains consistently non-dominant. This approach bears resemblance to the multiple-gradient descent algorithm (MGDA) utilised in multi-objective optimisation (Désidéri, 2012). The subsequent result demonstrates the specific selection of Q_t that results in C-Pareto improvement.

Theorem 4.1. *For $\lambda \in \Lambda$, suppose $\xi \mapsto \rho_\lambda[\boldsymbol{\mathcal{R}}](\bar{h}_\xi(\cdot, \lambda))$ is locally continuously differentiable in a neighbourhood of ξ. Define*

$$Q_t^* \in \operatorname*{arg\,min}_{Q \in \Delta(\Lambda)} \left\| \nabla_{\xi_{t-1}} J_Q(\xi_{t-1}) \right\|_2 \tag{7}$$

and $v_t(\xi_t) = \nabla_{\xi_t} J_{Q_t^*}(\xi_t)$. Then the update $\xi_t \leftarrow \xi_{t-1} - \eta \cdot v_t(\xi_t)$ for an appropriate choice of $\eta > 0$ always makes C-Pareto improvement.

4.1. Practical Algorithm with Theoretical Justification

In practice, we have access to data from d distinct domains. The empirical risk for the augmented hypothesis $\bar{h}_\xi \in \mathcal{H}_\Lambda$ for the i^{th} domain can be computed for each $\lambda \in \Lambda$ as

$$\widehat{\mathcal{R}}_i(\bar{h}_\xi(\cdot, \lambda)) = \frac{1}{n} \sum_{j=1}^{n} \ell(\bar{h}_\xi(x_j^{(i)}, \lambda), y_j^{(i)}), \tag{8}$$

where $(x_j^{(i)}, y_j^{(i)}) \sim \mathbb{P}_i$. In principle, the choice of λ determines how to aggregate the risk profile. However, in practice once λ is known, only then the corresponding $\bar{h}_\xi(\cdot, \lambda) \in \mathcal{H}_\Lambda$ can be used to compute the empirical risk profile. For a particular objective $\rho_\lambda[\widehat{\boldsymbol{\mathcal{R}}}](\bar{h}_\xi(\cdot, \lambda))$, we can compute the corresponding empirical risk profile as $\widehat{\boldsymbol{\mathcal{R}}}(\bar{h}_\xi(\cdot, \lambda)) = \{\widehat{\mathcal{R}}_i(\bar{h}_\xi(\cdot, \lambda)), \ldots, \widehat{\mathcal{R}}_d(\bar{h}_\xi(\cdot, \lambda))\}$. If Q is known to the learner, they can sample $\{\lambda_j\}_{j=1}^m \sim Q$ and compute the corresponding empirical risk profiles for each λ_j. However, for an imprecise learner, the right choice of distribution Q is unknown a priori. Therefore, we defer the computation of the empirical risk profile until the corresponding λ is known. That is, given a candidate distribution $Q \in \Delta(\Lambda)$ we compute the risk profile and aggregate it with $\{\lambda_j\}_{j=1}^m \sim Q$. We can then estimate Q_t^* with $\widehat{Q}_t$ using Monte Carlo estimate of (7), i.e.,

$$\widehat{Q}_t = \operatorname*{arg\,min}_{Q \in \Delta(\Lambda)} \left\| \frac{1}{m} \sum_{j=1}^{m} \nabla \rho_{\lambda_j}[\widehat{\boldsymbol{\mathcal{R}}}](\bar{h}_\xi(\cdot, \lambda_j)) \right\|_2, \tag{9}$$

where $\{\lambda_j\}_{j=1}^m \sim Q$. The direction of C-Pareto improvement is obtained by $\hat{v}_t(\xi_t) = \nabla_{\xi_t} J_{\widehat{Q}_t}(\xi_t)$. Algorithm 1 summarises the proposed algorithm.

Algorithm 1 Imprecise Risk Optimisation (IRO)

Input: Data from d distinct domains $\{x_i^{(d)}, y_i^{(d)}\}_{i=1}^n \sim \mathbb{P}_d(X, Y)$, a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, a probability space $\Delta(\Lambda)$, a (augmented) hypothesis class $\mathcal{H}_\Lambda$, risk aggregator $\rho_\lambda : \mathcal{L}^d(\mathcal{H}) \to \mathcal{L}(\mathcal{H})$, number of Monte Carlo samples m.
Initialise the parameter $\xi \in \Xi$.
repeat
  Estimate Q_t^* with $\widehat{Q}_t$ by solving (9) by computing $\widehat{Q}_t = \operatorname*{arg\,min}_{Q \in \Delta(\Lambda)} \| \frac{1}{m} \sum_{j=1}^m \nabla \rho_{\lambda_j}[\widehat{\boldsymbol{\mathcal{R}}}](\bar{h}_\xi(\cdot, \lambda_j)) \|_2$ where $\lambda_1, \ldots, \lambda_m \sim Q$.
  Compute $\hat{v}_t(\xi) = \frac{1}{m'} \sum_{k=1}^{m'} \nabla \rho_{\lambda_k}[\widehat{\boldsymbol{\mathcal{R}}}](\bar{h}_\xi(\cdot, \lambda_k))$ where $\lambda_1, \ldots, \lambda_{m'} \sim \widehat{Q}_t$.
  Update $\xi = \xi - \eta \hat{v}_t(\xi)$.
until $\|\hat{v}_t(\xi)\|_2 > \epsilon$

Proposition 4.2. *Let $Q \in \Delta(\Lambda)$ and let $\lambda_{\text{op}} \in \Lambda$ such that $Q(\lambda_{\text{op}}) > 0$. Assume that ρ_λ is a linear, idempotent aggregation operator and that the loss ℓ is upper bounded by $M \geq 0$. Let $n \geq 1$ be the number of samples we observe from each environment, assumed equal across environments. Then, there exists $q \in (0, 1)$ such that if*

$$\hat{g} \in \operatorname*{arg\,min}_{\bar{g} \in \mathcal{H}_\Lambda} \frac{1}{m} \sum_{i=1}^{m} \rho_{\lambda_i}[\widehat{\boldsymbol{\mathcal{R}}}](\bar{g}(\cdot, \lambda_i)) \tag{10}$$

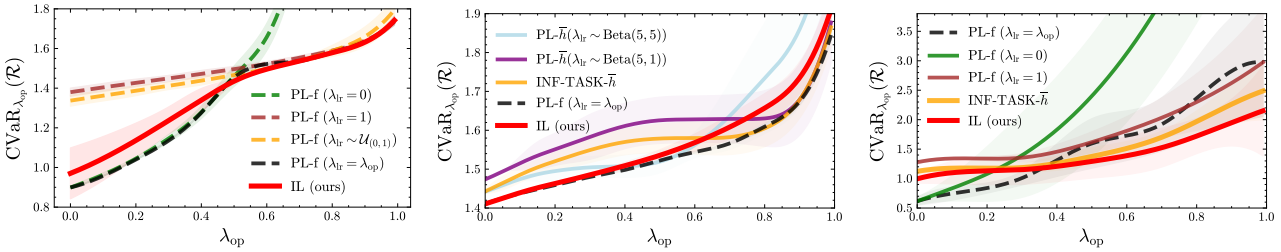
where $\lambda_1, \ldots, \lambda_m \sim Q$, then for any $\delta > q^m$, the following inequality holds with probability $1 - \delta$:

$$\begin{aligned} &\left| \rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}](\hat{g}(\cdot, \lambda_{\text{op}})) - \rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}](h^*(\cdot, \lambda_{\text{op}})) \right| \\ &\qquad \leq 2M \left(\sqrt{\frac{\log(6/\eta_\delta)}{2n}} + \sqrt{\frac{\log(6/\eta_\delta)}{2m(1-q)(1-q^m)}} \right), \end{aligned} \tag{11}$$

where $\eta_\delta = (\delta - q^m)/(1 - q^m)$.

This proposition shows that even when the learner does not know the operator's true preference λ_{op}, the operator excess risk on the solution of the empirical scalarised IRO problem $\hat{g}$ is bounded with high probability in $O(n^{-1/2} + m^{-1/2})$, provided Q has full support. This means in particular that, provided an unlimited budget on the number of samples (the λ_is) that can be drawn from Q, the operator excess risk has a bound that matches standard learning rates for ERM.

The constant $q \in (0, 1)$ depends on the choice of distribution Q and the operator's true preference λ_{op}. If Q has a high density around λ_{op}, then q can be chosen closer to zero. Conversely, if Q has a lower density around λ_{op}, the values of q will be closer to one, requiring a larger number of samples $\lambda_1, \ldots, \lambda_m$ to achieve a comparable bound.



(a) (Synthetic data) Comparing **IL** with **PL-**f trained using different λ_{lr}.

(b) (Synthetic data) Comparing **IL** with **PL-**$\bar{h}$ trained using different priors over λ_{lr}.

(c) (UCI Bike Rentals) Comparing **IL** with various **PL-**$\bar{h}$ and **PL-**f.

Figure 2: Experiments comparing imprecise learning (**IL**) with various precise learners with precise hypothesis (**PL-**f) and with augmented hypothesis (**PL-**$\bar{h}$). 1 standard deviation is included and experiments are repeated 5 times.

5. Experiments

Our framework features a learner, who trains the model, and an operator, who employs it, with their preferences denoted as λ_{lr} and λ_{op}. Due to the institutional separation, the operator's preferred generalisation strategy cannot be communicated to the learner. We assess our Imprecise Learning (IL) framework, allowing learners to train an augmented hypothesis $\bar{h}$ using our IRO algorithm (see Algorithm 1), enabling operators to provide λ_{op} at deployment. This contrasts with Precise Learners (PL-f) who commit to a fixed generalisation (λ_{lr}) during training, producing a precise hypothesis $f : \mathcal{X} \to \mathcal{Y}$, and PL-$\bar{h}$, who create an augmented hypothesis but with a pre-determined prior over λ_{lr}.

We evaluate using the objective $\rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}]$, comparing **IL**, **PL-**f with fixed (0 or 1) or uniform λ_{lr}, and **PL-**$\bar{h}$ with prior of λ_{lr} as Beta distributions (5,5), (5,1), and (1,1). The strategy aligning with Beta$(1, 1)$ corresponds to the approach in Brault et al. (2019), thus is termed **INF-TASK-**$\bar{h}$. We benchmark against an ideal scenario where λ_{lr} equals λ_{op} and also calculate the maximum regret, i.e., for any model $\bar{h}$ (or f), max-regret$(\bar{h}) \triangleq \sup_{\lambda_{\text{op}} \in \Lambda} \rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}](\bar{h}) - \rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}](\bar{h}^*_{\lambda_{\text{op}}})$, to gauge the models' deviation from optimality across all λ_{op}.

Synthetic data: Following Eastwood et al. (2022b), we construct a simulated experiment to compare learners. We consider a linear model for each domain d: $Y_d = \theta_d X + \epsilon$ with $X \sim \mathcal{N}(1, 0.5)$ and $\epsilon \sim \mathcal{N}(0, 0.1)$. We simulate different domains by drawing θ_d with probability $p = 0.5$ from Uniform distributions $\mathcal{U}_{(1,1.1)}$ and $\mathcal{U}_{(-1.1,-1)}$. This allows data to exhibit multi-modality, thus creating a discontinuous risk profile which becomes harder for a single augmented hypothesis to capture. We consider 250 train and 250 test domains with 100 samples from each domain.

CMNIST dataset. We also experiment on the CMNIST dataset (Arjovski, 2021), which is a modified version of the MNIST dataset. The task is to classify digits $\{0, 1, 2, 3, 4\}$ and $\{5, 6, 7, 8, 9\}$ into negative and positive classes, respectively. A color is introduced as an additional domain-specific predictive feature that varies across domains, e.g., $\mathbb{P}(Y = 1 \mid \text{color} = \text{red}) = 0.9$ for domains in which the true label is highly correlated with the color feature. As a result, the mechanism by which color influences the label changes across domains, but the shape has a stable mechanism across domains (see Figure 4a). We sample 10 training environments from a long-tailed Beta$(0.9, 1)$ distribution, resulting in over-represented (majority) and under-represented (minority) subgroups (see Figure 4b). Note that we do not make the IID assumption over environments since we evaluate all subgroups at test time. We further discuss the dataset and experiment setup in Appendix E.

Table 1: Reporting the maximum regret averaged over 5 repetitions for each experiment with one standard error included. **Top**: Comparing **IL** with **PL-**f (Synthetic). **Middle**: Comparing **IL** with **PL-**$\bar{h}$ (Synthetic). **Bottom**: Comparing **IL** with **PL-**f and **PL-**$\bar{h}$ (Bike Rentals).

PL-f ($\mathcal{U}(0,1)$)	**PL-**f ($\lambda_{lr}=0$)	**PL-**f ($\lambda_{lr}=1$)	**IL** (ours)
1.971 $\pm$ (0.0098)	6.177 $\pm$ (0.0617)	2.010 $\pm$ (0.0564)	**0.867** $\pm$ (0.0058)
PL-$\bar{h}$ (Beta(5,5))	**PL-**$\bar{h}$ (Beta(5,1))	INF-TASK-$\bar{h}$	**IL** (ours)
1.79 $\pm$ (0.12)	1.57 $\pm$ (0.03)	0.935 $\pm$ (0.04)	**0.56** $\pm$ (0.00)
PL-$f(\lambda_{lr}=0)$	**PL-**$f(\lambda_{lr}=1)$	INF-TASK-$\bar{h}$	**IL** (ours)
4.81$\pm$ 0.27	0.66 $\pm$ 0.01	0.72 $\pm$ 0.13	**0.42** $\pm$ 0.08

Real-world data: Following Rothenhäusler et al. (2021) and Subbaswamy et al. (2019), we use the UCI Bike Sharing dataset (Fanaee-T and Gama, 2014) to predict the number of hourly bike rentals R from various weather-related features. Here, R is transformed from count to continuous with normalization. The data contains $17,379$ observations with temporal information such as season and year. The data is partitioned by season (1-4) and year (1-2) to create 8 different domains. Domains from the first year are used for training and the subsequent year as test domains.

5.1. Insights from Experiments

Comparing IL with PL-f**.** Our initial experiment on synthetic data contrasts Imprecise Learning (**IL**) with Precise Learners (**PL-**f) across different λ_{lr} settings, including

average-case ($\lambda_{\text{lr}} = 0$) and worst-case ($\lambda_{\text{lr}} = 1$) scenarios. Results, shown in Figure 2a, indicate that **PL-f** models achieve the lowest aggregated risk compared to other learners when $\lambda_{\text{lr}} = \lambda_{\text{op}}$. However, when $\lambda_{\text{lr}} \neq \lambda_{\text{op}}$, **PL-$f$** then deviates from this ideal scenario, which is expected since **PL-f** models are finely tuned to their specific λ_{lr}. Conversely, **IL** achieves aggregated risks that remain close to the ideal scenario across the spectrum of λ_{op}, matching or exceeding the worst-case PL-f in risk-averse settings ($\lambda_{\text{op}} > 0.6$) and surpassing both average-case and worst-case **PL-f** as λ_{op} increases. Notably, **IL** achieved the lowest maximum regret (see Table 1), underscoring the advantage of imprecise learning in handling generalization uncertainty.

Comparing IL with PL-$\bar{h}$. In our second experiment using synthetic data, we evaluate **IL**'s augmented hypothesis trained using imprecise risk optimisation (IRO) against precise learners (PL-$\bar{h}$) employing various optimization strategies influenced by their subjective beliefs about λ_{op}. Results in Figure 2b indicate **IL**'s performance is close to the ideal baseline across most λ_{op} values, except at higher risk levels where INF-TASK and **PL-$\bar{h}$** trained under Beta(5,1) excel. This outcome aligns with expectations, as INF-TASK uniformly aggregates objectives, favoring higher-risk scenarios, similar to Beta(5,1)'s weighting towards higher λ. Despite this, **IL** outperforms these methods across other λ and achieves the lowest maximum regret (see Table 1), demonstrating the efficacy of the proposed method.

Comparing DG algorithms on CMNIST. In Table 2, we compare **IL** to other DG methods on three representative domains from minority and majority subgroups (see Figure 4). The domains $e \in \{0.0, 1.0\}$ demonstrate opposite mechanisms, i.e., in domain $e = 0.0$, the color red is fully predictive of the negative class, whereas for $e = 1.0$, it is fully predictive of the positive class. In domain $e = 0.5$, color is uncorrelated with the target. We can see that **IL** can learn relevant features in context with appropriate λ and generalises in all scenarios. By setting $\lambda = 0$, the model operator can be less risk averse and generalise better to domains from the majority subgroup, as noted in the performance of **IL** for $e = 0.0$. With $\lambda \to 1$, the model operator can be risk averse and generalise better to the minority subgroup and is also reflected in the performance of **IL** for $e \in \{0.5, 1.0\}$. Furthermore, with $\lambda \to 1$, it performs similarly to the invariant learners. We discuss the results on all test domains in Table 3 in Appendix E.

Real-world experiment. Figure 2c demonstrates similar comparisons between **IL** and various **PL-f** and **PL-$\bar{h}$** models as in previous experiments. Notably, **IL** surpassed the ideal scenario at higher risk levels. This can be attributed to the fact that CVAR as an objective discards data from lower-risk environments (see Section 3.2), thus the optimisation has lower statistical efficiency as risk level increases.

Table 2: Accuracy and maximal regret of different domain generalisation algorithms on the CMNIST test environments from $\mathbb{P}(Y = 1 \mid \text{color} = \text{red}) = e$ with $e \in \{0.0, 0.5, 1.0\}$, respectively. The hypothetical best invariant and Bayes classifier are listed in **bold**. Domain-wise best acc & regret are highlighted in **green**. Bayes classifier is defined w.r.t. the IID learner trained for a particular environment

Objective	Algorithm	$e = 0.0$	$e = 0.5$	$e = 1.0$	Regret
Average	ERM	**96.1**	59.2	28.3	72.7
Worst	GrpDRO	54.1	64.5	**75.5**	46.9
	SD	52.1	63.7	73.3	47.9
Invariance	IGA	71.8	65.2	50.3	49.7
	IRM	72.0	69.7	67.7	32.3
	VREx	72.7	69.5	68.5	31.5
	EQRM	67.8	69.1	72.1	32.2
	Oracle		**73.5**		**27.9**
PL-$\bar{h}$	Inf-Task	96.0	63.1	68.3	31.7
IL (Ours)	IRO	95.8	69.5	70.3	**29.7**
Bayes	ERM (IID)	**100.0**	**75.0**	**100.0**	

Augmented hypothesis mitigates this downside because it is smooth in the λ parameter by design, thus can "borrow" information from nearby risk regions. At last, **IL** consistently achieved the lowest maximum regret as shown in Table 1.

6. Conclusion

In out-of-distribution (OOD) generalisation, a clear institutional separation between machine learners and model operators creates generalisation uncertainty that prevents consensus on a specific generalisation approach during training. To overcome this, we presented imprecise domain generalisation. Our approach incorporates imprecise risk optimisation, allowing learners to maintain imprecision during training, coupled with a model framework that lets operators specify their generalisation strategy at deployment. Both theoretical analysis and experimental evaluations demonstrate the effectiveness of our proposed framework.

Our work faces two main limitations. First, it assumes that model operators are aware of their level of risk aversion. In practice, they may however struggle to precisely articulate their preferences. Consequently, this necessitates preference elicitation at test time, which may result in a probability distribution over λ rather than a single value. Second, imprecise learning is more computationally intensive compared to precise counterparts as it involves optimising for a continuum of objectives. In our future work, we aim to broaden the scope of imprecise learning by implementing methods to elicit user preferences more effectively, improving computational efficiency, and exploring alternative aggregation functions. This approach would empower operators to weigh various criteria such as fairness, privacy, and algorithmic performance effectively.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgement

We thank Kevin Murphy, Chris Holmes, Uri Shalit, Emtiyaz Khan, Hugo Monzón, Shai Ben-David, and Amartya Sanyal for fruitful discussion, and anonymous reviewers for their insightful feedback. We are indebted to Simon Föll for his contribution in conducting an initial set of experiments.

References

J. Abellan and A. R. Masegosa. An ensemble method using credal decision trees. *European journal of operational research*, 205(1):218–226, 2010.

M. Arjovski. *Out of Distribution Generalization in Machine Learning*. PhD thesis, New York University, 2021.

M. Arjovsky. *Out of Distribution Generalization in Machine Learning*. PhD thesis, Courant Institute of Mathematical Sciences, New York University, 2019.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *ArXiv*, abs/1907.02893, 2019.

P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, and L. M. Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1—-12, New York, NY, USA, 2020. Association for Computing Machinery.

S. Beery, G. Van Horn, and P. Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

S. Beery, E. Cole, and A. Gjoka. The iWildCam 2020 competition dataset. *CoRR*, 2020. URL https://arxiv.org/abs/2004.10340.

A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski. *Robust Optimization*, volume 28. Princeton University Press, 2009.

A. Ben-Tal, D. Bertsimas, and D. B. Brown. A soft robust model for optimization under ambiguity. *Operations Research*, 58:1220–1234, 2010.

G. Blanchard, G. Lee, and C. Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2178–2186. 2011.

G. Blanchard, A. A. Deshmukh, U. Dogan, G. Lee, and C. Scott. Domain generalization by marginal transfer learning. *Journal of Machine Learning Research*, 22(2): 1–55, 2021.

R. Brault, A. Lambert, Z. Szabo, M. Sangnier, and F. d'Alche Buc. Infinite task learning in rkhss. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pages 1294–1302. PMLR, 2019.

M. Caprio, S. Dutta, K. J. Jang, V. Lin, R. Ivanov, O. Sokolsky, and I. Lee. Imprecise Bayesian Neural Networks, May 2023. URL http://arxiv.org/abs/2302.09656. arXiv:2302.09656 [cs, stat].

M. Caprio, M. Sultana, E. Elia, and F. Cuzzolin. Credal learning theory, 2024.

S. L. Chau, S. Bouabid, and D. Sejdinovic. Deconditional Downscaling with Gaussian Processes. In *Advances in Neural Information Processing Systems*, volume 34, pages 17813–17825. Curran Associates, Inc., 2021a.

S. L. Chau, J.-F. Ton, J. González, Y. Teh, and D. Sejdinovic. BayesIMP: Uncertainty Quantification for Causal Data Fusion. In *Advances in Neural Information Processing Systems*, volume 34, pages 3466–3477. Curran Associates, Inc., 2021b.

S. L. Chau, M. Cucuringu, and D. Sejdinovic. Spectral ranking with covariates. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 70–86. Springer, 2022a.

S. L. Chau, J. Gonzalez, and D. Sejdinovic. Learning Inconsistent Preferences with Gaussian Processes. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 2266–2281. PMLR, May 2022b. ISSN: 2640-3498.

L. Chen, Y. Zhang, Y. Song, Y. Shan, and L. Liu. Improved test-time adaptation for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24172–24182, June 2023a.

Y. Chen, K. Zhou, Y. Bian, B. Xie, B. Wu, Y. Zhang, M. KAILI, H. Yang, P. Zhao, B. Han, and J. Cheng. Pareto invariant risk minimization: Towards mitigating the optimization dilemma in out-of-distribution generalization. In *The Eleventh International Conference on Learning Representations*, 2023b.

T. Cohen and M. Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016.

F. G. Cozman. Credal networks. *Artificial intelligence*, 120 (2):199–233, 2000.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2002.

J.-A. Désidéri. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5):313–318, 2012.

A. Dosovitskiy and J. Djolonga. You only train once: Loss-conditional training of deep networks. In *International Conference on Learning Representations*, 2020.

C. Eastwood, A. Robey, S. Singh, J. von Kügelgen, H. Hassani, G. J. Pappas, and B. Schölkopf. Probable domain generalization via quantile risk minimization. In *Advances in Neural Information Processing Systems*, volume 35, pages 17340–17358. Curran Associates, Inc., 2022a.

C. Eastwood, A. Robey, S. Singh, J. von Kügelgen, H. Hassani, G. J. Pappas, and B. Schölkopf. Probable Domain Generalization via Quantile Risk Minimization. In *Adv. Neural Inf. Process. Syst.*, volume 35. Curran Associates, Inc., Oct. 2022b.

C. Eastwood, S. Singh, A. L. Nicolicioiu, M. Vlastelica, J. von Kügelgen, and B. Schölkopf. Spuriosity didn't kill the classifier: Using invariant predictions to harness spurious features, 2023.

H. Fanaee-T and J. Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2:113–127, 2014.

S. Föll, A. Dubatovka, E. Ernst, S. L. Chau, M. Maritsch, P. Okanovic, G. Thäter, J. M. Buhmann, F. Wortmann, and K. Muandet. Gated Domain Units for Multi-source Domain Generalization, May 2023. URL http://arxiv.org/abs/2206.12444. arXiv:2206.12444 [cs].

J.-C. Gagnon-Audet, K. Ahuja, M. J. D. Bayazi, G. Dumas, and I. Rish. Woods: Benchmarks for out-of-distribution generalization in time series tasks. *ArXiv*, abs/2203.09978, 2022.

M. Ghifary, W. Kleijn, M. Zhang, and D. Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2551–2559, Los Alamitos, CA, USA, 2015. IEEE Computer Society.

F. J. Giron and S. Rios. Quasi-Bayesian Behaviour: A more realistic approach to decision making? *Trabajos de Estadistica Y de Investigacion Operativa*, 31 (1):17–38, Feb. 1980. ISSN 0041-0241. doi: 10.1007/BF02888345. URL http://link.springer.com/10.1007/BF02888345.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

I. J. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org.

E. L. Grab and I. R. Savage. Tables of the expected value of 1/x for positive bernoulli and poisson variables. *Journal of the American Statistical Association*, 49(265):169–177, 1954.

D. Ha, A. M. Dai, and Q. V. Le. Hypernetworks. In *International Conference on Learning Representations*, 2016.

M. Hardt, N. Megiddo, C. Papadimitriou, and M. Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 111–122, New York, NY, USA, 2016. Association for Computing Machinery.

H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

X. He, K. Zhao, and X. Chu. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622, 2021.

J. J. Heckman. Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture. *Journal of Political Economy*, 109(4):673–748, 2001.

C. Heinze-Deml and N. Meinshausen. Conditional variance penalties and domain shift robustness. *Machine Learning*, 110(2):303–348, 2021.

C. Heinze-Deml, J. Peters, and N. Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018.

T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008.

P. Kamath, A. Tangella, D. Sutherland, and N. Srebro. Does invariant risk minimization capture invariance? In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 4069–4077. PMLR, 2021.

P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. Earnshaw, I. Haque, S. M. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 18–24 Jul 2021.

M. Koyama and S. Yamaguchi. Out-of-distribution generalization with maximal invariant predictor. 2020.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. L. Priol, and A. Courville. Out-of-Distribution Generalization via Risk Extrapolation (REx), Feb. 2021. URL http://arxiv.org/abs/2003.00688. arXiv:2003.00688 [cs, stat].

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

T. Li, A. Beirami, M. Sanjabi, and V. Smith. On tilted losses in machine learning: Theory and applications. *Journal of Machine Learning Research*, 24(142):1–79, 2023.

X. Lin, H.-L. Zhen, Z. Li, Q.-F. Zhang, and S. Kwong. Pareto multi-task learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

J. Liu, Z. Hu, P. Cui, B. Li, and Z. Shen. Heterogeneous risk minimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 6804–6814. PMLR, 2021a.

Y. Liu, P. Kothari, B. van Delft, B. Bellot-Gurlet, T. Mordan, and A. Alahi. TTT++: When does self-supervised test-time training fail or thrive? In *Advances in Neural Information Processing Systems*, volume 34, pages 21808–21820. Curran Associates, Inc., 2021b.

P. Ma, T. Du, and W. Matusik. Efficient continuous pareto exploration in multi-task learning. In *International Conference on Machine Learning*, pages 6522–6531. PMLR, 2020.

D. Mahajan, S. Tople, and A. Sharma. Domain generalization using causal matching. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 7313–7324. PMLR, 2021.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Multiple Source Adaptation and the Renyi Divergence, May 2012. URL http://arxiv.org/abs/1205.2628. arXiv:1205.2628 [cs, stat].

K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 10–18, 2013.

V. Y. Nastl and M. Hardt. Predictors from causal features do not generalize better to new domains, 2024.

V. Pareto. The new theories of economics. *Journal of Political Economy*, 5, 1897.

J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 7599–7609. PMLR, 2020.

E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

J. Peters, P. BÃŒhlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

M. Pezeshki, O. Kaba, Y. Bengio, A. C. Courville, D. Precup, and G. Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021.

J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

H. Rahimian and S. Mehrotra. Distributionally Robust Optimization: A Review. *Open Journal of Mathematical Optimization*, 3:1–85, July 2022. ISSN 2777-5860. doi: 10.5802/ojmo.15. URL http://arxiv.org/abs/1908.05659. arXiv:1908.05659 [cs, math, stat].

A. Ratner, D. Alistarh, G. Alonso, D. G. Andersen, P. Bailis, S. Bird, N. Carlini, B. Catanzaro, E. S. Chung, B. Dally, J. Dean, I. S. Dhillon, A. G. Dimakis, P. Dubey, C. Elkan, G. Fursin, G. R. Ganger, L. Getoor, P. B. Gibbons, G. A. Gibson, J. E. Gonzalez, J. Gottschlich, S. Han, K. M. Hazelwood, F. Huang, M. Jaggi, K. G. Jamieson, M. I. Jordan, G. Joshi, R. Khalaf, J. Knight, J. Konečný, T. Kraska, A. Kumar, A. Kyrillidis, J. Li, S. Madden, H. B. McMahan, E. Meijer, I. Mitliagkas, R. Monga, D. G. Murray, D. S. Papailiopoulos, G. Pekhimenko, T. Rekatsinas, A. Rostamizadeh, C. Ré, C. D. Sa, H. Sedghi,

S. Sen, V. Smith, A. Smola, D. Song, E. R. Sparks, I. Stoica, V. Sze, M. Udell, J. Vanschoren, S. Venkataraman, R. Vinayak, M. Weimer, A. G. Wilson, E. P. Xing, M. Zaharia, C. Zhang, and A. Talwalkar. SysML: The new frontier of machine learning systems. *CoRR*, abs/1904.03257, 2019. URL http://arxiv.org/abs/1904.03257.

A. Robey, L. Chamon, G. J. Pappas, and H. Hassani. Probabilistically robust learning: Balancing average and worst-case performance. In *International Conference on Machine Learning*, pages 18667–18686. PMLR, 2022.

R. T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. 2002.

M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

E. Rosenfeld, P. K. Ravikumar, and A. Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021.

D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):215–246, 2021.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

S. Santurkar, D. Tsipras, and A. Madry. {BREEDS}: Benchmarks for subpopulation shift. In *International Conference on Learning Representations*, 2021.

O. Sener and V. Koltun. Multi-task learning as multi-objective optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 525–536, Red Hook, NY, USA, 2018. Curran Associates Inc.

G. Shafer. Dempster-shafer theory. *Encyclopedia of artificial intelligence*, 1:330–331, 1992.

Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui. Towards out-of-distribution generalization: A survey. *CoRR*, abs/2108.13624, 2021.

A. Subbaswamy, P. Schulam, and S. Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127. PMLR, 2019.

Y. Sun, X. Wang, Z. Liu, J. Miller, A. Efros, and M. Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 9229–9248. PMLR, 2020.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

L. V. Utkin, A. V. Konstantinov, and K. A. Vishniakov. An Imprecise SHAP as a Tool for Explaining the Class Probability Distributions under Limited Training Data, June 2021. URL http://arxiv.org/abs/2106.09111. arXiv:2106.09111 [cs, stat].

V. Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991.

V. Vapnik. *Statistical Learning Theory*. Wiley India Pvt Ltd, 1998.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin, Heidelberg, 1995.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

K. Q. H. Vo, M. Aadil, S. L. Chau, and K. Muandet. Causal Strategic Learning with Competitive Selection, Sept. 2023. arXiv:2308.16262 [cs].

P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, 1991.

D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021a.

J. Wang, C. Lan, C. Liu, Y. Ouyang, and T. Qin. Generalizing to unseen domains: A survey on domain generalization. *CoRR*, abs/2103.03097, 2021b. URL https://arxiv.org/abs/2103.03097.

R. Williamson and A. Menon. Fairness risk measures. In K. Chaudhuri and R. Salakhutdinov, editors, *Proc. 36th Int. Conf. Mach. Learn.*, volume 97 of *Proceedings of Machine Learning Research*, pages 6786–6797. PMLR, June 2019.

G. Wilson and D. J. Cook. A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol.*, 11 (5), 2020.

Y. Yang, H. Zhang, D. Katabi, and M. Ghassemi. Change is hard: A closer look at subpopulation shift. In *International Conference on Machine Learning*, 2023.

M. Zhang, H. Marklund, N. Dhawan, A. Gupta, S. Levine, and C. Finn. Adaptive risk minimization: Learning to adapt to domain shift. In *Advances in Neural Information Processing Systems*, volume 34, pages 23664–23678. Curran Associates, Inc., 2021.

R. Zhang and D. Golovin. Random hypervolume scalarizations for provable multi-objective black box optimization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 11096–11105. PMLR, 2020.

S. Zhao, X. Yue, S. Zhang, B. Li, H. Zhao, B. Wu, R. Krishna, J. E. Gonzalez, A. L. Sangiovanni-Vincentelli, S. A. Seshia, and K. Keutzer. A review of single-source deep unsupervised visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2): 473–493, 2022.

K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, apr 2023.

A. Proofs

This section provides the detailed proofs of our main results presented in the paper.

A.1. Proof of Proposition 3.6

Proof.
($\Leftarrow$)

We can easily verify that if $g^*(\cdot, \lambda)$ is Bayes optimal for all $\lambda \in \Lambda$, then it is C-Pareto optimal.

($\Rightarrow$)

Suppose $g^* \in \mathcal{H}_\Lambda$ is C-Pareto optimal. We have

$$\begin{aligned} g^* \text{ C-Pareto optimal} &\Leftrightarrow \nexists h \in \mathcal{H}_\Lambda, h \triangleright g^* \\ &\Leftrightarrow \forall h \in \mathcal{H}_\Lambda, \neg(h \triangleright g^*) \\ &\Leftrightarrow \forall h \in \mathcal{H}_\Lambda, \left[\exists \lambda \in \Lambda, \rho_\lambda[\boldsymbol{\mathcal{R}}](h(\cdot,\lambda)) > \rho_\lambda[\boldsymbol{\mathcal{R}}](g^*(\cdot,\lambda))\right] \\ &\qquad \vee \left[\forall \tilde{\lambda} \in \Lambda, \rho_{\tilde{\lambda}}[\boldsymbol{\mathcal{R}}](h(\cdot,\tilde{\lambda})) \geq \rho_{\tilde{\lambda}}[\boldsymbol{\mathcal{R}}](g^*(\cdot,\tilde{\lambda}))\right]. \end{aligned}$$

This implies in particular that

$$\left[\exists \lambda \in \Lambda, \rho_\lambda[\boldsymbol{\mathcal{R}}](h^*(\cdot,\lambda)) > \rho_\lambda[\boldsymbol{\mathcal{R}}](g^*(\cdot,\lambda))\right] \vee \left[\forall \tilde{\lambda} \in \Lambda, \rho_{\tilde{\lambda}}[\boldsymbol{\mathcal{R}}](h^*(\cdot,\tilde{\lambda})) \geq \rho_{\tilde{\lambda}}[\boldsymbol{\mathcal{R}}](g^*(\cdot,\tilde{\lambda}))\right].$$

Since $h^*(\cdot, \lambda)$ is Bayes optimal for all $\lambda \in \Lambda$, the first statement cannot be true. Therefore, the second statement must hold and we have

$$\begin{aligned} g^* \text{ C-Pareto optimal} &\Rightarrow \forall \tilde{\lambda} \in \Lambda, \rho_{\tilde{\lambda}}[\boldsymbol{\mathcal{R}}](h^*(\cdot,\tilde{\lambda})) \geq \rho_{\tilde{\lambda}}[\boldsymbol{\mathcal{R}}](g^*(\cdot,\tilde{\lambda})) \\ &\Rightarrow \forall \tilde{\lambda} \in \Lambda, \rho_{\tilde{\lambda}}[\boldsymbol{\mathcal{R}}](h^*(\cdot,\tilde{\lambda})) = \rho_{\tilde{\lambda}}[\mathcal{R}](g^*(\cdot,\tilde{\lambda})) \qquad (h^*(\cdot,\lambda) \text{ Bayes optimal}) \\ &\Rightarrow \forall \tilde{\lambda} \in \Lambda, g^*(\cdot,\tilde{\lambda}) \text{ Bayes optimal} . \end{aligned}$$

This concludes the proof. □

A.2. Proof of Proposition 3.7

Proof. Since $h^*(\cdot, \lambda)$ is a Bayes optimal model for all $\lambda \in \Lambda$, we have

$$\begin{aligned} &\rho_\lambda[\boldsymbol{\mathcal{R}}](g^*(\cdot,\lambda)) - \rho_\lambda[\boldsymbol{\mathcal{R}}](h^*(\cdot,\lambda)) \geq 0, \ \ \forall \lambda \in \Lambda \\ \Rightarrow\ &\mathbb{E}_{\lambda\sim Q}[\rho_\lambda[\boldsymbol{\mathcal{R}}](g^*(\cdot,\lambda)) - \rho_\lambda[\boldsymbol{\mathcal{R}}](h^*(\cdot,\lambda))] \geq 0. \end{aligned}$$

But by definition of g^* we also have

$$J_Q(g^*) \leq J_Q(h^*) \Rightarrow \mathbb{E}_{\lambda\sim Q}[\rho_\lambda[\boldsymbol{\mathcal{R}}](g^*(\cdot,\lambda)) - \rho_\lambda[\boldsymbol{\mathcal{R}}](h^*(\cdot,\lambda))] \leq 0.$$

Therefore,

$$\int_\Lambda \left[\rho_\lambda[\boldsymbol{\mathcal{R}}](g^*(\cdot,\lambda)) - \rho_\lambda[\boldsymbol{\mathcal{R}}](h^*(\cdot,\lambda))\right] Q(\lambda)\, \mathrm{d}\lambda = 0.$$

Since the integrand is positive, it implies that for all $\lambda \in \Lambda$ such that $Q(\lambda) > 0$, $\rho_\lambda[\boldsymbol{\mathcal{R}}](g^*(\cdot,\lambda)) = \rho_\lambda[\boldsymbol{\mathcal{R}}](h^*(\cdot,\lambda))$ which concludes the proof. □

A.3. Proof of Theorem 4.1 and C-Pareto Improvement

When the choice of scalarisation, i.e., Q improves some objectives at the cost of degrading other objectives, it induces a preference. Therefore, the problem becomes multi-objective again as there will be a trade-off among these objectives. The imprecise choice of scalarization will be the distribution Q^* such that it improves at least one of the objectives without degrading any other objective, i.e., it ensures C-Pareto improvement. Formally,

Proposition A.1. *Suppose a learning algorithm $\mathfrak{P}$ learns an augmented hypothesis $g \in \mathcal{H}_\Lambda$ using aggregated objective $J_Q(g)$ (ref. Eq 5). Then, the learner does not induce an additional preference over $\mathcal{H}_\Lambda$ if it makes Pareto improvement.*

Proof. Consider a learner $\mathfrak{P}$ which learns an augmented hypothesis $g \in \mathcal{H}_\Lambda$ which aggregates the objectives $\rho_\lambda[\boldsymbol{\mathcal{R}}](g(\cdot,\lambda))$ for all $\lambda \in \Lambda$ with respect to Q to obtain the aggregated objective $\mathbb{E}_{\lambda\sim Q}[\rho_\lambda[\boldsymbol{\mathcal{R}}](g(\cdot,\lambda))]$. Assume that $\mathcal{G} = \{g_i\}_{i=0}^n$ denotes the sequence of models that the learner obtains at every update while minimizing the aggregated objective. We know that $\forall i, j$ such that $i < j$ and $g_i, g_j \in \mathcal{G}$

$$\mathbb{E}_{\lambda\sim Q}[\rho_\lambda[\boldsymbol{\mathcal{R}}](g_i(\cdot,\lambda))] > \mathbb{E}_{\lambda\sim Q}[\rho_\lambda[\boldsymbol{\mathcal{R}}](g_j(\cdot,\lambda))]$$

which defines a preference on $\mathcal{G} \subset \mathcal{H}_\Lambda$ with aggregated objective as the utility function $u_Q(g) := -\mathbb{E}_{\lambda\sim Q}[\rho_\lambda[\boldsymbol{\mathcal{R}}](g(\cdot,\lambda))]$. This additional preference relation agrees with the original binary preference relation ($\succeq$) on $\mathcal{H}_\Lambda$ which defines dominance and C-Pareto optimality if there does not exist $g_i, g_j \in \mathcal{G}$ such that $i < j$, $g_i \not\succeq g_j$ and $g_j \not\succeq g_i$ with $u_Q(g_i) > Q(g_j)$. This implies that aggregated objective u_Q must be such that for all $g_i, g_j \in \mathcal{G}$ and $i < j$, $g_j \succeq g_i$. That is, u_Q should make C-Pareto improvement to not induce any additional preference. □

Therefore, we propose an alternate characterization of C-Pareto optimality based on the concept of C-Pareto improvement with the idea of local gradients.

Proposition A.2. *An augmented hypothesis $\bar{h}_\xi$ is C-Pareto optimal if and only if there exists no $w \in \Xi$ such that for an $\epsilon > 0$, $\rho_\lambda[\boldsymbol{\mathcal{R}}](\bar{h}_{\xi-\epsilon w}(\cdot,\lambda)) \leq \rho_\lambda[\boldsymbol{\mathcal{R}}](\bar{h}_\xi(\cdot,\lambda))$ for all $\lambda \in \Lambda$ and $\rho_{\tilde{\lambda}}[\boldsymbol{\mathcal{R}}](\bar{h}_{\xi-\epsilon w}(\cdot,\lambda)) < \rho_{\tilde{\lambda}}[\boldsymbol{\mathcal{R}}](\bar{h}_\xi(\cdot,\lambda))$ for some $\tilde{\lambda} \in \Lambda$.*

Proof. We prove the forward direction using contradiction. Assume h is C-Pareto optimal and there exists $w \in \Xi$ such that for an $\epsilon > 0$, $\rho_\lambda[\mathcal{R}](h_{\xi-\epsilon w}(\cdot,\lambda)) \leq \rho_\lambda[\mathcal{R}](h_\xi(\cdot,\lambda))$ for all $\lambda \in \Lambda$ and $\rho_{\tilde{\lambda}}[\mathcal{R}](h_{\xi-\epsilon w}(\cdot,\lambda)) < \rho_{\tilde{\lambda}}[\mathcal{R}](h_\xi(\cdot,\lambda))$ for some $\tilde{\lambda} \in \Lambda$. Then $h_{\xi-\epsilon w}$ strictly dominates h_ξ according to our definition of C-Pareto optimality for augmented hypothesis Def **??** which contradicts that h_ξ is C-Pareto optimal. We prove the reverse direction using the contraposition. Assume h_ξ is not C-Pareto optimal. Then there exists $h_{\xi'}$ that strictly dominates h_ξ, i.e., $\rho_\lambda[\mathcal{R}](h_{\xi'}(\cdot,\lambda)) \leq \rho_\lambda[\mathcal{R}](h_\xi(\cdot,\lambda))$ for all $\lambda \in \Lambda$ and $\rho_{\tilde{\lambda}}[\mathcal{R}](h_{\xi'}(\cdot,\lambda)) < \rho_{\tilde{\lambda}}[\mathcal{R}](h_\xi(\cdot,\lambda))$ for some $\tilde{\lambda} \in \Lambda$. Then there exists $w = \xi - \xi'$ and $\epsilon = 1$ such that $h_{\xi-\epsilon w}$ strictly dominates h_ξ. □

Proposition A.2 shows that for C-Pareto optimality there must not be any direction $w \in \Xi$ for Pareto improvement. The non-existence of a direction for Pareto improvement is an if and only-if condition for C-Pareto optimality.

Proposition A.3. *In an ϵ-neighbourhood of ξ let $\rho_\xi[\boldsymbol{\mathcal{R}}](h_\xi(\cdot,\lambda))$ be a smooth function of ξ and the local gradient is defined as $v_\lambda(h_\xi) := \nabla_\xi \rho_\lambda[\mathcal{R}](h_\xi(\cdot,\lambda))$. If h_ξ is not pareto optimal then there exists a local pareto improvement direction $-w \in \Xi$ such that for all $\lambda \in \Lambda$ $w^\top v_\lambda(h_\xi) \geq 0$ and for some $\tilde{\lambda} \in \Lambda$ $w^\top v_{\tilde{\lambda}}(h_\xi) > 0$.*

Proof. From Proposition A.2, when h_ξ is not Pareto optimal, there exists a $w \in \Xi$ such that for an $\epsilon > 0$, $h_{\xi-\epsilon w} \succ h_\xi$. Then for all $\lambda \in \Lambda$,

$$\begin{aligned}
\rho_\lambda[\mathcal{R}](h_{\xi-\epsilon w}) &\leq \rho_\lambda[\mathcal{R}](h_\xi) \\
\rho_\lambda[\mathcal{R}](h_\xi) - \epsilon w^\top v_\lambda(h_\xi) + \epsilon^2 \mathbf{R} &\leq \rho_\lambda[\mathcal{R}](h_\xi) \\
-\epsilon w^\top v_\lambda(h_\xi) + \epsilon^2 \mathbf{R} &\leq 0 \qquad (\mathbf{R} : \text{Remainder Higher order terms}) \\
\epsilon \mathbf{R} &\leq w^\top v_\lambda(h_\xi)
\end{aligned}$$

Since $\rho_\lambda[\mathcal{R}](h_{\xi-\epsilon w}) \leq \rho_\lambda[\mathcal{R}](h_\xi)$, then $w^\top v_\lambda(h_\xi) \geq 0$ as $\epsilon \to 0$ otherwise a contradiction would arise for sufficiently small ϵ. Similarly for an $\tilde{\lambda} \in \Lambda$, since $\rho_{\tilde{\lambda}}[\mathcal{R}](h_{\xi-\epsilon w}) < \rho_{\tilde{\lambda}}[\mathcal{R}](h_\xi)$, then $w^\top v_{\tilde{\lambda}}(h_\xi) > 0$. □

Proposition A.3 extends the argument from Proposition A.2 that when h_ξ is not C-Pareto optimal, a direction for Pareto improvement must exist. Remark explains that a local Pareto improvement direction must align with the gradient of all objectives. Since the direction opposite to the local gradient of an objective shows us the direction of the improvement for the objective, then the direction opposite to local Pareto improvement $w \in \Xi$ must align with the local gradient if $-w \in \Xi$ improves the corresponding objective. Note that the local gradient of aggregated objective (5) is

$$\nabla_\xi J_Q(\xi) := \mathbb{E}_{\lambda\sim Q}[v_\lambda(h_\xi)] \tag{12}$$

Where $v_\lambda(h_\xi) := \nabla_\xi \rho_\lambda[\boldsymbol{\mathcal{R}}](h_\xi(\cdot,\lambda))$ denotes the local gradient of $\rho_\lambda[\boldsymbol{\mathcal{R}}](h_\xi(\cdot,\lambda))$. Then the choice of Q such that $\nabla_\xi J_Q(\xi)$ is the the direction of local Pareto improvement is given by

Proposition A.4. *For $\lambda \in \Lambda$, suppose $\xi \mapsto \rho_\lambda[\boldsymbol{\mathcal{R}}](\bar{h}_\xi(\cdot,\lambda))$ is locally continuously differentiable in a neighbourhood of ξ. Define*

$$Q_t^* \in \arg\min_{Q\in\Delta(\Lambda)} \left\|\nabla_{\xi_{t-1}} J_Q(\xi_{t-1})\right\|_2 \tag{13}$$

and $v_t(\xi_t) = \nabla_{\xi_t} J_{Q_t^*}(\xi_t)$. Then the update $\xi_t \leftarrow \xi_{t-1} - \eta \cdot v_t(\xi_t)$ for an appropriate choice of $\eta > 0$ always makes C-Pareto improvement. i.e., $-v_t(\xi_t)$ for all objectives $\rho_\lambda[\mathcal{R}](h_\beta(\cdot,\lambda))$, $\lambda \in \Lambda$ such that $v_t(\xi_t)^T v_\lambda(h_\xi) \geq ||v_t(\xi_t)||_2^2$.

Proof. We start by assuming that a given Q_t^* exists then the update $\xi_t \leftarrow \xi_{t-1} - \eta \cdot v_t(\xi_t)$ performs local C-Pareto improvement. First we show that $\forall\lambda \in \Lambda$ the $v_t(\xi_t)^T v_\lambda(h_\xi) \geq ||v_t(\xi_t)||_2^2$. For any distribution $Q \in \Delta(\Lambda)$, $v = \mathbb{E}_{\lambda\sim Q}[v_\lambda(h_\xi)] - v_t(\xi_t)$. We can say that $\forall\epsilon \in [0,1]$ $v_t(\xi_t) + \epsilon v$ is essentially

$$\begin{aligned} v_t(\xi_t) + \epsilon v &= v_t(\xi_t) + \epsilon(\mathbb{E}_{\lambda\sim Q}[v_\lambda(h_\xi)] - v_t(\xi_t)) \\ &= (1-\epsilon)v_t(\xi_t) + \epsilon\mathbb{E}_{\lambda\sim Q}[v_\lambda(h_\xi)] \\ &= \mathbb{E}_{\lambda\sim\epsilon Q+(1-\epsilon)Q_t^*}[v_\lambda(h_\xi)] \end{aligned}$$

Where $\epsilon Q + (1-\epsilon)Q_t^*$ is some other valid probability distribution. Therefore the norm of $v_t(\xi_t) + \epsilon v$ must be larger than or equal to the minimum norm obtained from Equation (13).

$$\begin{aligned} (v_t(\xi_t) + \epsilon v)^T (v_t(\xi_t) + \epsilon v) &\geq v_t(\xi_t)^T v_t(\xi_t) \\ 2\epsilon v_t(\xi_t)^T v + \epsilon^2 v^T v &\geq 0 \\ \epsilon &\geq \frac{-2v_t(\xi_t)^T v}{v^T v} \end{aligned}$$

Since the above statement must be true for all $\epsilon \in (0,1]$. For $\epsilon = 0$ equality must hold that $v_t(\xi_t)^T v_t(\xi_t) = v_t(\xi_t)^T v_t(\xi_t)$. Therefore, the lower bound from above must be less than or equal to 0.

$$\begin{aligned} \frac{-2v_t(\xi_t)^T v}{v^T v} &\leq 0 \\ v_t(\xi_t)^T v &\geq 0 \end{aligned}$$

Replacing v by $\mathbb{E}_{\lambda\sim Q}[v_\lambda(h_\xi)] - v_t(\xi_t)$ then gives us that

$$\begin{aligned} v_t(\xi_t)^T(\mathbb{E}_{\lambda\sim Q}[v_\lambda(h_\xi)] - v_t(\xi_t)) &\geq 0 \\ v_t(\xi_t)^T\mathbb{E}_{\lambda\sim Q}[v_\lambda(h_\xi)] &\geq v_t(\xi_t)^T v_t(\xi_t) \end{aligned}$$

Thus we obtain that $\forall\lambda \in \Lambda$ the $v_\lambda(h_\xi)^T v_t(\xi_t) \geq ||v_t(\xi_t)||_2^2$ by setting Q to be dirac delta function at λ. Therefore from Proposition A.3 we can say that $h_{\xi_{t-1}-\eta v_t(\xi_t)} \succ h_{\xi_{t-1}}$. This makes $w \in \Xi$ the common direction for local C-Pareto improvement. □

Analogous to the definition 3.3 we define C-Pareto stationarity for augmented hypothesis as

Definition A.5. Let $\rho_\lambda[\mathcal{R}](\bar{h}(\cdot,\lambda))$ be a smooth function of augmented hypothesis $\bar{h}$ and $v_\lambda(\bar{h}_\xi) := \nabla\rho_\lambda[\boldsymbol{\mathcal{R}}](\bar{h}_\xi(\cdot,\lambda))$ be the local gradient then the augmented hypothesis is said to be C-Pareto Stationary if and only if there exists a probability density q such that $\int v_\lambda(\bar{h}_\xi)\, dq(\lambda) = 0$.

Intuitively, C-Pareto Stationarity corresponds to local C-Pareto Optimality. For a single objective, C-Pareto stationarity is equivalent to the first-order derivative being zero. Therefore, If an augmented hypothesis h is C-Pareto optimal, it is C-Pareto stationary. This means that C-Pareto stationarity is a necessary condition for C-Pareto optimality. From Proposition A.2 we know that for a C-Pareto optimal point, no direction for Pareto improvement must exist, which implies that no direction for local Pareto improvement must also not exist. From theorem 4.1 we know that a local direction for pareto improvement is $v_t(\xi_t) = \int v_\lambda(h_\xi) dQ_t^*(\lambda)$ where $Q_t^* = arg\min_{Q\in\Delta(\Lambda)} ||\mathbb{E}_{\lambda\sim Q}[v_\lambda(h_\xi)]||$. Given that no direction for local C-Pareto improvement must exist implies that $v_t(\xi_t) = 0$. This means that there exists a distribution Q such that $\int v_\lambda(h_\xi) dQ(\lambda) = 0$. This illustrates that C-Pareto stationarity is a necessary condition for C-Pareto optimality which intuitively illustrates that local C-Pareto optimality is necessary for C-Pareto optimality.

A.4. Proof of Proposition 4.2

A.4.1. USEFUL RESULTS

Proposition A.6. *Let X be a random variable taking values in $\mathcal{X}$ and let $f : \mathcal{X} \to \mathbb{R}_+$ and $g : \mathcal{X} \to \mathbb{R}_+$ be non-negative functions. Define $Z = (f(X), g(X))$ and suppose that it admits a continuous density p_Z with respect to the Lebesgue measure on $\mathbb{R}^2$.*

Let $\alpha, \beta > 0$ such that $p_Z(\alpha, \beta) > 0$ and let $Z_1, \ldots, Z_n$ be independent copies of Z. Then there exists $r \geq 1$ and a random subsampling operator π such that $\pi([n]) \in 2^{\{1,\ldots,n\}}$, $|\pi([n])| \sim \text{Binomial}(n, 1/r)$, and for any index $i \in \pi([n])$

$$\mathbb{E}[Z_i] = (\alpha, \beta), \tag{14}$$

where the expectation is taken against both the variable and the index.

Proof. The proof consists in showing that the assumptions made are sufficient to construct a rejection sampling procedure where the proposal density is the density of Z and the target density is a uniform centered over (α, β).

Since $p_Z(\alpha, \beta) > 0$ and p_Z is continuous, there exists an open neighbourhood of (α, β) where p_Z is strictly positive. Therefore, there exists $\eta > 0$ such that if we define the closed rectangle

$$\begin{aligned} A_\alpha &= [\alpha - \eta/2, \alpha + \eta/2] \\ A_\beta &= [\beta - \eta/2, \beta + \eta/2] \\ A &= A_\alpha \times A_\beta, \end{aligned}$$

then $p_Z(x, x') > 0$ for any $(x, x') \in A$ and admits a positive lower bound on A. Further, we can define the uniform random variable $U \sim \text{Uniform}(A)$ with probability density

$$p_U(x, x') = \frac{1}{\eta^2}, \ \ \forall (x, x') \in A.$$

Then, by upper boundedness of p_U over A and lower-boundedness of p_Z over A, there exists $r \geq 1$ such that for any $(x, x') \in A$,

$$\frac{p_U(x, x')}{p_Z(x, x')} \leq r.$$

As a result, we can formally construct a rejection sampling procedure to sample from U using samples from Z with acceptance rate $1/r$. It is important to note this is only a formal construction to show the existence of an appropriate subsampling procedure. In practice, we may not be able to evaluate p_Z and therefore may be unable to effectively implement the procedure.

Algorithm 2 Algorithmic definition of the random subsampling operator π

Input: p_U, p_Z, r, $Z_1, \ldots, Z_n$
Initialise `subsampled` = $\{\}$
for $i \in \{1, \ldots, n\}$ **do**
  Let $U_i \sim \text{Uniform}([0, 1])$
  if $U_i \leq p_U(Z_i)/r p_Z(Z_i)$ **then**
    Append i to `subsampled`
  end if
end for
Return `subsampled`

Algorithm 2 outlines an algorithmic definition of a random subsampling operator $\pi : 2^{[n]} \to 2^{[n]}$ based on rejection sampling. We emphasise the random nature of the operator π as $Z_1, \ldots, Z_n$ are treated throughout as random variables. By property of rejection sampling, the number of accepted samples $|\pi([n])|$ or $|\texttt{subsampled}|$ follows a Binomial distribution with n trials and probability of success $1/r$. Finally, we have by construction that for any $i \in \pi([n])$

$$\mathbb{E}[Z_i] = \mathbb{E}[\text{Uniform}(A)] = (\alpha, \beta)$$

which concludes the proof. □

A.4.2. PROOF OF THE MAIN RESULT

We begin by introducing notations which will be used in this proof. Suppose we observe $n \in \mathbb{N}$ of IID observations from each environment, i.e., we observe $(x_1^{(i)}, y_1^{(i)}), \ldots, (x_n^{(i)}, y_n^{(i)}) \sim \mathbb{P}_i$ for every $i \in \{1, \ldots, d\}$. Furthermore, let $Q \in \Delta(\Lambda)$ be the scalarisation density the learner chooses and let $\lambda_1, \ldots, \lambda_m \sim Q$ be independent samples from this distribution.

For each environment $i \in \{1, \ldots, d\}$, we define an empirical risk

$$\hat{\mathcal{R}}_i(f) = \frac{1}{n} \sum_{j=1}^{n} \ell(y_j^{(i)}, f(x_j^{(i)})), \quad f \in \mathcal{H},$$

which we concatenate into an empirical risk profile $\hat{\boldsymbol{\mathcal{R}}} = (\hat{\mathcal{R}}_1, \ldots, \hat{\mathcal{R}}_d)$. We can easily verify that for any $i \in \{1, \ldots, d\}$, $\mathbb{E}[\hat{\mathcal{R}}_i] = \mathcal{R}_i$ where the expectation is taken against $\mathbb{P}_i$, thus $\mathbb{E}[\hat{\boldsymbol{\mathcal{R}}}] = \boldsymbol{\mathcal{R}}$. Therefore, if we take the empirical aggregated risk to be $\rho_\lambda[\hat{\boldsymbol{\mathcal{R}}}]$ for $\lambda \in \Lambda$ and assume that $\rho_\lambda : \mathcal{L}_2^d(\mathcal{H}) \to \mathcal{L}_2(\mathcal{H})$ is a linear risk aggregation function, it follows that

$$\mathbb{E}\left[\rho_\lambda[\hat{\boldsymbol{\mathcal{R}}}]\right] = \rho_\lambda\left[\mathbb{E}[\hat{\boldsymbol{\mathcal{R}}}]\right] = \rho_\lambda[\boldsymbol{\mathcal{R}}].$$

Finally, define the empirical scalarised risk using the values $\lambda_1, \ldots, \lambda_m$ sampled above, for $g \in \mathcal{H}_\Lambda$ as

$$\hat{J}_Q(g) = \frac{1}{m} \sum_{i=1}^{m} \rho_{\lambda_i}[\hat{\boldsymbol{\mathcal{R}}}](g(\cdot, \lambda_i)).$$

In what follows, we will always assume there exists a function $\hat{h} \in \mathcal{H}_\Lambda$ such that for any $\lambda \in \Lambda$, $\hat{h}(\cdot, \lambda)$ is a minimiser of the empirical aggregated risk $\rho_\lambda[\hat{\boldsymbol{\mathcal{R}}}]$, i.e.,

$$\hat{h}(\cdot, \lambda) \in \operatorname*{arg\,min}_{f \in \mathcal{H}} \rho_\lambda[\hat{\boldsymbol{\mathcal{R}}}](f), \quad \forall \lambda \in \Lambda,$$

and that the empirical scalarised risk also admits a minimiser which we denote $\hat{g} \in \mathcal{H}_\Lambda$, i.e.,

$$\hat{g} \in \operatorname*{arg\,min}_{g \in \mathcal{H}_\Lambda} \hat{J}_Q(g).$$

The following lemma shows that when such minimisers exists, then $\hat{g}(\cdot, \lambda_i)$ is automatically a minimiser of the empirical aggregated risk $\rho_{\lambda_i}[\hat{\boldsymbol{\mathcal{R}}}]$.

Lemma A.7. *Suppose there exists $\hat{h}, \hat{g}$ defined as above. Then $\hat{g}(\cdot, \lambda_i)$ minimises $\rho_{\lambda_i}[\hat{\boldsymbol{\mathcal{R}}}]$ for all $i \in \{1, \ldots, m\}$.*

Proof. Let $\hat{h} \in \mathcal{H}_\Lambda$ such that $\hat{h}(\cdot, \lambda) \in \operatorname*{arg\,min}_{f \in \mathcal{H}} \rho_\lambda[\hat{\boldsymbol{\mathcal{R}}}](f)$ for any $\lambda \in \Lambda$. Then, we have

$$\begin{aligned}
& \rho_\lambda[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot, \lambda)) \geq \rho_\lambda[\hat{\boldsymbol{\mathcal{R}}}](\hat{h}(\cdot, \lambda)), \quad \forall \lambda \in \Lambda \\
\Rightarrow\ & \rho_{\lambda_i}[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot, \lambda_i)) \geq \rho_{\lambda_i}[\hat{\boldsymbol{\mathcal{R}}}](\hat{h}(\cdot, \lambda_i)), \quad \forall i \in \{1, \ldots, m\} \\
\Rightarrow\ & \hat{J}_Q(\hat{g}) \geq \hat{J}_Q(\hat{h}) \\
\Rightarrow\ & \hat{J}_Q(\hat{g}) = \hat{J}_Q(\hat{h}) && (\hat{g} \in \arg\min \hat{J}_Q) \\
\Rightarrow\ & \frac{1}{m} \sum_{i=1}^{m} \rho_{\lambda_i}[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot, \lambda_i)) - \rho_{\lambda_i}[\hat{\boldsymbol{\mathcal{R}}}](\hat{h}(\cdot, \lambda_i)) = 0 \\
\Rightarrow\ & \rho_{\lambda_i}[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot, \lambda_i)) = \rho_{\lambda_i}[\hat{\boldsymbol{\mathcal{R}}}](\hat{h}(\cdot, \lambda_i)), \quad \forall i \in \{1, \ldots, m\} && (\text{sum of positives}) \\
\Rightarrow\ & \hat{g}(\cdot, \lambda_i) \in \operatorname*{arg\,min}_{f \in \mathcal{H}} \rho_{\lambda_i}[\hat{\boldsymbol{\mathcal{R}}}](f), \quad \forall i \in \{1, \ldots, m\}.
\end{aligned}$$

This concludes the proof. □

Finally, before we turn to the main result, recall that we assume there exists $h^* \in \mathcal{H}_\Lambda$ such that $h^*(\cdot, \lambda) \in \mathcal{H}$ is a Bayes optimal model for any $\lambda \in \Lambda$, i.e., $h^*(\cdot, \lambda) \in \arg\min_{f \in \mathcal{H}} \rho_\lambda[\boldsymbol{\mathcal{R}}](f)$. For any $\lambda \in \Lambda$, we denote the resulting Bayes risk as

$$\rho_\lambda[\boldsymbol{\mathcal{R}}]^\star = \rho_\lambda[\boldsymbol{\mathcal{R}}](h^*(\cdot, \lambda)).$$

Let $\lambda_{\text{op}} \in \Lambda$ be the choice of λ which reflects the operator's preference, but is unknown to the learner. The following result provides a bound on the excess risk at λ_{op} when using $\hat{g}$ as a hypothesis.

Proposition A.8. *Let $Q \in \Delta(\Lambda)$ and let $\lambda_{op} \in \Lambda$ such that $Q(\lambda_{op}) > 0$. Suppose that ρ_λ is a linear, idempotent aggregation operator and that the loss ℓ is upper bounded by $M \geq 0$. Then there exists $q \in (0, 1)$ such that for any $\delta > q^m$, the following inequality holds with probability $1 - \delta$:*

$$\left|\rho_{\lambda_{op}}[\boldsymbol{\mathcal{R}}](\hat{g}(\cdot, \lambda_{op})) - \rho_{\lambda_{op}}[\boldsymbol{\mathcal{R}}]^\star\right| \leq 2M \left(\sqrt{\frac{\log(2/\eta_\delta)}{2n}} + \sqrt{\frac{\log(2/\eta_\delta)}{2m(1-q)(1-q^m)}} \right),$$

where $\eta_\delta = (\delta - q^m)/(1 - q^m)$.

Proof. The proof consists in (1) constructing a subsequence from $\lambda_1, \ldots, \lambda_m$ such that the empirical scalarised risks converge to appropriate limits, (2) using these subsequences to apply concentration inequalities to the excess risk when the subsequence exists and (3) combining the results together in the general case.

(1) – Constructing an appropriate subsampling procedure

Let λ be a random variable with probability density function Q over Λ. It induces a real-valued distribution over the risks $\rho_\lambda[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot, \lambda))$ and $\rho_\lambda[\hat{\boldsymbol{\mathcal{R}}}](\hat{h}(\cdot, \lambda))$. We assume that $\left(\rho_\lambda[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot, \lambda)), \rho_\lambda[\hat{\boldsymbol{\mathcal{R}}}](\hat{h}(\cdot, \lambda))\right)$ admits a continuous density with respect to the Lebesgue measure in $\mathbb{R}^2$ we denote p. Further, define

$$\alpha_{\text{op}} = \rho_{\lambda_{\text{op}}}[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot, \lambda_{\text{op}}))$$
$$\beta_{\text{op}} = \rho_{\lambda_{\text{op}}}[\hat{\boldsymbol{\mathcal{R}}}](\hat{h}(\cdot, \lambda_{\text{op}})).$$

Since $Q(\lambda_{\text{op}}) > 0$, we have $p(\alpha_{\text{op}}, \beta_{\text{op}}) > 0$. Then by Proposition A.6, given $\lambda_1, \ldots, \lambda_m \sim Q(\lambda)$ IID, there exists $r \geq 1$ and a random subsampling $\pi([m]) \in 2^{[m]}$ such that for any index $i \in \pi([m])$ we have

$$\mathbb{E}\left[\left(\rho_{\lambda_i}[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot, \lambda_i)), \rho_{\lambda_i}[\hat{\boldsymbol{\mathcal{R}}}](\hat{h}(\cdot, \lambda_i))\right)\right] = (\alpha_{\text{op}}, \beta_{\text{op}}).$$

In particular, let $p = |\pi([m])| \sim \text{Binomial}(m, 1/r)$ denote the number of subsampled elements and assume without loss of generality these are the first p ones. Then, conditionally on $p \geq 1$, we have that $\frac{1}{p}\sum_{i=1}^p \rho_{\lambda_i}[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot, \lambda_i))$ and $\frac{1}{p}\sum_{i=1}^p \rho_{\lambda_i}[\hat{\boldsymbol{\mathcal{R}}}](\hat{h}(\cdot, \lambda_i))$ are respectively unbiased estimators of $\alpha_{\text{op}} = \rho_{\lambda_{\text{op}}}[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot, \lambda_{\text{op}}))$ and $\beta_{\text{op}} = \rho_{\lambda_{\text{op}}}[\hat{\boldsymbol{\mathcal{R}}}](\hat{h}(\cdot, \lambda_{\text{op}}))$.

(2.1) – Bounding the regret when $p \geq 1$ is fixed

Suppose we are in a fixed setting where $p \geq 1$. Then we can decompose and upper bound the regret following

$$\begin{aligned}
\rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}](\hat{g}(\cdot,\lambda_{\text{op}})) - \rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}]^\star &= \rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}](\hat{g}(\cdot,\lambda_{\text{op}})) - \rho_{\lambda_{\text{op}}}[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot,\lambda_{\text{op}})) \\
&\quad + \rho_{\lambda_{\text{op}}}[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot,\lambda_{\text{op}})) - \frac{1}{p}\sum_{i=1}^{p} \rho_{\lambda_i}[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot,\lambda_i)) \\
&\quad + \frac{1}{p}\sum_{i=1}^{p} \rho_{\lambda_i}[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot,\lambda_i)) - \rho_{\lambda_{\text{op}}}[\hat{\boldsymbol{\mathcal{R}}}](\hat{h}(\cdot,\lambda_{\text{op}})) \\
&\quad + \rho_{\lambda_{\text{op}}}[\hat{\boldsymbol{\mathcal{R}}}](\hat{h}(\cdot,\lambda_{\text{op}})) - \rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}]^\star \\
&\leq \rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}](\hat{g}(\cdot,\lambda_{\text{op}})) - \rho_{\lambda_{\text{op}}}[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot,\lambda_{\text{op}})) \\
&\quad + \rho_{\lambda_{\text{op}}}[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot,\lambda_{\text{op}})) - \frac{1}{p}\sum_{i=1}^{p} \rho_{\lambda_i}[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot,\lambda_i)) \\
&\quad + \frac{1}{p}\sum_{i=1}^{p} \rho_{\lambda_i}[\hat{\boldsymbol{\mathcal{R}}}](\hat{h}(\cdot,\lambda_i)) - \rho_{\lambda_{\text{op}}}[\hat{\boldsymbol{\mathcal{R}}}](\hat{h}(\cdot,\lambda_{\text{op}})) && \text{(Lemma A.7)} \\
&\quad + \rho_{\lambda_{\text{op}}}[\hat{\boldsymbol{\mathcal{R}}}](h^*(\cdot,\lambda_{\text{op}})) - \rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}]^\star && \left(\hat{h}(\cdot,\lambda_{\text{op}}) \in \arg\min \rho_{\lambda_{\text{op}}}[\hat{\boldsymbol{\mathcal{R}}}]\right) \\
&\leq 2\sup_{f\in\mathcal{H}} \left|\rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}](f) - \rho_{\lambda_{\text{op}}}[\hat{\boldsymbol{\mathcal{R}}}](f)\right| \\
&\quad + \left|\rho_{\lambda_{\text{op}}}[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot,\lambda_{\text{op}})) - \frac{1}{p}\sum_{i=1}^{p} \rho_{\lambda_i}[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot,\lambda_i))\right| \\
&\quad + \left|\frac{1}{p}\sum_{i=1}^{p} \rho_{\lambda_i}[\hat{\boldsymbol{\mathcal{R}}}](\hat{h}(\cdot,\lambda_i)) - \rho_{\lambda_{\text{op}}}[\hat{\boldsymbol{\mathcal{R}}}](\hat{h}(\cdot,\lambda_{\text{op}}))\right|
\end{aligned}$$

Let $\eta \in (0,1)$ fixed. By linearity of ρ_λ, we have shown that $\rho_\lambda[\hat{\boldsymbol{\mathcal{R}}}](f)$ is an unbiased estimator of $\rho_\lambda[\boldsymbol{\mathcal{R}}](f)$. Therefore, McDiarmid's inequality gives us that we have with probability at least $1-\eta/3$

$$\left|\rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}](f) - \rho_{\lambda_{\text{op}}}[\hat{\boldsymbol{\mathcal{R}}}](f)\right| \leq M\sqrt{\frac{\log(6/\eta)}{2n}}.$$

If we denote $Z_{\lambda_i} = \rho_{\lambda_i}[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot,\lambda_i))$ for the p accepted samples from the rejection sampling procedure, then we have by construction that $\frac{1}{p}\sum_{i=1}^{p} Z_{\lambda_i}$ is an unbiased estimator of $\rho_{\lambda_{\text{op}}}[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot,\lambda_{\text{op}}))$. Therefore, we can also apply McDiarmid's inequality to obtain that with probability at least $1-\eta/3$

$$\left|\rho_{\lambda_{\text{op}}}[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot,\lambda_{\text{op}})) - \frac{1}{p}\sum_{i=1}^{p} \rho_{\lambda_i}[\hat{\boldsymbol{\mathcal{R}}}](\hat{g}(\cdot,\lambda_i))\right| \leq M\sqrt{\frac{\log(6/\eta)}{2p}}.$$

Applying the same reasoning to the last line and combining the bounds together using the union bound we get that with probability at least $1-\eta$

$$\rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}](\hat{g}(\cdot,\lambda_{\text{op}})) - \rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}]^\star \leq 2M\sqrt{\frac{\log(6/\eta)}{2n}} + 2M\sqrt{\frac{\log(6/\eta)}{2p}}.$$

(2.2) – Integrating the upper bound against p given $p \geq 1$

We now consider the random setting, conditional on $p \geq 1$. Recall that the number of accepted samples from the rejection sampling procedure p follows a $\mathrm{Binomial}(n, 1/r)$ distribution. We want to take the expectation of the established probabilistic upper bound with respect to p given that $p \geq 1$. Let $q = 1 - 1/r$ denote the rejection rate, then we have for any $k \geq 1$

$$\mathbb{P}(p = k \mid p \geq 1) = \frac{1}{1 - q^m} \binom{m}{k} (1 - q)^k q^{m-k}.$$

This corresponds to a positive Bernoulli distribution (Grab and Savage, 1954), and in particular if we denote $B_{m,r}(k) = \mathbb{P}(\mathrm{Bernoulli}(m, 1/r) \leq k)$, we have from (Grab and Savage, 1954) Eq. (12) that

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{p} \mid p \geq 1\right] &\leq \frac{1}{(m+1)(1-q)(1-q^m)} \left([1 - B_{m+1,r}(1)] + \frac{3}{(1-q)(m+2)} [1 - B_{m+2,r}(2)] \right) \\
&\leq \frac{1}{m(1-q)(1-q^m)}.
\end{aligned}$$

Therefore, it follows that

$$\begin{aligned}
\mathbb{E}\left[\sqrt{\frac{\log(6/\eta)}{2p}} \mid p \geq 1\right] &= \mathbb{E}\left[\sqrt{\frac{\log(6/\eta)1/p}{2}} \mid p \geq 1\right] \\
&\leq \sqrt{\frac{\log(6/\eta)\mathbb{E}[1/p \mid p \geq 1]}{2}} && \text{(Jensen)} \\
&\leq \sqrt{\frac{\log(6/\eta)}{2m(1-q)(1-q^m)}},
\end{aligned}$$

and by applying this to the probabilistic upper bound on the excess risk we have obtained earlier, we get that with probability at least $1 - \eta$

$$\rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}](\hat{g}(\cdot, \lambda_{\text{op}})) - \rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}]^\star \leq 2M\sqrt{\frac{\log(6/\eta)}{2n}} + 2M\sqrt{\frac{\log(6/\eta)}{2m(1-q)(1-q^m)}}.$$

(3) – Combining things together

Now that we have established a probabilistic upper-bound on the excess risk when at least one sample is accepted by π, we set out to obtain a general probabilistic bound on the excess risk. Let $q = 1 - 1/r$ be the rejection rate of the rejection sampling procedure and fix $\delta \in (q^m, 1)$.

Take $\eta_\delta = (\delta - q^m)/(1 - q^m)$ and $\varepsilon_\delta = 2M\sqrt{\frac{\log(6/\eta_\delta)}{2n}} + 2M\sqrt{\frac{\log(6/\eta_\delta)}{2m(1-q)(1-q^m)}}$, then we have

$$\begin{aligned}
\mathbb{P}\left(\rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}](\hat{g}(\cdot, \lambda_{\text{op}})) - \rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}]^\star > \varepsilon_\delta\right) &= \underbrace{\mathbb{P}\left(\rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}](\hat{g}(\cdot, \lambda_{\text{op}})) - \rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}]^\star > \varepsilon_\delta \mid p = 0\right)}_{\leq 1} \underbrace{\mathbb{P}\left(p = 0\right)}_{= q^m} \\
&\quad + \mathbb{P}\left(\rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}](\hat{g}(\cdot, \lambda_{\text{op}})) - \rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}]^\star > \varepsilon_\delta \mid p \geq 1\right) \mathbb{P}\left(p \geq 1\right) \\
&\leq q^m + (1 - q^m)\mathbb{P}\left(\rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}](\hat{g}(\cdot, \lambda_{\text{op}})) - \rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}]^\star > \varepsilon_\delta \mid p \geq 1\right) \\
&\leq q^m + (1 - q^m)\eta_\delta \\
&= q^m + (1 - q^m)\frac{\delta - q^m}{1 - q^m} = \delta,
\end{aligned}$$

where the last derivations follow from the construction of ε_δ and η_δ. This shows that for any $\delta \in (q^m, 1)$, the following inequality holds with probability $1 - \delta$

$$\rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}](\hat{g}(\cdot, \lambda_{\text{op}})) - \rho_{\lambda_{\text{op}}}[\boldsymbol{\mathcal{R}}]^\star \leq 2M\sqrt{\frac{\log(6/\eta_\delta)}{2n}} + 2M\sqrt{\frac{\log(6/\eta_\delta)}{2m(1-q)(1-q^m)}},$$

where $\eta_\delta = (\delta - q^m)/(1 - q^m)$. This concludes the proof.

□

B. Conditional Value-at-Risk (CVaR)

Proposition B.1. *Let $I = \{1, \dots, m\}$ be an index set and $R : I \to \mathbb{R}_+$ such that $R(i) = \hat{R}_i$ for $i \in I$. Denote $C(I)$ as the space of real-valued, continuous function on I and $C(I)^*$ its dual, i.e., $\{T : C(I) \to \mathbb{R}\}$. Then there is a finite measure μ on I such that for any $T \in C(I)^*$ and $R \in C(I)$, we have*

$$T(R) = \sum_{i \in I} R_i \mu_i.$$

Sketch Proof. The key is to notice I is a compact metric space because it is bounded. Furthermore, all functions on discrete space are automatically continuous. This allows us to directly apply the Riesz-Markov-Kakuani representation theorem. □

The proposition implies that no matter how we aggregate a risk profile, it will always correspond to some kind of weighted average. From the perspective of optimisation, since these weights are always convex (noramlising the weights does not change the optimisation), it can then be understood that whenever we aggregate the risk profile, we are picking a particular weighted distribution to perform the standard ERM.

C. Single-Domain Scenario

In a single-domain setting, we envision two possible approaches to imprecise learning. The first approach treats each training data point as an individual domain, estimating the risk profile through point-wise loss functions, denoted as $\boldsymbol{\mathcal{R}}(f) = (\ell(f(x_1), y_1), \dots, \ell(f(x_n), y_n))$. The second approach delineates a credal set by an ϵ-ball around the empirical distribution of the training data, akin to Distributionally Robust Optimisation (DRO). Subsequently, it extracts a finite number of extreme points from this credal set, which then represent the risk profile. While the first approach can be directly implemented within the current framework, the second approach entails a non-trivial extension of the existing setup.

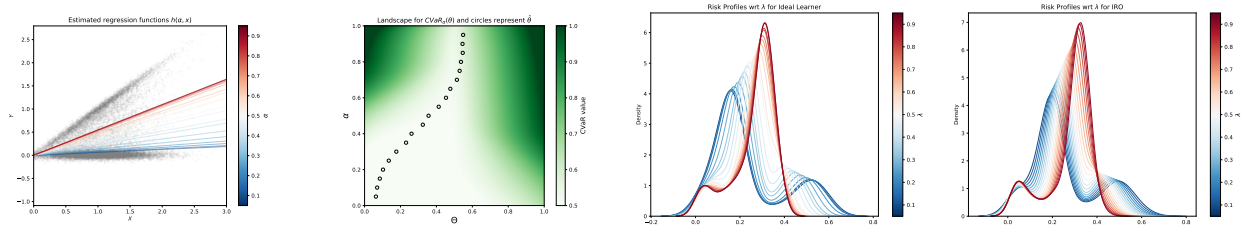
D. Risk Profiles of Simulation

Simulation of Risk Profile: In economic theory, risk aversion explains the inclination to accept a situation with a more predictable but possibly lower payoff than another situation with a very unpredictable but possibly higher payoff. In OOD research, the term risk averseness has been conceptually used to describe the operator's risk perception for the model's risk profile (i.e., the distribution of $\hat{R}$). A risk-averse operator prefers models whose risk is more predictable but possibly higher than models whose risk is less predictable but possibly lower. Given that the operator at test time have a risk averseness between "less risk averse" and "risk averse" and by having $h(x, \lambda)$ we can cover this spectrum of the operator's risk averseness. Given that we use CVaR, the entire spectrum of an ML Operators potential risk averseness is encoded in the interval of λ being between 0 and 1. By construction, $h(x, \lambda)$ can cover the spectrum of "risk averseness" because it corresponds to the prediction function we obtain at $CVaR(\lambda)$. Hence, we verify this hypothesis.

Experiment 1A: Assume a linear model $Y_e = \theta_e X + \epsilon$, where $X \sim \mathcal{N}(2, 0.2)$ and $\epsilon \sim \mathcal{N}(0, 0.1)$. We simulate different environments by drawing θ from a Beta distribution $Beta(0.1, 0.2)$. In total, we generate for 250 train and test domains 100 observations each.

Each data line corresponds to a domain in Figure 3a. Hence the domains differ in their slope. Since we take θ from the bimodal distribution $Beta(0.1, 0.2)$, we observe that the domains form two clusters. The more dominant cluster includes the domains with smaller θ. Subsequently, we aim to find the optimal $\hat{\theta}$ for all $\lambda \in \{0.05, \dots, 0.95\}$ by solving the corresponding CVaR objective. As we can see from this plot, the optimal lines for small values of λ cluster around the dominant cluster of the environments. We consider the dark blue line (λ=0.05) as the "average case". When increasing λ, the lines get closer to the second cluster of domains, which could be considered as the "worst-case". Hence, the dark red line (λ=0.95) could be somewhat considered to be the estimated θ that works well in the worst cases.

Figure 3: Figure 3a illustrated the data and the ideal learner $f_\lambda(\hat{\theta}) \in \mathcal{H}$ for $\lambda \in \{0.05, \ldots, 0.95\}$. Figure 3b describes the landscape of the objective function ρ (CVaR) for the ideal learner. We plot $\hat{\theta}$ as circles.Figure 3c describes the Risk profile for $\lambda \in \{0.05, \ldots, 0.95\}$ for the ideal learner. Figure 3d describes the Risk profile for $\lambda \in \{0.05, \ldots, 0.95\}$ Imprecise Learner.



(a) (b) (c) (d)

In Figure 3b, we observe that for higher values of λ, the optimal solutions for θ vary a lot, while for smaller values of λ, the optimal solutions for θ do not vary significantly. As an interpretation, we can say that it is likely that the problem becomes harder for small λ. This interpretation is supported by the fact that when we choose λ to be high, we condition on the tail of the $\boldsymbol{\mathcal{R}}$, thus considering only a subset of the domains (i.e., lesser data for optimization). When looking at the optimal $\hat{\theta}$, they form a smooth curve across all $\lambda \in \{0.05, \ldots, 0.95\}$.

Lastly, in Figure 3c, we see how the distribution of the risk changes across all λ. As expected, when choosing higher λ we consider higher risks from the risk profile $\boldsymbol{\mathcal{R}}$ and minimize these parts in the optimization. We observe that for higher values of λ the risk profile does not transition smoothly contrary to the case of IRO in Figure 3d. We postulate that this is because an ideal learner essentially throws away the data from low-risk domains when focusing on high-risk domains due to the formulation of CVaR as an aggregator. However, since IRO learns all the objectives simultaneously it can implicitly address this issue of a finite number of domains for training in λ corresponding to higher risks. This observation is consistent with our observation from real-world experiments on UCI bike rentals in Figure 2c.

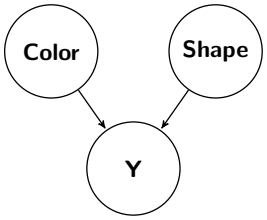
E. Experiments on CMNIST

E.1. Dataset Setup

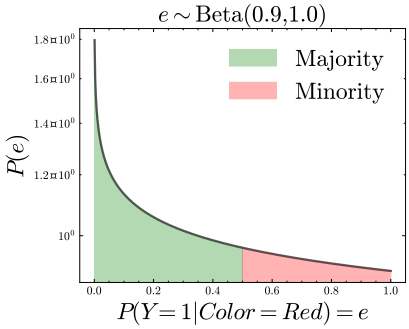
We conduct a large-scale experiment using an extension of the CMNIST dataset (Arjovski, 2021). The CMNIST comprises data from the MNIST dataset modified to the task of binary classification. For the standard task in CMNIST, the digits (0-4) and (5-9) have to be classified into two labels 0 and 1. Another 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 $P(Y = 1 \,|\, \text{shape} \in \{5, 6, \ldots, 9\}) = 0.75$.

E.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 $\{0.01, 0.02, 0.05, 0.07, 0.09, 0.12, 0.14, 0.58, 0.7, 0.99\}$. 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



(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.

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).

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

Table 3: 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 $\mathbb{P}(Y=1 \mid \text{color}=\text{red})=e$											**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**	

In Table 3 we compare **IL** to other methods, showing that **IL** can learn relevant features in context. This also allows us to guide model operators on selecting 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 $e \in \{0.0, \ldots, 0.4\}$ it performs similar to average case learner and for relatively less seen i.e. minority domains $e \in \{0.5, \ldots, 1.0\}$ it performs similar to the invariant learner.

F. Limitations of Imprecise Learner

F.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.

F.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, model operators may encounter challenges in precisely articulating their level of risk aversion. Additionally, bridging the operator's concept of generalization to a specific domain with an appropriate risk level remains ambiguous. In our experiments on modified CMNIST, we address this by allowing the model operator to be more risk-averse to generalize to minority environments. In contrast, for generalizing to a domain from majority environments users can be more risk-seeking.

F.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.

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

In Table 4, 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 4: CMNIST Test Accuracy. Training Environments are $\{0.1, 0.2\}$ & Test Environment $\{0.9\}$

Objective	**Algorithm**	**Initialization**		
		Rand.	ERM	Best Case
PL-f	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
	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	$\mathbf{71.6 \pm 0.5}$	$\mathbf{71.6 \pm 0.5}$
	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	$\mathbf{71.4 \pm 0.2}$	65.4 ± 0.1	71.4 ± 0.2
Invariant Pred.	Oracle		$\mathbf{73.5 \pm 0.2}$	

G. Implementation Details

This section provides the details of specific implementations used in our experiments.

G.1. Augmented Hypothesis

For implementing the augmented hypothesis, we use hypernetworks (Ha et al., 2016) to realize the dependence of h on model operator's preference, i.e., λ. In this scenario, the weights of the augmented model are dependent on λ, i.e., $h_\xi(x, \lambda) := f_{g_w(\lambda)}(x)$ where $g_w(\lambda)$ is the hypernetwork and $\xi := \{w, g_w(\lambda)\}$. For neural networks with multiple layers, we use FILM layers (Perez et al., 2018) to augment the network such that it can be conditioned upon λ.

G.2. Imprecise Risk Optimisation

To operationalise the imprecise risk optimization, we need to minimise (9) with respect to the family of probability distributions $\Delta(\Lambda)$. Since for our case $\Lambda = [0, 1]$, we parameterise the family of distributions with Beta(α, β). We sample λ from Q via uniform sampling from the inverse CDF of Q, which we denote as F^{-1}. We approximate the gradient of F^{-1} by first-order difference as described in Algorithm 3.

Algorithm 3 Sampling from a Beta Distribution using ICDF with Gradient Computation

class ICDFBeta:
  def forward(u): # Compute ICDF
    return $\mathrm{F}^{-1}(\alpha,\beta)(u)$
  def backward(u): # Compute Gradient
    $\delta := 1e-6$
    $\nabla_\theta F^{-1}(\alpha, \beta)(u) := \frac{(F^{-1}(\alpha+\delta,\beta)(u)-F^{-1}(\alpha,\beta)(u))}{\delta}$
    $\nabla_\phi F^{-1}(\alpha, \beta)(u) := \frac{(F^{-1}(\alpha,\beta+\delta)(u)-F^{-1}(\alpha,\beta)(u))}{\delta}$
    return $\nabla_\alpha F^{-1}(\alpha, \beta)(u)$, $\nabla_\beta F^{-1}(\alpha, \beta)(u)$
Initialize: $\alpha, \beta \leftarrow 1.0, 1.0$
icdfbeta = ICDFBeta(α, β)
for epoch $= 1$ **to** k **do**
  for $i = 1$ **to** m **do**
    $u_i \sim \mathrm{Uniform}([0, 1])$
    λ_i = icdfbeta.forward(u_i)
  end for
end for
Return Set of samples $\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ and gradients for each epoch