UNIVERSITY OF CAMBRIDGE

Department of Engineering

Set Valued Predictions for Human-AI Teams

Author Name: Varun Babbar

Supervisor: Dr Adrian Weller

Date: 1st June 2022

I hereby declare that, except where specifically indicated, the work submitted herin is my own original work.

Signed ______________ *date* 1st June 2022

Acknowledgements

I'm truly grateful to my supervisor, Dr Adrian Weller for his guidance and mentorship throughout my 4^{th} year and for all the fruitful and insightful discussions we have had. I also want to express my heartfelt gratitude to Umang Bhatt, Dr Weller's PhD student, who has helped me *immensely* throughout the year. Our meetings have always left me feeling challenged, humbled, motivated, and inspired. Under Umang's and Dr Weller's guidance, I was able to submit a part of this thesis to the International Joint Conference in Artificial Intelligence 2022, where it has been accepted as a conference paper with a long oral presentation.

I also want to give a quick shout out to Anastasios Angeloupoulos, the author of some seminal papers in distribution free uncertainty quantification (which served as a major inspiration for this thesis), whom I had the pleasure of meeting in person. Thank you for the joyful punt ride, Jack's Gelato, and the interesting conversations!

This section would, of course, be incomplete without an appreciation for the pivotal role my parents, my sister, and my friends have played in my life. Thank you for all the learnings, the experiences, and the joy you have given me throughout my time at university and in my life.

Declaration

A part of this thesis, specifically most of Section 4, has been submitted to the International Joint Conference in Artificial Intelligence, Vienna, 2022, where it has been accepted as a conference paper with a long oral presentation (~15% of submissions were accepted and 25% of accepted submissions were given a long oral presentation slot). The *ArXiv* version of the paper can be found at [6].

I hereby declare that except where specific reference is made to the work of others, the contents of this report are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This report is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and in Acknowledgements. This report contains fewer than 50 pages including footnotes, figures, tables, equations, appendices, and references.

Technical Abstract

Machine learning models are increasingly being used in many real world settings involving high stakes decision making, such as medical diagnostics, computational drug discovery, and hazard detection in industrial applications. In these situations, they are often expected to work in conjunction with the human, providing complementary and useful information that assists the human in deciding the best course of action. To this end, several works have explored the benefits of leveraging the complementary strengths of the human and AI to achieve improved accuracy, trust, and fairness. However, to the best of our knowledge, research on human-AI teams in multi-class problems has so far provided experts with a single label, which ignores the uncertainty in a model's recommendation. As it is crucial for the human to be able to gauge and interpret the uncertainty of a model in order to facilitate robust decision making, we explore one notion of model uncertainty in this thesis - set valued predictions. Specifically, we explore how we can generate predictive sets in multi-class problems that are not only useful to a human in the sense of providing actionable uncertainty quantification, but also satisfy attractive theoretical properties such as provable marginal risk control (for example, false negative coverage) at any chosen level. Through our exploration, we make the underlying argument that principled risk control is a multi-faceted problem that needs to be tackled from the perspective of both the black box machine learning model and the downstream decision maker insofar as deployment in high stakes situations is concerned.

We consider a class of non-parametric, distribution free, risk controlling techniques called Conformal Prediction (CP) that are widely studied in literature. These techniques construct predictive sets that are guaranteed to control the false negative risk to a user specified level $1-\alpha$, with some extensions such as Risk Controlling Predictive Sets (RCPS) that are able to control for any user defined risk function to a level γ with high probability $1-\delta$. We first study the utility of CP sets in human-AI teams through human-subject experiments, uncovering that principled, calibrated set predictions are perceived by human experts to be statistically significantly more useful and trustworthy than Top-1 predictions or any randomly generated set prediction. We then improve upon the CP baseline by introducing a scheme we call D-CP - a method where the model learns to defer some examples to an expert and output CP sets on others.

- Our empirical experiments on CIFAR-10H and CIFAR-100 using synthetic experts suggest that D-CP not only improves the overall human-AI team accuracy, but also yields smaller predictive set sizes on non-deferred examples for the same error tolerance.

- Our human subject experiments show that there is a statistically significant increase in perceived utility, trust, overall expert accuracy, and a reduction in bias towards incorrect labels when humans are shown D-CP sets vs CP sets. However, these improvements come at a small cost in terms of a reduction in statistical efficiency of the scheme.

- We also prove some theoretical properties that complement our experiments:
  - A principled deferral policy that defers examples a model is less confident on can guarantee smaller set sizes on non-deferred examples. (Theorem 1)
  - We can generate predictive sets that can simultaneously guarantee risk control on non-deferred examples and control for the misclassification rate of the expert on deferred examples. (Theorem 2)

We evaluate more general predictive risk control properties of D-CP through experimentation with a real world dataset where we want to use information about the radicalisation process and socioeconomic background of documented extremists to predict their tendency to commit violent crimes. We empirically verify Theorem 2 for the case where we want to control for the cost of misclassification of a violent individual.

Lastly, we consider the fact that that, in many real world applications, it may be impractical to obtain instance-level human annotations for large training datasets, limiting the utility of a scheme like D-CP. Thus, we analyse the case where a human expert can provide global, non-instance specific information that the model can use to generate more useful predictive sets. Hereto, we develop a post-hoc heuristic that enables users to construct predictive sets that are as similar as possible, where the notion of label similarity, which may be context dependent, is provided by the user in the form of a label similarity matrix that encodes similarity distances between labels. We evaluate our method using BERT embedding distances between labels on the CIFAR-100 dataset, exploring the relationship between predictive set size, controlled risk, and the dissimilarity cost. We show that we can generate sets that contain semantically similar labels, whilst maintaining the desired risk guarantees.

This project leaves many directions for future work in terms of human subject evaluation, theoretical and statistical analysis, and algorithmic development.

- We need to quantify and qualify the properties of predictive sets that humans may find useful. Knowledge of these properties can then help us evaluate the effect of generating sets shaped using a label similarity matrix on the performance of human-AI teams.
- From an algorithmic perspective, we may be able to generate more useful D-CP sets if we develop improved deferral policies that are designed with the model's predictive uncertainty in mind.
- Lastly, future theoretical work could generalise CP to provide conditional coverage guarantees for any label, which would be useful because the presence of dangerous labels (e.g. `cancer`) in a set could potentially bias the human's predictions. Thus, we may want to bound the probability a CP set should contain any chosen label, i.e. the "error" of a CP scheme. We provide a starting point for this in Theorem 3.

Glossary of Terms

Abbreviations

CP	Conformal Prediction
D-CP	Deferral - Conformal Prediction (also more generally D-RCPS)
ICP	Inductive Conformal Prediction
LTT	Learn Then Test Framework
MoG	Mixture of Gaussians
PIRUS	Profiles of Individual Radicalization in the United States
RCPS	Risk Controlling Prediction Sets
TCP	Transductive Conformal Prediction
UCB	Upper Confidence Bound

Machine Learning

$\mathcal{D}$	Dataset
$\mathcal{D}_{cal}$	Calibration Dataset
$\mathcal{D}_{val}$	Validation Dataset
$\mathcal{P}$	Input / Example Space
$\mathcal{Y}$	Label Space
$g_y(X)$	Unnormalized Classifier output for label y
$r(X)$	Deferral Policy
X, x	Input
y, y'	Any Label (Not Necessarily Target)
Y	Target Label

Miscellaneous Notation

μ	Dissimilarity Penalty
$d(y, y')$	Label Distance
K	Number of Classes
M	Label Similarity Matrix
N	Size of Dataset / Number of Samples
N_{val}	Size of the Validation Dataset

Probability and Statistics

α	CP Error Tolerance
δ	Excess Risk Probability
$\Gamma(X)$	CP Set Valued Predictor
γ	Desired Risk Tolerance
$\Gamma_\lambda(X)$	RCPS Set Valued Predictor
$\hat{R}(\lambda)$	MC Estimate of Risk
λ	RCPS Threshold Parameter
$\mathbb{E}(.)$	Expectation Operator
$\mathbb{I}(.)$	Indicator Function
$\pi_y(X)$	Softmax output for label y
$\tau(X, y)$	Conformity Score Function
$F_\tau(.)$	Cumulative Distribution Function
H_0	Null Hypothesis
H_1	Alternate Hypothesis
$L(Y, \Gamma_\lambda(X))$	Set Loss Function
$R(\lambda)$	Risk: Expected Loss
$R^+(\lambda)$	UCB of Risk
$u_{(n)}$	n^{th} order statistic of a collection of random variables $\{u_i\}_{i=1}^N$

List of Figures

List of Tables

Table of Contents

1 Introduction

> Machines can never think as humans do but just because something thinks differently from you, does it mean it's not thinking?
>
> Alan Turing, *The Imitation Game, 2014*

1.1 Motivation

In a world where decision making is becoming increasingly data driven, there is a need to design human-AI teams driven by trust and performance. To this end, several works have explored the benefits of leveraging the complementary strengths of the human and AI by devising loss functions or training procedures that optimize for joint utility [41], accuracy [25, 32], and trust [8]. However, to the best of our knowledge, research on human-AI teams has so far provided experts with a single label, which ignores the uncertainty in a model's recommendation. In time and cost sensitive domains such as medical diagnostics, computational drug discovery, and hazard detection in industrial applications, it is important for human operators to be able to gauge and interpret the uncertainty in a model's predictions in order to facilitate robust decision making. To this end, in this report, we explore one notion of transparency - the uncertainty in an AI's predictions [12]. Specifically, we explore the idea of prediction sets in the context of classification models working alongside a human expert. For a classification model, we define a set valued model prediction $\Gamma(X)$ as a mapping from the input space $\mathcal{X}$ to the power set of the label space $\mathcal{Y}$, i.e. $\Gamma : \mathcal{X} \to 2^{\mathcal{Y}}$. Our ultimate aim is to generate predictive sets that satisfy certain guarantees and are maximally useful to humans. The problem henceforth is two-fold:

- How do we define what a useful set is? One definition of set utility is the size of the set, but smaller set sizes are less likely to contain the true label, especially if they are not certified to control for the false negative rate. In situations where the model is overly confident about a wrong example, the resulting predictive set, although small in size, can accidentally rule out dangerous labels, incur high costs as a result, (e.g. compromising the health and well being of a patient) and ultimately erode the expert's trust in the model. While there are trade-offs in terms of risk control and size associated with any set prediction scheme, we ultimately need to provide uncertainty quantification to an expert in a manner that is *actionable* and does not compromise the reliability of the model.

- How do we learn to generate useful sets? We want to be able to construct sets that take into account the strengths of an expert and satisfy guarantees pertaining to risk / coverage whilst being maximally useful (where sets satisfy the notion of utility prescribed by the expert). For example, if a doctor is specialized in a certain medical domain, we want the AI assistant to adapt to the competency of the doctor and a) either output prediction sets that are consistent with the doctor's evaluations or b) avoid making predictions and *defer* to the doctor.

1.2 Outline of Report

In this report, we develop tools for constructing predictive sets that are useful to humans.

- In Section 2, we introduce Conformal Prediction - a well established line of research that aims to build a theoretically grounded, calibrated prediction set which contains the true label with high probability - and a related scheme in literature: Risk Controlling Predictive Sets.

- In Section 3, we provide a short overview of the rejection learning literature in machine learning and discuss a scheme by [26] that enables a model to defer some examples to an expert.

- In Section 4, we introduce a paradigm we call D-CP that *Defers* some examples to an expert and performs *Conformal Prediction (or RCPS)* on others. We provide a theoretical and empirical analysis of the scheme and validate its benefits in Human-AI teams through human subject experiments. *This work has been accepted at IJCAI'22 (see [6]).*

- In Section 5, we apply D-CP to a real world dataset where the task is to predict whether a radicalised person will commit violent crimes. Here, we generate predictive sets that can simultaneously control for the expert's misclassification rate on deferred examples and any risk associated with non-deferred examples.

- Lastly, in Section 6, we consider a broader notion of useful sets: those that are as similar as possible, where we the notion of similarity is provided by the user. As a first step in this direction, we develop and evaluate a method for generating predictive sets that contain labels that are as semantically similar as possible whilst satisfying desired coverage / risk guarantees.

- We then draw the main conclusions of this project and provide directions for future research in Section 7.

2 Background: Generating Calibrated Set-Valued Predictions

With our ultimate aim being to generate efficient, calibrated, and useful set valued predictions, we examine some earlier approaches to set construction in this section. These methods are non-parametric, make no assumptions on the type of machine learning model being employed, the distribution of training examples, and the accuracy of the model. Instead, they serve as a wrapper that can be added on top of existing models, making them lightweight, interpretable, and computationally efficient. They are guided by 3 desiderata [4]:

- **Coverage**: The main class of set valued predictors considered in this report require the user to specify an error tolerance α. Given this, we construct sets that are guaranteed to contain the true label with at least $1-\alpha$ probability.
- **Size**: For a given coverage level, the sets must be as small as possible in order to be most useful.
- **Adaptiveness**: The sets must convey instance-wise uncertainty, i.e. they should generally be larger for examples the model found difficult and smaller for easy examples.

Definition 2.1 (Coverage)**.** Coverage is the probability that a set valued prediction contains the true label, i.e.

$$C = P(Y \in \Gamma(X)) = \mathbb{E}[\mathbb{I}_{Y\in\Gamma(X)}] \tag{2.1}$$

We estimate it empirically using

$$\hat{C} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{Y_i \in \Gamma(X_i)} \approx C \tag{2.2}$$

This is an intrinsic property of a set valued predictor which we seek to control in expectation.

2.1 The Naive Procedure

Consider a classification model $m_\theta(x) : \mathcal{X} \to \mathcal{Y}$ parameterised by weights θ acting on an input x. The output of the model is a softmax distribution over the label space $\mathcal{Y}$

$$\pi_y(x) = \frac{\exp(g_y(x))}{\sum_{y'=1}^{K}\exp(g_{y'}(x))} \tag{2.3}$$

where $g_y(x)$ is the raw model output corresponding to label y. Assume we have access to an oracle classifier with perfect knowledge of the conditional label distribution $P_{Y|X}$. Equivalently, the softmax probabilities of this classifier are perfectly *calibrated.*

Definition 2.2 (Calibrated Distribution [16])**.** A softmax distribution is *calibrated* when, for any label $y \in \mathcal{Y}$, $P(Y = y|\pi_y(X) = p) = p$.

For the oracle classifier, one way to construct a prediction set is to add the most probable labels until the probability mass of the set is equal to $1-\alpha$, i.e.

$$\Gamma(x) = \{y : \sum_{y'=1}^{K} \pi'_y(x)\mathbb{I}_{\pi'_y(x)\geq\pi_y(x)} \leq 1-\alpha\} \tag{2.4}$$

In practice, there are two problems with this approach:

- The probabilities output by a predictive model are rarely calibrated, so the sets do not achieve coverage in practice.
- For examples where the model is not confident, this naive strategy must select many classes before it reaches the desired confidence level, leading to a very large set size.

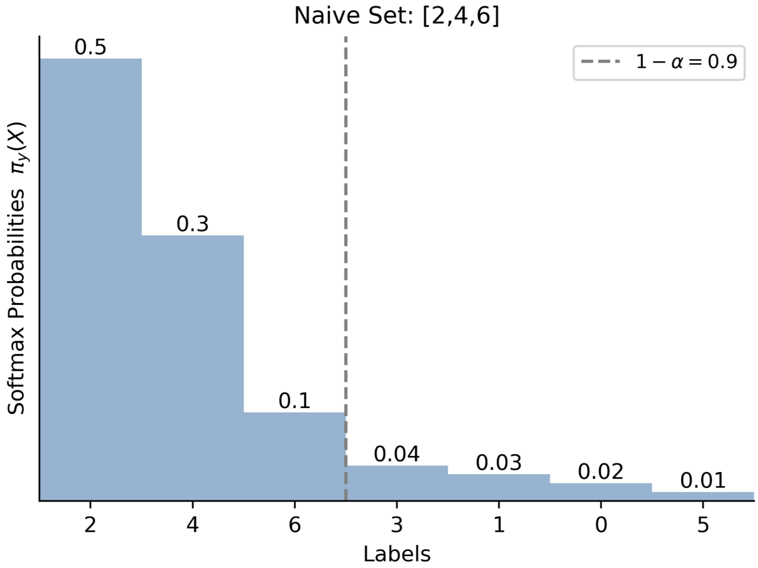


Figure 2.1: The naive set construction procedure: Since softmax scores are not calibrated in practice, the threshold at $1-\alpha=0.9$ does not provide coverage

As we cannot employ the naive procedure to generate calibrated confidence sets that truly reflect model uncertainty, we turn towards more recent, state of the art methods found in literature.

2.2 Conformal Prediction (CP)

2.2.1 Introduction

One way to construct a set valued predictor is through a technique called conformal prediction (CP) [40]. CP provides rigorous uncertainty quantification by generating a prediction set that may contain multiple labels, but contains the true label with a user defined probability $1-\alpha$. The goal of CP is to construct predictive sets that are as small as possible for any given error-rate (or false negative rate) α. Formally, suppose we have a trained classifier $m_\theta(X) \in [0,1]^K$ that outputs softmax scores for each of the K classes $\{\pi_1(X), \pi_2(X), ..., \pi_K(X)\}$. We aim to construct a set valued prediction of the following form:

$$1-\alpha \leq P(Y_{test} \in \Gamma(X_{test})) \tag{2.5}$$

which holds in expectation for any (X_{test}, Y_{test}) that originate from the same distribution as the validation and training datasets. Here, α is the user defined error rate and $\Gamma(X_{test}) \subset \{1, ...K\}$. To construct a conformal predictor, we need to define a conformity score function $\tau(X_{test}, y) : \mathcal{X} \times \mathcal{Y} \to \mathcal{R}$.

Definition 2.3 (Conformity Score)**.** The conformity score $\tau(X, y)$ not only encodes the agreement between the example X_{test} and class y, but can also measure how similar the pair (X_{test}, y) is to previously observed training data $\mathcal{D}_{train} = \{(X_i, Y_i)\}_{i=1}^{N}$. While this function is an important design choice, a common example of a conformity score is $\tau(X, y) = \pi_y(X)$ [34], i.e. the softmax probability of class y.

We now discuss contemporary paradigms for performing CP on any black box classification or regression model. All of the methods explored below rest on the following idea: to include label y in a predictive set for a test example, we require that the conformity score $\tau(X_{test}, y)$ is at least α-common with respect to conformity scores on previously observed data, i.e. $\text{Quantile}(\tau(X_{test}, y), \{\tau(X_i, Y_i)\}_{i=1}^{n}) \geq \alpha^1$.

2.2.2 Transductive Conformal Prediction (TCP)

We define a conformity measure in a Leave-One-Out (LOO) sense. Given an unseen example X_{test} and a test label $\hat{y}$, we define $D' = \{(X_i, Y_i)\}_{i=1}^{N} \cup \{X_{test}, \hat{y}\}$. For each $(X_i, Y_i) \in D'$:

$$\tau(X_i, Y_i) = \tau_{\theta_{D' \setminus X_i, Y_i}}(X_i, Y_i) \tag{2.6}$$

That is, given a model $m_\theta(x)$, for each test label y', we retrain the model $N + 1$ times on $D' \setminus (X_i, Y_i)$, computing the conformity score of each held out datapoint $(X_i, Y_i) \in D'$. The threshold is determined for each y' as:

$$\tau_{y'} = \text{Quantile}\Bigg(\alpha, \{\tau(X_i, Y_i)\}_{i=1}^{N} \cup \tau(X_{test}, y')\Bigg) \tag{2.7}$$

and the set is constructed in the following manner:

$$\Gamma(X) = \{y | \tau(X_{test}, y) \geq \tau_y\} \tag{2.8}$$

As expected, this is a computationally expensive procedure, with the computational cost of prediction + calibration for a single test point being $\mathcal{O}((N + 1)KT)$ where $\mathcal{O}(T)$ is the complexity of training the model once and K is the number of labels. While this ensures the highest statistical power of **predictions** [40], we discuss this scheme here only to motivate more computationally efficient prediction methods and don't evaluate it in this thesis.

2.2.3 Inductive Conformal Prediction (ICP)

Given the computationally expensive nature of TCP, contemporary CP methods [4, 34, 33, 37] have focused mainly on ICP - a more efficient scheme that provides the same guarantees as TCP. Hereto, we train the model only once on the given training data $\mathcal{D}_{train}$ and then use a held out *calibration dataset* $\mathcal{D}_{cal} = \{(X_i, Y_i)\}_{i=1}^{N}$ to determine a threshold:

$$\tau_{cal} = \inf\{\tau | F_\tau(\tau) \geq \alpha\} \approx \text{Quantile}(\alpha, \{\tau(X_i, Y_i)\}_{i=1}^{N}) \tag{2.9}$$

$$\Gamma(X) = \{y : \tau(X, y) \geq \tau_{cal}\} \tag{2.10}$$

where $F_\tau(\tau(X, y))$ is the empirical CDF of $\tau(X, y)$ as constructed from $\mathcal{D}_{cal}{}^2$. This threshold implies that the conformity score $\tau(X_{test}, Y_{test})$ of a test example (X_{test}, Y_{test}) drawn from the same distribution as the training and calibration examples will be greater than τ_{cal}

[1] While *we stick to the conformity score convention in this report*, CP literature sometimes uses the convention of non-conformity scores (e.g. [10]), which is equivalent to the negation of conformity scores and $1 - \alpha$ commonness requirement for label inclusion.

[2] Note that in order to retain coverage guarantees, we have to make a small finite sample correction in practice, i.e. we choose τ_{cal} as the $N - \lceil (N + 1)(1 - \alpha) \rceil$ smallest value in $\{\tau(X_i, Y_i)\}_{i=1}^{N}$. However, this makes no difference to the properties of such predictors analysed henceforth.

with probability at least $1-\alpha$. Figure 2.2 illustrates the pipeline for performing ICP on any dataset. Note that throughout this report, we will use the phrase CP and ICP synonymously.

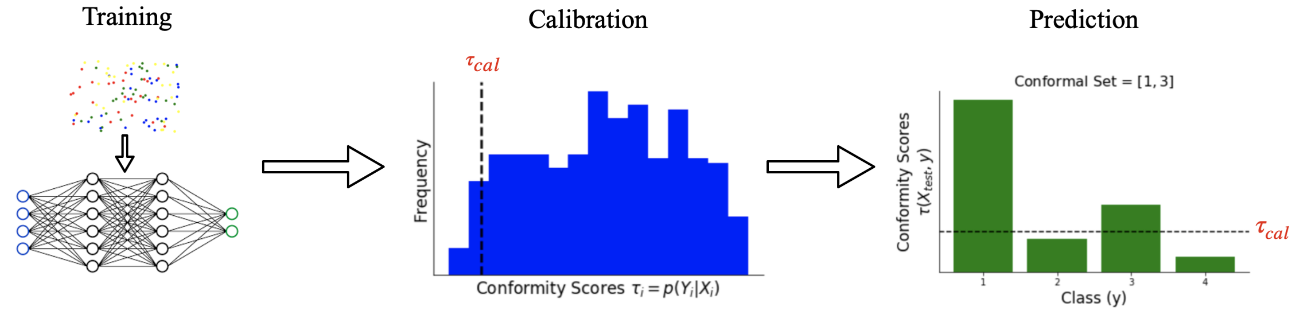


Figure 2.2: Pipeline for Performing ICP on any Dataset

2.2.4 Mondrian Conformal Prediction

The coverage guarantee in Equation 4.8 applies marginally across the dataset and only holds in expectation over possible sets of training data and test points. This is often undesirable and in many cases unfair. If our set predictor is valid at the significance level 5% but makes an error of 10% for certain groups (e.g. stratified by race, gender, income, etc), this can lead to prediction sets that are group-wise unfair and / or ignore the fact that misrepresenting some groups can incur greater cost than others (e.g. misdiagnosing a dangerous disease). A more useful definition of coverage is *conditional coverage*, i.e.

$$1-\alpha \leq P(Y \in \Gamma(X)|X) \tag{2.11}$$

However, this is impossible to achieve in practice, as it requires access to a perfect classifier with calibrated probabilities [33]. In practice, we can obtain an alternate notion of conditional coverage by *group-wise conditioning*, e.g in [24]. That is, given a stratification of the calibration dataset according to known groups $\mathcal{G} = \{G_1, ...G_k\}$, we can obtain the guarantee:

$$1-\alpha \leq P(Y \in \Gamma(X)|X \in G_i) \tag{2.12}$$

or

$$1-\alpha \leq P(Y \in \Gamma(X)|Y \in G_i) \tag{2.13}$$

depending on whether the user decides to stratify the label space or the example space. To apply ICP in a Mondrian sense, we perform groupwise calibration to obtain $|\mathcal{G}|$ thresholds. For the i^{th} group, the threshold is determined as:

$$\tau_{cal}^{G_i} = \inf\{\tau | F_\tau^{G_i}(\tau) \geq \alpha\} \approx \text{Quantile}(\alpha, \{\tau(X_j, Y_j) | X_j \in G_i\}) \tag{2.14}$$

This allows set construction in the following manner:

$$\begin{align} \Gamma(X|X \in \mathcal{G}_i) &= \{y | F_\tau^{G_i}(\tau(X, y)) \geq \alpha\} \tag{2.15} \\ &= \{y | \tau(X, y) \geq \tau_{cal}^{G_i}\} \tag{2.16} \end{align}$$

If we have groups in the label space

$$\tau_{cal}^{G_i} = \inf\{\tau | F_\tau^{G_i}(\tau) \geq \alpha\} \approx \text{Quantile}(\alpha, \{\tau(X_j, Y_j) | Y_j \in G_i\}) \tag{2.17}$$

and we construct sets as follows:

$$\begin{align} \Gamma(X) &= \{y | F_\tau^{G(y)}(\tau(X, y)) \geq \alpha\} \tag{2.18} \\ &= \{y | \tau(X, y) \geq \tau_{cal}^{G(y)}\} \tag{2.19} \end{align}$$

where $G_y \in \mathcal{G}$ is the group label y belongs to. One possible definition of $\mathcal{G}$ in the label space is $\mathcal{G} = \mathcal{Y}$, which guarantees us equal coverage for all labels. Note that so far we specified a single error-rate α that is applicable across groups. We can also specify group-wise error-rates $= \{\alpha_1, ...\alpha_{|G|}\}$ and apply CP as before. This is useful if, for example, we require a set that guarantees us lower error rates on riskier labels (such as `cancer`) whilst sacrificing some accuracy on less risky labels (such as `headache`).

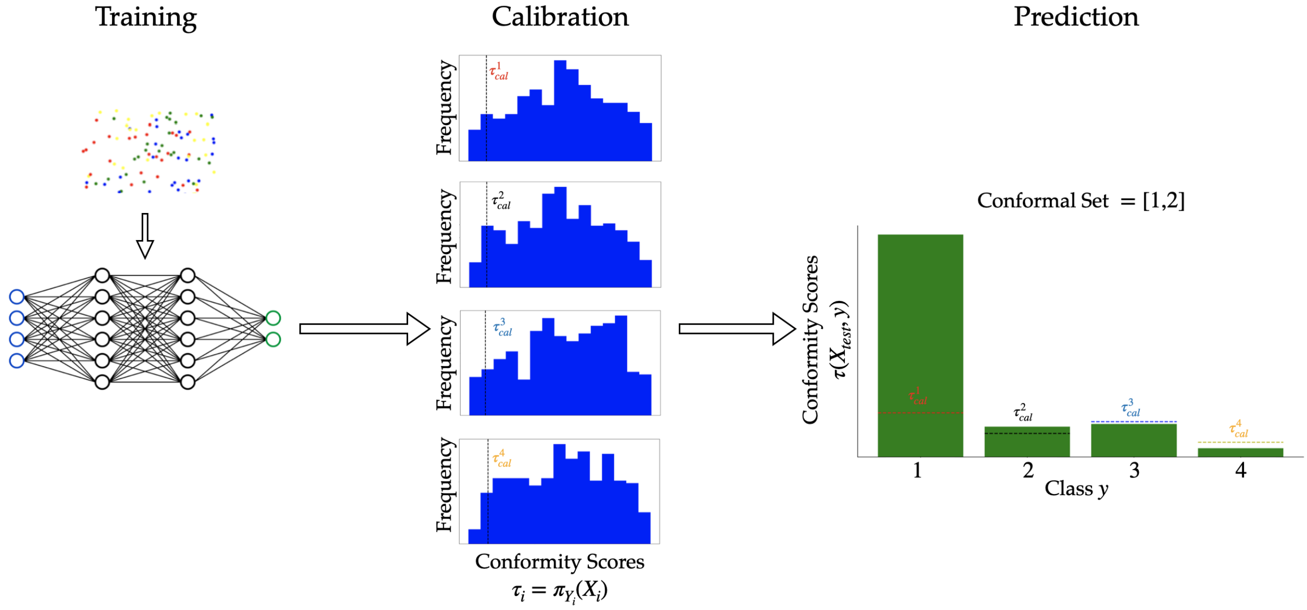


Figure 2.3: Pipeline for Performing Label Specific Mondrian ICP on any Dataset

2.2.5 Comparison Between Different Conformal Predictors

Remarkably, CP gives prediction sets that are guaranteed to satisfy Equation 2.5, no matter what (possibly incorrect) model is used, what the (unknown) distribution of the data is [2], or what calibration algorithm is used. The only requirement is that the samples be exchangeable.

Definition 2.4 (Exchangeability)**.** A set of samples $\{z_1 = (X_1, Y_1), ..., z_{n+1} = (X_{n+1}, Y_{n+1})\}$ is exchangeable when any permutation of the samples is equally likely. That is, $p(z_1 = (X_1, Y_1), ..., z_{n+1} = (X_{n+1}, Y_{n+1})) = p(z_1 = (X_{\rho(1)}, Y_{\rho(1)}), ..., z_{n+1} = (X_{\rho(n+1)}, Y_{\rho(n+1)}))$ for any permutation $\rho(.)$. This is a weaker condition than having i.i.d samples.

While different conformal predictors may satisfy the size and adaptiveness desiderata to varying extents, the main aim of contemporary CP literature has been to minimize the *predictive inefficiency* - which [11] define as the size of the predictive set - whilst maintaining the desired coverage guarantees. In this report, as we explore the utility of prediction sets in human-AI teams and develop algorithms to construct more *useful* sets, we will make use of the following ICP schemes, applied both marginally or in a Mondrian sense:

- **Least Ambiguous Classifiers (LAC)**: [34] introduce a framework for multi-class set-valued classification, where the classifiers guarantee user-defined levels of coverage while minimizing the predictive inefficiency. All that is required is an estimator of the conditional distribution $P_{Y|X}$ which is provided by any classification model's softmax function. The conformity score is defined as

$$\tau(X, y) = \pi_y(X) \tag{2.20}$$

and we apply the normal ICP algorithm as above.

- **Adaptive Prediction Sets (APS)**: [33] define a slightly more involved conformity score function. Denoting by $\pi_{(1)}(x) \geq \pi_{(2)}(x) \geq ...\pi_{(K)}(x)$ the order statistics of the softmax probabilities $\pi_y(x)$ (where the k^{th} order statistic is the k^{th} largest probability value), they define the generalized conditional quantile function:

$$L(X, \pi, \gamma) = \min\{k \in \{1, ...K\} : \sum_{i=1}^{k} \pi_{(i)}(x) \geq \gamma\} \tag{2.21}$$

and its generalized inverse

$$S(x, u, \gamma) = \begin{cases} \text{`y' indices of the } L(x, \pi, \gamma) - 1 \text{ largest } \pi_y(x), & \text{if } u \leq V(x, \pi, \gamma) \\ \text{`y' indices of the } L(x, \pi, \gamma) \text{ largest } \pi_y(x), & \text{otherwise} \end{cases}$$

where

$$V(x, \pi, \gamma) = \frac{1}{\pi_{(L(x,\pi,\gamma))}(x)} \left[\sum_{k=1}^{L(x,\pi,\gamma)} \pi_{(k)}(x) - \gamma \right] \tag{2.22}$$

and $u \sim \mathcal{U}[0, 1]$ is the uniform random variable. The conformity score is now stochastic and is defined as:

$$\tau(X, y, u) = -\min\{\gamma \in [0, 1] : y \in S(x, u, \pi, \gamma)\} \tag{2.23}$$

In the absence of stochasticity (say if $u = 1$), this conformity score would be equivalent to the sum of softmax probabilities of all labels that are more probable than y up to and including y. [33] show that this choice of conformity score leads to predictive sets that better approximate conditional coverage (where we condition on a small region around X) and provide nearly tight marginal coverage.

- **Regularised Adaptive Prediction Sets (RAPS)**: [4] propose an extension of the APS procedure by introducing regularisation terms that serve to penalise large set sizes. The intuition is that the tail softmax probabilities produced by a model are often noisy and inaccurate - we would be better off not including the corresponding labels in our predictive sets as far as possible. To temper their influence, [4] define the following conformity score function:

$$\tau(X, y, u) = -\left[\sum_{y'=1}^{K} \pi'_y(x) \mathbb{I}_{\pi'_y(x) > \pi_y(x)} + \pi_y(x) \cdot u + \lambda[o_y(x) - k_{reg}]^+ \right] \tag{2.24}$$

where $o_y(x) = |y' : \pi'_y(x) > \pi_y(x)|$ is the ranking of label y according to order statistics. For a more probable label y, the conformity score will be quite large, making it more likely that it is included in the predictive set. The randomisation term $u \sim \mathcal{U}[0, 1]$ handles the fact that the conformity score jumps discretely with the inclusion of each y and is similar to the randomization in Equation 2.23. This is of little practical importance - as the predictive set output by a randomized procedure will differ from that of a non-randomized procedure by at most one element - but [4] show that this is needed to maintain exact $1 - \alpha$ coverage. Lastly, the regularization promotes small set sizes. For a less probable label y, the ranking of y will be larger as it is further down the ordered list of classes. Thus, the term $\lambda[o_y(x) - k_{reg}]^+$ makes y require a higher value of τ before it is included in the set.

Note that if the conformity scores are almost surely distinct (e.g. if they take values in $\mathbb{R}$), [4] show that, given a calibration dataset $D_{cal} = \{(X_i, Y_i)\}_{i=1}^{n}$, we can provide the guarantee:

$$1 - \alpha \leq P(Y_{test} \in \Gamma(X_{test})) \leq 1 - \alpha + \frac{1}{n+1} \tag{2.25}$$

for any unseen example X_{test} sampled from the same distribution as the examples in the calibration dataset. This is useful because, given a sufficiently large calibration dataset, we can guarantee almost tight marginal coverage.

2.3 Risk Controlling Predictive Sets (RCPS)

2.3.1 Introduction

So far, we have introduced predictive sets that guarantee user-defined control of the one type of risk - the false negative rate. However, in many situations, we may want to control for a broader class of risks. Furthermore, we may seek to limit the violation of this risk, not just in expectation, but with a predefined probability. This is not possible with CP, whose guarantees only hold in expectation. To address these points, [9] develop a procedure called Risk Controlling Predictive Sets (RCPS) that is an extension of CP. While both the RCPS and CP procedure are similar in the sense that they *involve learning thresholds for including labels in a set*, RCPS achieves a broader notion of error control - namely guaranteeing control of any user-defined risk function.

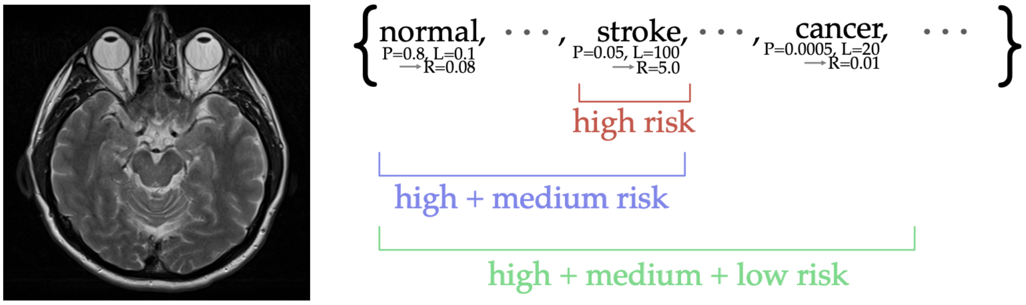


Figure 2.4: An example of the RCPS procedure applied on a medical diagnostics example. (from [9]). P = estimated class probability, L = loss associated with the corresponding class. R = risk - defined as the loss times probability. The red, blue, and green brackets represent possible sets of labels the procedure can output.

Figure 2.4 [9] illustrates the utility of controlling for a given risk. A notion of consequence is encoded by assigning the most severe loss to `stroke` ($L = 100$) and the least severe loss to `normal` ($L = 0.1$). Depending on the level of risk required, one can form sets that are low, medium, or high risk as those denoted by the red, blue, and green brackets. This procedure guarantees that the average loss or risk on future data will be less than γ with probability at least $1 - \delta$ where γ and δ are chosen by the user. *That is, there is at most a δ probability that we encounter a 'bad' set of datapoints where risk control is violated. [2].*

2.3.2 The Procedure

Formally, given a set valued prediction $\Gamma_\lambda(X) : \mathcal{X} \to 2^{\mathcal{Y}}$ and a loss function $L(Y, \Gamma_\lambda(X)) : \mathcal{Y} \times 2^{\mathcal{Y}} \to \mathbb{R}$, we control for the risk $R(\lambda) = \mathbb{E}[L(Y, \Gamma_\lambda(X))]$ such that:

$$P(R(\lambda) \leq \gamma) \geq 1 - \delta \tag{2.26}$$

Here, λ is a threshold parameter that determines the labels present in the set and satisfies the property

$$\lambda_1 < \lambda_2 \Rightarrow \Gamma_{\lambda_1}(X) \subset \Gamma_{\lambda_2}(X) \tag{2.27}$$

In order to construct such a set, [9] show that we need access to a pointwise upper confidence bound (UCB) for the risk function, i.e. we need an $\hat{R}^+(\lambda)$ such that

$$P(R(\lambda) \leq \hat{R}^+(\lambda)) \geq 1 - \delta \tag{2.28}$$

To obtain the UCB, we can use a *concentration inequality.*

Definition 2.5 (Concentration Inequality)**.** A concentration inequality bounds how much a random variable U deviates from some value, typically, its expected value $\mathbb{E}[U]$. In this report, we use concentration inequalities of the form:

$$P(U - \mathbb{E}[U] \geq \gamma) \leq \delta \tag{2.29}$$

where γ and δ are specified by the user.

Now, consider the empirical risk:

$$R(\lambda) \approx \hat{R}(\lambda) = \frac{1}{N}\sum_{i=1}^{N} L(Y_i, \Gamma_\lambda(X_i)) \tag{2.30}$$

One possible UCB of this risk can be obtained using Hoeffding's concentration inequality, which states:

$$P(\hat{R}(\lambda) - R(\lambda) \leq -x) \leq \exp(-2Nx^2) \tag{2.31}$$

If we set $\delta = \exp(-2Nx^2)$, we get

$$x = \sqrt{\frac{1}{2N}\log\frac{1}{\delta}} \tag{2.32}$$

and the corresponding UCB

$$\hat{R}^+(\lambda) = \hat{R}(\lambda) + \sqrt{\frac{1}{2N}\log\frac{1}{\delta}} \tag{2.33}$$

We obtain an (γ, δ) RCPS by tuning λ such that:

$$\hat{\lambda} = \inf\{\lambda : \hat{R}^+(\lambda) \leq \gamma\} = \inf\{\lambda : \hat{R}(\lambda) + \sqrt{\frac{1}{2N}\log\frac{1}{\delta}} \leq \gamma\} \tag{2.34}$$

which in turn gives us the guarantee in Equation 2.26. This is illustrated in Figure 2.5.

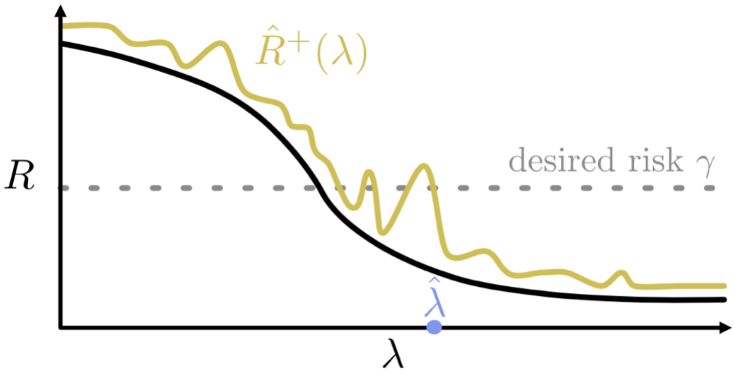


Figure 2.5: Upper Confidence Bound Calibration for the RCPS procedure (from [9]). The smallest parameter $\hat{\lambda}$ (a threshold that determines the labels added to the set) that ensures that the $1 - \delta$ UCB of the empirical risk is less than γ is selected

While we can use RCPS for any prediction problem, including classification with single, multiple, or hierarchical labels, regression, and object segmentation, we focus on multi-class problems in this report. Also note that we need not restrict ourselves to using only Hoeffding's concentration inequality - indeed [9] employ tighter concentration bounds to control the risk depending on the properties of the risk function. This has the advantage that we can find a lower optimal value of λ to control the risk and by the monotonicity property in Equation 2.27, this gives smaller set sizes. In this report, we employ the Central Limit Theorem concentration bound found in [9], which is of the form:

$$\hat{R}^{+}(\lambda) = \hat{R}(\lambda) + \Phi^{-1}(1-\delta)\frac{\hat{\sigma}(\lambda)}{\sqrt{N}} \tag{2.35}$$

where $\hat{\sigma}(\lambda)$ is the empirical variance of $L(Y, \Gamma_\lambda(X))$ on the calibration dataset and Φ^{-1} is the inverse Gaussian CDF. While this is technically valid asymptotically, we found that a) it gave reasonable coverage coverage close to the target level for the risks controlled in our experiments and b) it was tighter than the Hoeffding bound, giving smaller set sizes for the same risk control.

2.3.3 RCPS in Multi-Class Classification

For multi-class problems, we have a single, correct label Y whose cost of miscoverage may vary. That is,

$$L(Y, \Gamma_\lambda(X)) = L_y \mathbb{I}_{Y \notin \Gamma_\lambda(X)} \tag{2.36}$$

and we want to control $R(\lambda) = \mathbb{E}[L(Y, \Gamma_\lambda(X))]$. We can construct sets

$$\Gamma_\lambda(X) = \{y : \pi_y(x) > -\lambda\} \tag{2.37}$$

or more generally

$$\Gamma_\lambda(X) = \{y : \tau(x, y) > -\lambda\} \tag{2.38}$$

and tune λ (e.g. through a grid search scheme) to find the optimal value as in Equation 2.34. In a sense, RCPS in the multi-class setting is reminiscent of the Mondrian CP paradigm in that we can allow customized control of risky labels, i.e. we can assign low error tolerances in the former and set L_y to be large in the latter.

However, RCPS is a generalization of pure CP in that if $L_y = 1 \ \forall y$, then we recover the same risk controlled by CP, i.e. $R(\lambda) = P(Y \notin \Gamma_\lambda(X))$. However, we achieve Probably Approximately Correct (PAC) control of this risk (i.e. we control for the expected false negative rate on future data with probability $1-\delta$), whereas in CP the false negative rate on future data will be constrained to the desired level *on average*.

2.3.4 Evaluation of Coverage

In order to evaluate coverage on RCPS, we randomly split the validation set into calibration and test sets $N = 1000$ times. For each trial, we first perform calibration and then determine the expected risk on the validation set. As one would expect, this risk would exceed the pre-determined level γ in approximately δN trials[3].

[3] If we were to perform the equivalent procedure for CP, we would notice a variation in coverage but the average coverage would be centered around $1-\alpha$ (e.g. see Figure 4.9)

3 Background: A Policy for Learning to Defer to an Expert

3.1 Introduction

Machine learning models are increasingly being used to complement human-decision making in many real world applications such as healthcare and criminal justice [44]. In this section, we explore a simple method that can be used to control the level of automation offered to the model. This paradigm is referred to as *predictive triage* or *learning to defer* [28]. While machine learning models have surpassed or matched human-level performance in many supervised learning tasks in image classification [39], object detection [31], and medical image diagnostics [42], they are still less accurate on some instances the human may find easy (e.g. see Figure 4.8). The main promise is that, by working together, human experts and predictive models are likely to achieve a considerably better performance than each of them would achieve on their own. In the next section, we will employ this paradigm in order to provide smaller and more useful sets to human-AI teams.

3.2 Rejection Learning

The motivation for a loss function taking into account expert predictions stems from the idea of rejection learning, i.e. given an expert $h(x, z)$, we learn a classifier $m_\theta(x) \in \mathcal{Y}$ and a rejector $r(x) \in \{0, 1\}$ with the following loss function:

$$\mathcal{L}_{0-1}(h, r, x, y, m) = \mathbb{E}[l(x, y, m_\theta(x))\mathbb{I}_{r(x)=0} + l_{exp}(x, y, h(x, z))\mathbb{I}_{r(x)=1}] \tag{3.1}$$

That is, we incur a cost whenever a) the model defers to the expert and they make a mistake and b) when the model doesn't defer and makes a mistake. This is further illustrated in Figure 3.1.

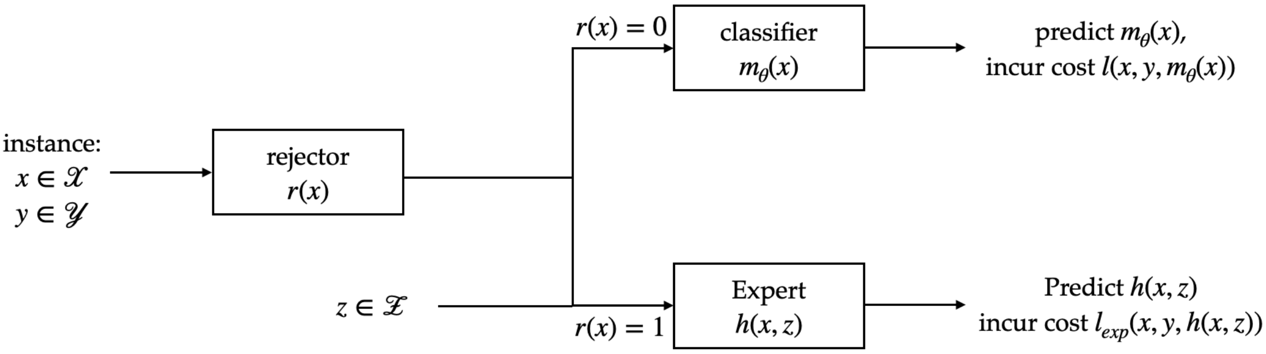


Figure 3.1: The expert deferral system (from [26]) - the rejector $r(x)$ allocates predictive responsibility between the classifier $m_\theta(x)$ and the expert $h(x, z)$, each with their own cost of misclassification. In this very general setup, the expert can have access to prior knowledge z that they use to faciltate decision making. This setup also allows for cost sensitive learning, i.e. where the cost of making a mistake on some labels may be greater than others

This utility of this method becomes apparent in situations where there are constraints on the classifier (e.g. linear decision boundaries), making it ill equipped to handle more complex decision spaces. In a binary classification example, if training examples are linearly separable in one region of the input space but inseparable in another, we may want to perform linear classification only on the separable region and let the expert generate their own non-linear decision boundaries. By using rejection learning, we let the rejector learn regions of the input space where deferral can resolve class ambiguity.

While numerous works have explored rejection learning from a theoretical perspective [27, 14, 19], we focus on a more practically focused scheme proposed by [26] for our experiments in this thesis.

3.3 Surrogate Loss Function for Learning to Defer

Instead of training rejector and classifier networks separately as in [5], a more modular, computationally inexpensive, and theoretically sound approach followed by [26] is to construct a joint system that acts upon an augmented label space $\mathcal{Y} \cup \perp$, where $\perp$ refers to the action of deferral. This is illustrated in Figure 3.2. We cannot use the loss function in Equation 3.1 to train this system as it is not convex and is computationally infeasible to optimize with gradient based methods, which require a differentiable loss. To overcome this, [26] propose a convex loss function of the following form:

$$\mathcal{L}_{\text{CE}}(h, r, x, y, m) = -(\alpha \mathbb{I}_{m=y} + \mathbb{I}_{m \neq y} \log \left(\frac{\exp(g_y(x))}{\sum_{y' \in \mathcal{Y} \cup \perp} \exp(g'_y(x))} \right)) - \mathbb{I}_{m=y} \log \left(\frac{\exp(g_\perp(x))}{\sum_{y' \in \mathcal{Y} \cup \perp} \exp(g'_y(x))} \right) \tag{3.2}$$

where $g_y(x)$ is the output of the model m_θ corresponding to label y before softmax transformation is applied and $\perp$ is the deferral class. This loss function can be viewed as a convex surrogate to the loss in Equation 3.1 when the costs $l(x, y, h(x))$ and $l_{exp}(x, y, h(x, z))$ are indicator functions - indeed [26] show that such a loss upper bounds Equation 3.1. Intuitively:

- The loss in Equation 3.2 allows for deferral in instances where the expert is correct. This capacity to defer is also controlled by the α parameter, which also allows the model to try and learn the target label for itself.
- The loss enables the model to learn to predict the target label in instances where the expert is wrong.

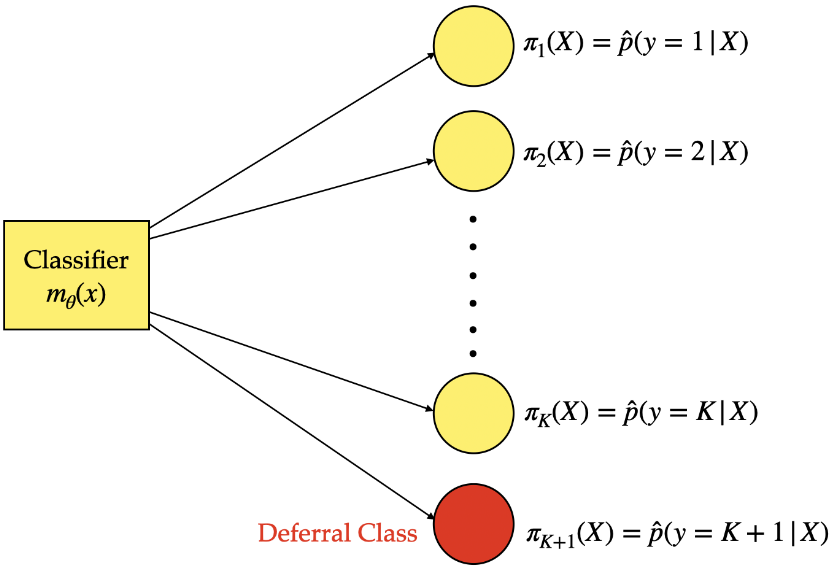

$\pi_1(X) = \hat{p}(y = 1 \mid X)$
$\pi_2(X) = \hat{p}(y = 2 \mid X)$
Classifier
$m_\theta(x)$
$\pi_K(X) = \hat{p}(y = K \mid X)$
Deferral Class
$\pi_{K+1}(X) = \hat{p}(y = K + 1 \mid X)$

Figure 3.2: The classfier trained on the loss function in Equation 3.2 casts softmax probabilities over an augmented label space wherein an additional deferral class is added

4 Generating Relevant Sets

We now design techniques for generating relevant sets in the context of Human-AI teams, using ideas developed in the previous sections. We define a *relevant* set below and argue that this should be an additional desideratum that a useful set-valued prediction should satisfy in alongside desiderata outlined in Section 3:

Definition 4.1 (Set Relevancy)**.**

- The model should provide useful predictive sets that provide actionable uncertainty quantification, however, different scenarios entail different notions of utility, so this is a more subjective desiderata. One measure of utility is the size of the set, as smaller set sizes are obviously more useful.

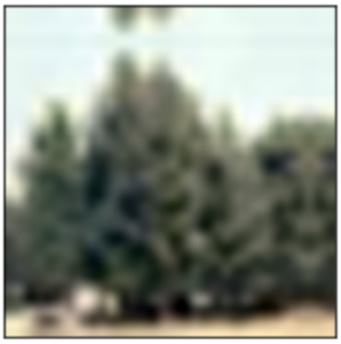


Figure 4.1: A sample image from the CIFAR-100 dataset. We argue that a predictive set such as {`pine tree, willow tree, forest`} can be perceived as being more useful than a predictive set such as {`pine tree, house, sky`}, even if both sets provide the same coverage guarantees and have the same size.

  However, consider the image in Figure 4.1 and the aforementioned predictive sets in the caption. We argue that another measure of set utility is how similar the labels in a predictive set are, where a similarity measure can be defined by a human expert. Thus, with appropriate human intervention, we can model the precise relationships between labels in order to shape predictive sets with certain kinds of labels in them.

- In general, we want the predictive set to be shaped by different human preferences and competencies. This means that it should not only be amenable to different notions of utility of an expert, but also consider the fact that an expert may have orthogonal competencies relative to the model. For example, a doctor specialised in oncology may not require confident sets for detecting `brain tumours` from MRI data (or may not require model predictions at all!), but may require highly confident, small sets that can predict `brain strokes`.

In this report, we consider the problem of generating maximally relevant sets (either CP or RCPS sets) when user specified parameters such as the error tolerance α, risk level γ, loss function $L(Y, \Gamma(X))$, and concentration probability δ are held constant. *All the work presented in this section henceforth has been accepted at IJCAI'22 and can be found at [6].*

4.1 Are Prediction Sets Better for Human-AI Teams than Top-1 Predictions?

Before designing relevant sets, we perform a preliminary investigation in order to establish the value of CP set predictions. Our main aims for this section are the following:

- Demonstrate that Human-AI teams benefit from principled, calibrated set valued predictions sets compared to Top-1 predictions.

- Show that Human-AI teams cannot expect to benefit from any set valued prediction - these predictions must capture the inherent uncertainty present in the model in some sense. This also relates to the relevancy desiderata outlined above - for example, a set containing random labels + the Top-1 label should almost certainly be less useful than a CP set.

4.1.1 Methodology

For our experiments, we focus on the Regularised Adaptive Prediction Sets (RAPS) scheme [4].

- We first train a WideResNet [43] classifier on the CIFAR-100 [22] dataset for 10 epochs, employing the training scheme found in [26]. We then recruit 30 participants on Prolific, paying them at a rate of £10 per hour prorated, and divide them into 2 equal groups. The first group is shown 18 images from the CIFAR-100 dataset alongside the model's most probable prediction (Top-1). The second group is shown the same images but alongside a RAPS prediction set with error rate $\alpha = 0.1$.

- To understand the effect of set valued predictions on examples of varying difficulty, we divide the CIFAR-100 test dataset into 3 difficulty quantiles, where difficulty is defined as the entropy of the model predictive distribution. We select 6 images from each difficulty quantile. For each quantile, we show 2 images whose Top-1 prediction is incorrect but whose RAPS set contains the true label. This is consistent with the accuracy of the model ($\approx 65\%$) and lets us determine the effect of set valued predictions on examples on which the model is *almost* (i.e. Top-K) correct.

- Given these model predictions for each image, we ask participants in both groups to predict the correct class, rate their confidence in their predictions, and rate how useful they found the model predictions on that example. At the end of the survey, we ask participants to rate their overall trust in the model's predictions. All ratings are on a scale from $1 - 10$. We employ preliminary attention checks by first asking them to classify 3 easy examples, rejecting any participants who classify these examples incorrectly. We recruit participants until we have 15 accepted participants for each group.

- For each accepted participant, we calculate their average utility and confidence scores across all examples (note the trust scores are not example specific, so we have only one value for each participant) to obtain N scores for each metric.

Given N scores from the Top-1 and RAPS group (with corresponding unknown means μ_1 and μ_2), we test the null hypothesis H_0 where

$$H_0 : \mu_2 = \mu_1 \tag{4.1}$$

$$H_1 : \mu_2 \neq \mu_1 \tag{4.2}$$

We use the two-sample t-test and report the two-tailed p-values. In this method, given a metric $\in \{\text{Trust}, \text{Accuracy}, \text{Utility}, \text{Confidence}\}$, we first compute the empirical means and variances of the two groups for that metric:

$$\hat{\mu}_i = \frac{1}{N}\sum_{j=1}^{N} r_{ij} \tag{4.3}$$

$$\hat{\sigma}_i^2 = \frac{1}{N}\sum_{j=1}^{N} (r_{ij}^2 - \hat{\mu}_i)^2 \tag{4.4}$$

where r_{ij} is the response of the j^{th} participant for the i^{th} group. We then calculate the pooled standard deviation:

$$\sigma_p^2 = \frac{(N-1)(\sigma_1^2 + \sigma_2^2)}{2N-2} = \frac{\sigma_1^2 + \sigma_2^2}{2} \tag{4.5}$$

The test statistic is then calculated as

$$\hat{t} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sigma_p \sqrt{\frac{2}{N}}} \tag{4.6}$$

and the corresponding two-tailed p value is $2(1 - \Phi_{t,2N-1}(\hat{t}))$, where $\Phi_{t,2N-1}(\hat{t})$ is the CDF of the Student-T distribution with $2N - 1$ degrees of freedom. Lastly, we also calculate the effect size d of the observation. This measures the magnitude of the experimental effect, i.e. the strength of the relationship between 2 variables.

$$d = \frac{|\hat{\mu}_1 - \hat{\mu}_2|}{\sigma_p} \tag{4.7}$$

Note that, while we report two-tailed p-values for all sections (which are generally recommended, as they allow for testing of significance in either direction), we can obtain one-sided p-values $1 - \Phi_{t,2N-1}(\hat{t})$ for the alternate hypotheses $H_1 : \mu_2 > \mu_1$ or $H_1 : \mu_2 < \mu_1$. However, in our case, this makes no difference to our resulting conclusions as we obtain sufficiently small p-values for all statistically significant effects, with no borderline cases.

4.1.2 Human Subject Evaluation

Metric	Top-1	RAPS	p value	Effect Size
Accuracy	0.76 ± 0.05	0.76 ± 0.05	0.999	0.000
Reported Utility	5.43 ± 0.69	6.94 ± 0.69	**0.003**	1.160
Reported Confidence	7.21 ± 0.55	7.88 ± 0.29	0.082	0.674
Reported Trust in Model	5.87 ± 0.81	8.00 ± 0.69	< **0.001**	1.487

Table 4.1: Top-1 vs RAPS: All Examples

Metric	Top-1	RAPS	p value	Effect Size
Accuracy	0.90 ± 0.05	0.87 ± 0.07	0.486	0.273
Reported Utility	6.07 ± 0.94	6.35 ± 1.00	0.438	0.195
Reported Confidence	7.88 ± 0.65	8.82 ± 0.31	**0.013**	1.019

Table 4.2: Top-1 vs RAPS: Lowest Difficulty Quantile

Metric	Top-1	RAPS	p value	Effect Size
Accuracy	0.64 ± 0.07	0.66 ± 0.10	0.828	0.068
Reported Utility	5.30 ± 0.75	7.28 ± 0.69	**0.001**	1.432
Reported Confidence	6.64 ± 0.64	6.96 ± 0.78	0.4888	0.280

Table 4.3: Top-1 vs RAPS: Highest Difficulty Quantile

From Table 4.4, we see that Top-1 predictions result in statistically significant lower levels of trust ($p < 0.001$) and perceived utility ($p = 0.003$) compared to RAPS, with large effect sizes. However, both schemes result in similar accuracy and confidence in predictions for all levels of difficulty. We also see that users find Top-1 and RAPS predictions equally useful for easy examples (Table 4.5). This makes sense because in such cases, the predictive set will be small and therefore comparable to a Top-1 prediction. However, users are more confident about their answers when they observe RAPS predictions. On the other hand, from Table 4.6, RAPS sets are perceived to be much more useful on hard examples, where Top-1 predictions will often be wrong.

In order to eliminate any differential bias participants may have when shown a set vs a single prediction, we also compared the performance of the RAPS scheme with a variant called Top-1 + Random, using the same methodology outlined in Section 5.1.1 with the same images shown (to a new participant group consisting of 15 people). Here, we construct a predictive set consisting of the Top-1 prediction of the model + 3 random classes chosen uniformly from the label space of CIFAR-100. We chose 3 random classes in order to approximately control for the set size of both schemes - the average size of a RAPS set for CIFAR-100 for $\alpha = 0.1$ is 3.75.

Metric	Top-1 + Random	RAPS	p value	Effect Size
Accuracy	0.72 ± 0.05	0.76 ± 0.05	0.427	0.338
Reported Utility	5.01 ± 0.65	6.94 ± 0.69	**0.003**	1.432
Reported Confidence	7.29 ± 0.47	7.88 ± 0.29	0.082	0.098
Reported Trust in Model	5.73 ± 1.07	8.00 ± 0.69	**0.008**	1.316

Table 4.4: Top-1 + Random vs RAPS: All Examples

Metric	Top-1 + Random	RAPS	p value	Effect Size
Accuracy	0.81 ± 0.07	0.87 ± 0.07	0.313	0.388
Reported Utility	5.48 ± 0.95	6.35 ± 1.00	0.204	0.501
Reported Confidence	8.74 ± 0.42	8.82 ± 0.31	0.743	0.125

Table 4.5: Top-1 + Random vs RAPS: Lowest Difficulty Quantile

Metric	Top-1 + Random	RAPS	p value	Effect Size
Accuracy	0.63 ± 0.14	0.66 ± 0.10	0.827	0.068
Reported Utility	5.08 ± 0.88	7.28 ± 0.69	$<$ **0.001**	1.453
Reported Confidence	6.46 ± 0.62	6.96 ± 0.78	0.270	0.425

Table 4.6: Top-1 + Random vs RAPS: Highest Difficulty Quantile

We observe similar results as the Top-1 vs RAPS experiments. For easy examples, there are no statistically significant differences in accuracy, utility, or confidence, implying that users express equal preference for either kind of predictive set. However, RAPS sets provide statistically significant increases in utility and trust compared to Top-1 + Random sets overall. This effect is particularly visible for difficult examples, where the effect size of reported utility is the highest out of the 3 tables presented. These results suggest that set valued predictions must be principled - the model cannot add value in a human-AI team if its set valued predictions don't capture inherent uncertainty.

4.2 Combining the Advantages of CP and Learning to Defer: D-CP

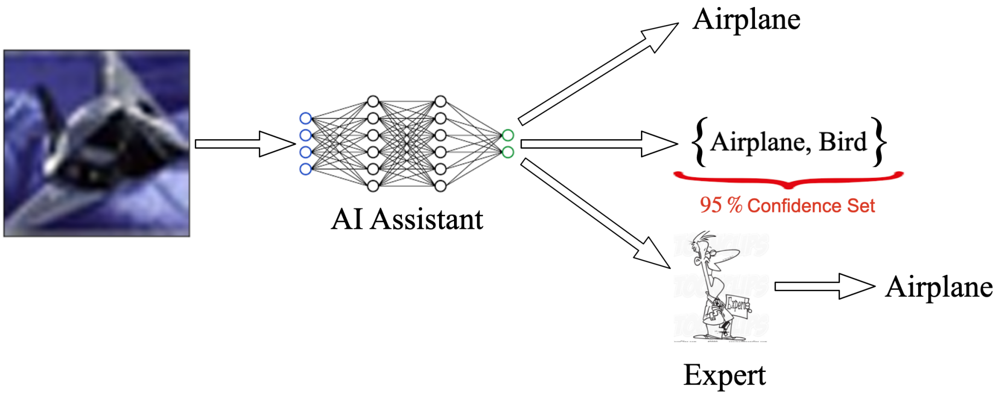


Figure 4.2: An AI assistant working alongside an expert can output one of three things: the most likely label, a set valued prediction with a predetermined error probability, or a deferral token indicating that the example should be labelled by the expert. The precise nature of the AI's prediction should be dependent on acquired knowledge of the expert's capabilities. Generally speaking, because the size of the predictive set is a reflection of the model's confidence, deferring examples on which an expert is more confident than the model would prevent an expert from using large, incoherent prediction sets.

4.2.1 The Problem with CP Sets

In our experiments above, we showed users examples where the set sizes on CIFAR-100 are small enough to be considered useful. However, this may not always be the case, especially on tasks with large label spaces. For instance, a standard WideResNet model trained on CIFAR-100 ($\approx$ 65% accuracy) with APS conformal prediction yields prediction sets with greater than 15 labels for over 20% of examples. One option to mitigate this issue is to defer examples with CP set sizes above a threshold to an expert. However, this provides no guarantee that the expert will be able to classify them with sufficient accuracy. Furthermore, we also lose the finite sample coverage guarantees provided by contemporary CP methods, i.e. we cannot ascertain that $P(Y_{test} \notin \Gamma(X_{test})) \leq \alpha$. This is because the calibration dataset for CP and the resulting test dataset (i.e. the non-deferred examples) are not identically distributed and hence not exchangeable (Definition 2.4).

4.2.2 Proposed Approach

Our scheme, described in Algorithm 1, is centered around two components: a deferral policy $r(x) : \mathcal{X} \to \{0, 1\}$ and a CP method. The deferral policy is based on our knowledge of the expert's strengths either acquired during training or a-priori. For example, if an expert is better at identifying brain tumors than our model, our policy should learn to defer those examples with high probability. Using this black box policy, we first prune our calibration dataset, removing all examples where our deferral policy recommends deferral. One could use any scheme in [26, 28, 41] to learn a deferral policy. While Algorithm 1 specifies a deferral policy as an input, for some deferral methods (such as [26]), the policy is trained alongside the model. In others, such as [28], the policy is applied post-hoc. In this report, we consider the former deferral policy: the D-CP algorithm for this is outlined in Algorithm 4 in the Appendix. After training a model and a suitable deferral policy, we perform conformal calibration on this pruned dataset of non-deferred examples. In this procedure, for any predictive set $\Gamma(X_{test}, \tau_{cal})$ for an example X_{test} we can guarantee that:

$$1 - \alpha \leq P(Y \in \Gamma(X_{test}, \tau_{cal}) | r(X_{test}) = 0) \tag{4.8}$$

where $r(X_{test}) = 1$ represents the action of deferral. From [4], when the conformity scores are known to be almost surely distinct and continuous, we can also guarantee:

$$P(Y \in \Gamma(X_{test}, \tau_{cal}) | r(X_{test}) = 0) \leq 1 - \alpha + \frac{1}{n+1} \tag{4.9}$$

where n is the size of the non-deferred calibration dataset. Because the deferral policy r probabilistically decides which unseen examples to defer, all non-deferred examples can be thought of as being generated from a data generating distribution $X \sim p(X|r(X) = 0)$ an an i.i.d manner. Any new test example X_{test} that is not deferred is therefore independently drawn from this distribution. Thus, $\{X_i\}_{i=1}^N \cup \{X_{test}\} \sim p(X_1, ..X_{test}|r(X_1), ..r(X_{test}) = 0)$ are exchangeable, thereby satisfying the coverage guarantee in Equation 4.8. **Note that while this section focuses on CP, we can also suitably define D-CP in the context of RCPS sets**. Here, the pipeline would be similar, except that the calibration step on non-deferred examples would follow the RCPS algorithm instead of CP.

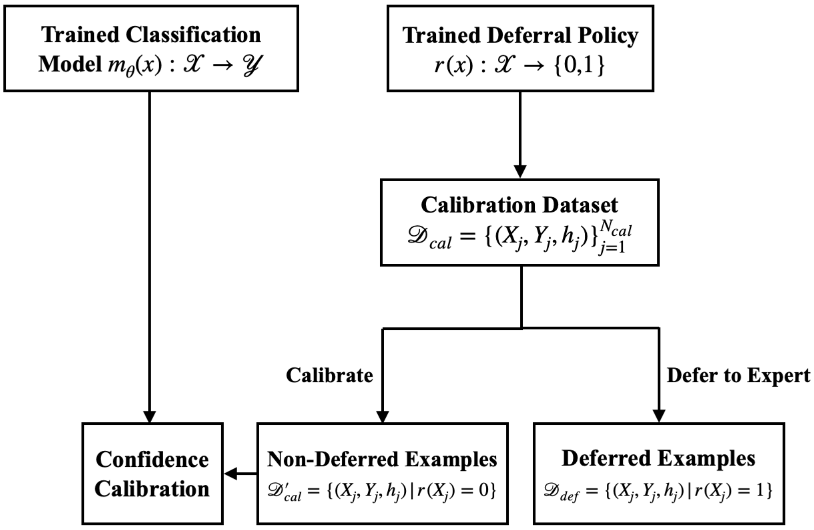

Trained Classification Model $m_\theta(x) : \mathcal{X} \rightarrow \mathcal{Y}$
Trained Deferral Policy $r(x) : \mathcal{X} \rightarrow \{0,1\}$
Calibration Dataset $\mathcal{D}_{cal} = \{(X_j, Y_j, h_j)\}_{j=1}^{N_{cal}}$
Calibrate
Defer to Expert
Confidence Calibration
Non-Deferred Examples $\mathcal{D}'_{cal} = \{(X_j, Y_j, h_j) | r(X_j) = 0\}$
Deferred Examples $\mathcal{D}_{def} = \{(X_j, Y_j, h_j) | r(X_j) = 1\}$

Figure 4.3: D-CP: Training and Calibration Phase

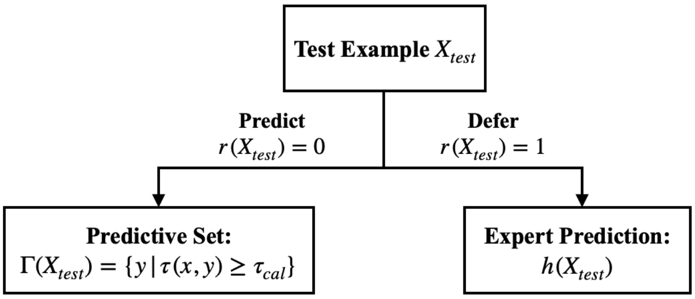


Figure 4.4: D-CP: Test Phase given a deferral policy $r(X)$

Algorithm 1 General D-CP

Input: Classifier $m_\theta(x) \in \mathcal{R}^{|\mathcal{Y}|}$, Deferral Policy $r_\omega(x) \in \{0, 1\}$, Training Set $\mathcal{D}$, Expert $h(x) \in \mathcal{Y}$, Calibration Set $\mathcal{D}_{\text{cal}}$, Validation Set $\mathcal{D}_{\text{val}}$, Test Example x_{test}, Conformity Score Function $\tau(X, y)$, Loss function $l(m_\theta(x), y, h(x))$

Parameter: Number of Epochs N, Learning Rate γ, Error Tolerance α

```
for $i \in \{1, ...N\}$ do
    for Batch $\mathcal{B} \in \mathcal{D}$ do
        $\theta = \theta - \gamma \mathbb{E}_{(x,y)\in\mathcal{B}}[\nabla_\theta l(m_\theta(x), y, h(x))]$
        $\omega = \omega - \gamma \mathbb{E}_{(x,y)\in\mathcal{B}}[\nabla_\omega l(r_\omega(x), y, h(x))]$        ▷ Train Model and Deferral Policy
$\mathcal{D}'_{\text{cal}} = \{(X, Y) | r_\omega(X) = 0, (X, Y) \in \mathcal{D}_{\text{cal}}\}$
$\tau_{cal} = \text{Quantile}(\alpha, \{\tau(X_i, Y_i) | (X_i, Y_i) \in \mathcal{D}'_{\text{cal}}\})$        ▷ Calibrate on Non-Deferred Examples
if $r(X_{test}) = 0$ then
    return $\Gamma(X_{test}, \tau_{cal}) = \{y | \tau(X_{test}, y) \geq \tau_{cal}\}$
else if $r(X_{test}) = 1$ then
    Defer to Expert $h(X_{test})$
```

- To show the utility of our scheme, a good deferral policy would guarantee that resulting predictive sets on non-deferred examples will contain fewer incorrect labels than before. We prove this formally in Theorem 1 in the Appendix. While this technically applies to conformity scores $\tau(x, y)$ that are monotonic with respect to softmax probabilities (such as LAC, see Equation 2.20), our subsequent experiments with other CP methods such as RAPS and APS suggest that our scheme generalises well across other classes of conformity score functions.

- We also prove in Theorem 2 in the Appendix a distribution free result that simultaneously guarantees control of the expected loss on deferred examples and the mis-coverage of the CP set on non-deferred examples with high probability for any deferral policy, model, and expert. This is empirically evaluated in Section 6.

4.2.3 Toy Example

One way to combine conformal prediction and deferral is to only perform CP on “easy” examples and defer the “hard” examples. An “easy” example would be one which the model is confident on and a “hard” example is the converse. This can lead to smaller sets. To demonstrate the intuition, we generate equiprobable synthetic data using a Mixture of Gaussians (MoG) model.

Definition 4.2 (Mixture of Gaussians (MoG))**.** Given K clusters, a MoG is a probabilility distribution s.t.

$$p(x) = \sum_{i=1}^{K} u_k \mathcal{N}(x; \mu_k, \Sigma_k) \tag{4.10}$$

where $\sum_{i=1}^{K} u_k = 1$ and $\mathcal{N}(x; \mu_k, \Sigma_k)$ is a Gaussian with mean μ_k and covariance matrix Σ_k.

We first generate 1000 training datapoints (not shown) sampled from $p(x) = \frac{1}{4}[\mathcal{N}(1, 1) + \mathcal{N}(1, -1) + \mathcal{N}(-1, 1) + \mathcal{N}(-1, -1)]$ and train a simple multilayer perceptron (MLP) model [17] to infer class memberships and the corresponding decision boundaries. Then, using a held out calibration set generated using the same procedure, we perform CP with error tolerance $\alpha = 5\%$ using the Least Ambiguous Classifiers (LAC) method [34]. Recall that we use the model softmax probabilities $\pi_y(X) = \tau(X, y)$ as conformity scores in this method.

- Figure 4.5 (left) shows a 1-D scatter plot of conformity scores assigned to ground truth labels in the toy calibration dataset. Figure 4.6 shows the resulting test datapoints colored according to their true classes with model decision boundaries overlaid. We see that points closer to the decision boundary have larger predictive set sizes, reflecting their inherent uncertainty.

- If we defer points with conformity scores in the bottom 15^{th} percentile (naive decision policy) as in Figure 4.5, the α threshold conformity score will increase. From Figure 4.5 (Right), for non-deferred examples, this increases the threshold for including labels in the set, resulting in more confident sets for the same error control.

- However, this naive deferral method, whilst ensuring small set sizes on the remaining examples, does not take into account the expertise of the expert involved. Furthermore, we assumed access to ground truth labels for test examples, which is not practical.

- We can engage the expert in a better manner and approximate the idea of the toy example by learning a *deferral policy* which incorporates estimates of expert ability as well as machine difficulty. This scheme makes an implicit assumption that the expert is a) either better than the model on average or b) not necessarily better than the model on average, but is proficient in classifying certain subgroups of examples. We make use of this assumption in Section 5 where we illustrate the risk control properties of D-CP. This is a reasonable assumption in many real world settings, as humans will not deploy a black box AI assistant when they themselves have no knowledge of the task at hand.

- Given these assumptions about an expert, our deferral policy is more likely to defer examples that a model is less confident on. Theorem 1 henceforth assures us of lower predictive set sizes.

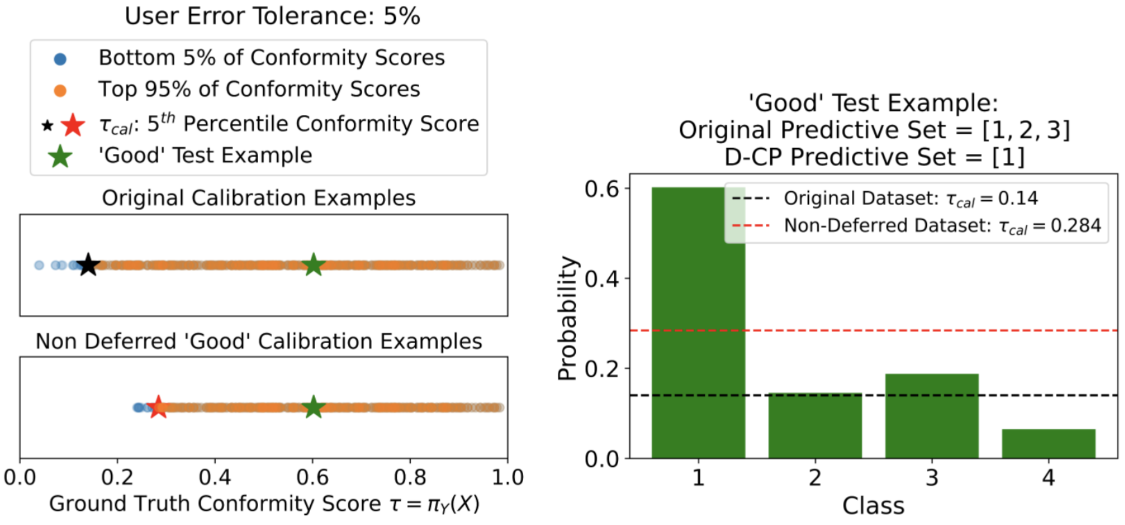


Figure 4.5: (Left): 1-D scatter plot of all ground truth conformity scores $\tau = \pi_{Y_i}(X_i)$ on a toy calibration dataset. We assume an oracle deferral policy that defers $\beta = 15\%$ of examples with the lowest τ. Both values of τ_{cal} maintain 95% coverage on their respective datasets. (Right): Class probabilities for the test **green starred example**. For the predictive set, we include all scores which are greater than the threshold τ_{cal}. Thus, the predictive set $\{1, 2, 3\}$ gives 95% coverage for the original dataset. On the non-deferred dataset, the set $\{1\}$ gives 95% coverage.

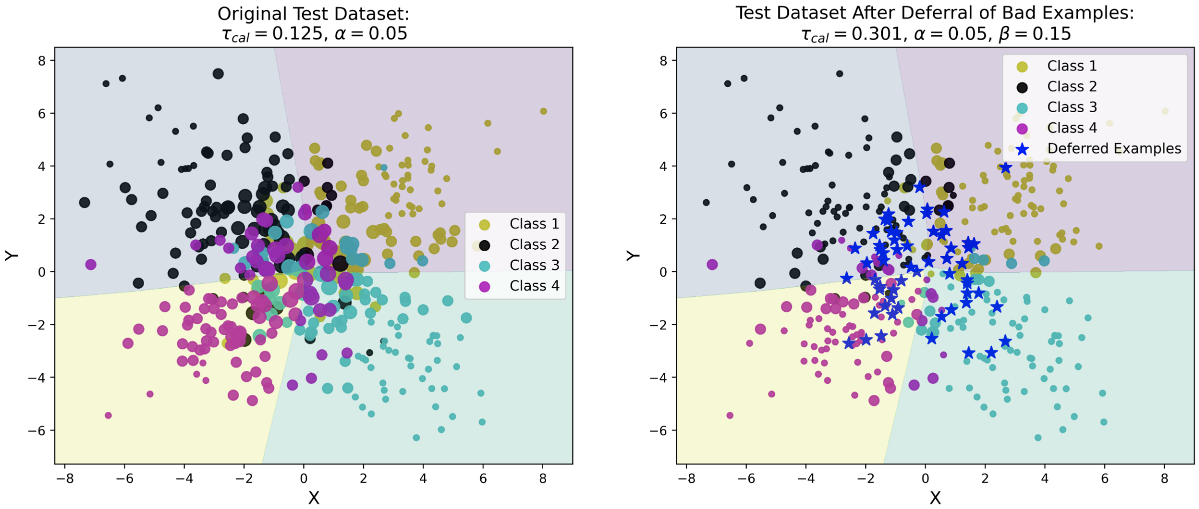


Figure 4.6: (Left): Toy dataset comprising of datapoints belonging to one of 4 classes along with overlaid model decision boundaries. **The size of the datapoints indicates their predictive set sizes**. (Right) We defer the $\beta = 0.15$ proportion of examples with the lowest ground truth conformity scores. Doing so increases the value of the 5^{th} percentile conformity score of the remaining examples in Figure 4.5, causing CP set sizes of examples to be smaller. Note that we have not changed the model in this process.

4.2.4 Experimental Evaluation

To validate our approach, we perform experiments with synthetic expert labels on the CIFAR-100 dataset and real expert labels on the CIFAR-10H [29] dataset. The latter is essentially the CIFAR-10 dataset except that there are additional annotations made by 50 humans on the validation set, forming a distribution over the label space that reflects human perceptual uncertainty. Because the CIFAR-10H dataset contains no expert labels on the training set, we employ the approach in [26] and train a binary classifier to predict examples where the expert is correct. We then provide synthetic $0-1$ expert labels $\mathbb{I}_{h(x)=y}$ or $\mathbb{I}_{h(x)\neq y}$ for examples in the training set according to whether the expert errs on them. Note that, in line with the assumption made in this paper, the experts chosen in this setting are, on average, better than the model trained. We consider 2 scenarios:

- We have access to a single expert's annotations. For CIFAR-100, we generate a synthetic expert with 70% accuracy. To motivate this choice, we ran a control study where we asked 20 participants to classify 15 randomly chosen CIFAR-100 examples. We found participants had average accuracy of 69% with a standard error of $\approx 2.5\%$. For the CIFAR-10H dataset, we randomly sample a label from the predictive distribution provided. This gave an accuracy of $\approx 95\%$.

- We have access to multiple expert annotations. This is an ensemble of the above experts, and the predicted class is chosen through majority voting for both datasets. For the CIFAR-100, we generate predictions from 5 synthetic experts.

Deferral Rate b	Classifier Accuracy	System Accuracy (Single Expert)	System Accuracy (Multiple Experts)	Predictive Set Size of Non-Deferred Examples		
				RAPS	APS	LAC
0	65.18 ± 0.30	65.18 ± 0.30	65.18 ± 0.30	3.75 ± 0.06	4.61 ± 0.08	3.26 ± 0.03
0.05	68.39 ± 0.31	68.04 ± 0.32	68.91 ± 0.33	3.22 ± 0.05	4.16 ± 0.06	2.48 ± 0.03
0.10	69.92 ± 0.24	69.95 ± 0.31	71.53 ± 0.35	2.81 ± 0.05	4.05 ± 0.06	2.13 ± 0.04
0.20	72.98 ± 0.30	72.25 ± 0.30	78.99 ± 0.40	2.36 ± 0.07	2.93 ± 0.10	2.07 ± 0.03

Deferral Rate b	Classifier Accuracy	System Accuracy (Single Expert)	System Accuracy (Multiple Experts)	Predictive Set Size of Non-Deferred Examples		
				RAPS	APS	LAC
0	82.02 ± 0.55	82.02 ± 0.55	82.02 ± 0.55	1.91 ± 0.03	2.83 ± 0.05	2.47 ± 0.12
0.05	84.41 ± 0.69	84.31 ± 0.65	84.64 ± 0.30	1.87 ± 0.08	2.76 ± 0.10	2.25 ± 0.08
0.10	86.12 ± 0.67	86.53 ± 0.68	88.12 ± 0.61	1.73 ± 0.08	2.56 ± 0.07	1.90 ± 0.15
0.20	88.97 ± 0.50	89.43 ± 0.64	91.46 ± 0.32	1.49 ± 0.06	2.13 ± 0.11	1.51 ± 0.09

Table 4.7: D-CP Predictive Set Size, Overall System Accuracy, and Classifier Accuracy on non-deferred examples on the CIFAR-100 (top) and CIFAR-10H (bottom) datasets for the deferral scheme in [26] and the 3 CP schemes ($\alpha = 0.1$, 5 Trials, 95% CI). Even in the low deferral rate regime, we not only obtain smaller set sizes across all CP schemes tested, but also benefit from increased Human-AI system accuracy and classifier accuracy on non-deferred examples (relative to the baseline where the deferral rate $b = 0$). While having multiple experts does not further improve the predictive set size for this deferral policy, we obtain improved system accuracy.

We train a WideResNet [43] classifier $m_\theta(x) : \mathcal{X} \to \mathcal{Y} \cup \perp$ with softmax probabilites $\pi_y(x)$ on CIFAR-10H and CIFAR-100 for 5 and 10 epochs respectively using the learning rate schedule in [26]. $\perp$ represents the action of deferral to an expert $h(x)$. We employ the following loss function:

$$\mathcal{L}_{\text{CE}}(h, x, y, m_\theta) = -(\mathcal{I}_{h(x)\neq y} + \alpha \mathcal{I}_{h(x)=y})\pi_y(x) \tag{4.11}$$

$$- \beta \mathcal{I}_{h(x)=y} \log \pi_\perp(x) \tag{4.12}$$

This is similar to the loss in Equation 3.2 except that we use the term $\beta \log \pi_\perp(x)$ instead of $\log \pi_\perp(x)$. We empirically observed that, by varying β, we can obtain fine grained control of the deferral rate, which was not the case when we varied α. Thus, for our experiments, we set $\alpha = 1$ and we vary β. The policy $r(x)$ is therefore:

$$r(x) = \begin{cases} 1 & \text{argmax}_{y \in \mathcal{Y} \cup \perp} \pi_y(x) = |\mathcal{Y} \cup \perp| \\ 0 & \text{otherwise} \end{cases}$$

To compute conformity scores for the RAPS, APS, and LAC schemes, we first renormalize the softmax probabilities for examples where $r(x) = 0$ (i.e. where we don't defer) using Bayes' rule:

$$\pi'_y(x) = p(y|x, r(x) = 0, \theta) = \frac{p(y \neq |\mathcal{Y} \cup \perp||x, y, \theta) p(y|x, \theta)}{p(y \neq |\mathcal{Y} \cup \perp||x, \theta)} \tag{4.13}$$

$$= \frac{p(y|x, \theta)}{p(y \neq |\mathcal{Y} \cup \perp||x, \theta)} \tag{4.14}$$

$$= \frac{\pi_y(x)}{1 - \pi_\perp(x)} \tag{4.15}$$

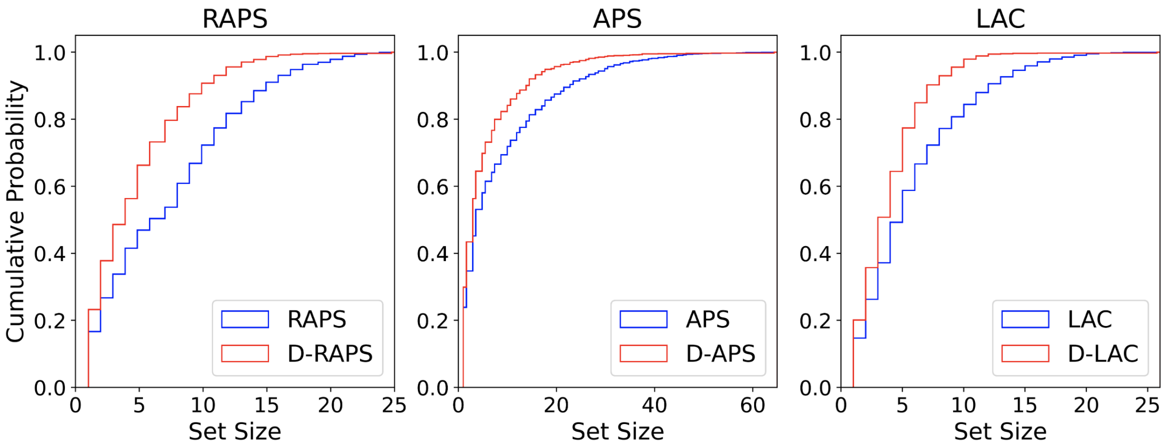
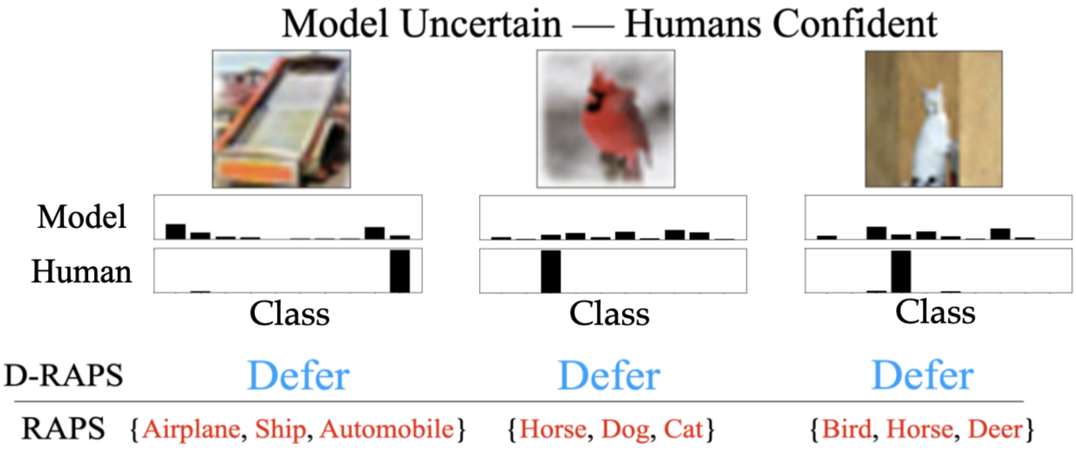


Figure 4.7: Cumulative CP and D-CP Set Size Distribution of Non-Deferred Examples in the CIFAR-100 dataset ($\alpha = 0.05$, deferral rate $b = 0.2$, Single Expert).

- In our experiments, we did not notice any statistically significant difference in accuracy of non-deferred examples or predictive set sizes when employing multiple experts as opposed to a singular expert, at least in the deferral rate regimes tested.

- Because we are performing experiments in the low deferral rate regime, it is likely that the deferral scheme defers similar examples to both expert types - examples the model is sure the expert(s) will get right. *Thus, in Table 4.7, the classifier accuracy on non-deferred examples and predictive set sizes are representative for both singular and multiple experts.* This is also the reason that we get only small, statistically insignificant improvements in system accuracy between single and multiple experts for small deferral rates.

- However, we benefit from increased system accuracy by using ensemble voting across multiple experts. In addition, from Table 4.7 and Figure 4.7, our scheme ensures smaller set sizes across all conformal methods and deferral rates tested. Increasing the deferral rate reduces the predictive set size. In Figure 4.8, the model and expert have a mutually beneficial relationship: the model provides smaller predictive sets on examples the expert is more uncertain on and defers examples it is less certain of than an expert.



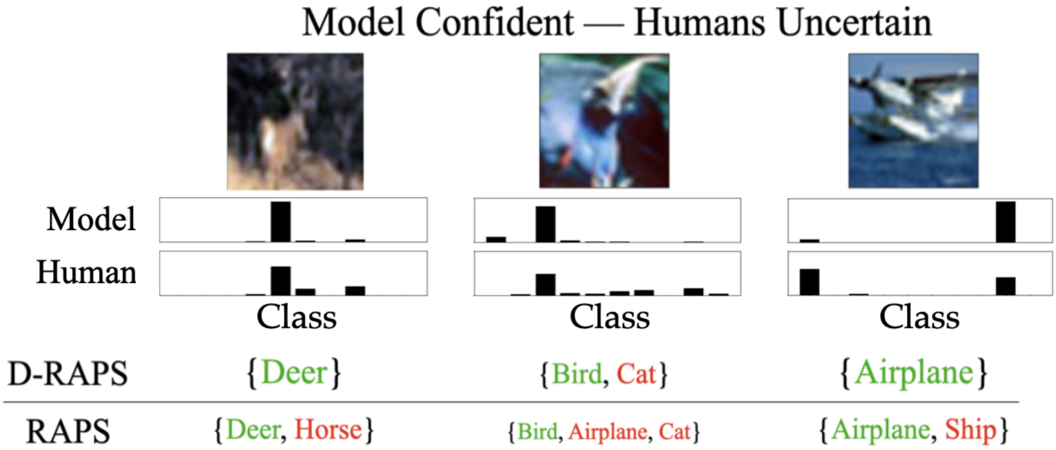


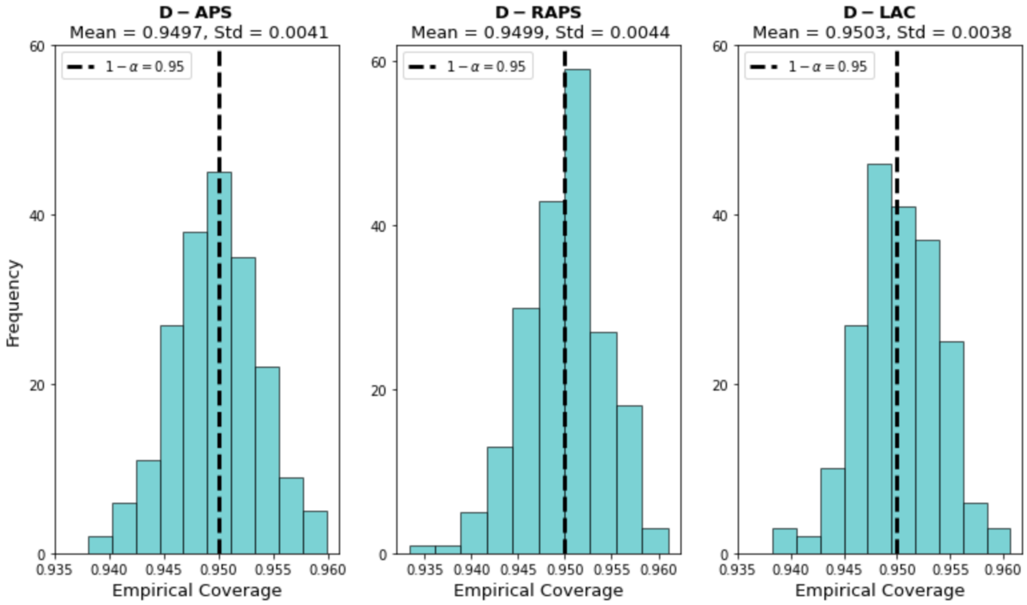
Figure 4.8: D-RAPS vs RAPS on some examples in the CIFAR-10H dataset ($\alpha = 0.05$, $b = 0.2$). Deferring on examples where experts are more confident than the model provides smaller sets on examples where the model is more confident than the expert. Thus, we leverage the strengths of both the model and the expert.

4.2.5 Coverage Guarantees and Statistical Efficiency of D-CP

Recall that for an Inductive Conformal Predictor (ICP), the coverage on the validation set is defined as:

$$C = \frac{1}{N_{val}} \sum_{i=1}^{N_{val}} \mathbb{I}_{Y_i \in \Gamma(X_i)} \tag{4.16}$$

While conformal prediction provides theoretical guarantees of the form in Equations 4.8 and 4.9, due to the finite number of samples and variations in test and training data distributions, ICP does not result in exact coverage in practice. [38] and [23] report that the coverage of conformal intervals is highly concentrated around $1 - \alpha$. Because D-CP ensures that samples in the calibration and validation sets remain exchangeable, we get similar coverage distributions for D-CP as we would for any CP method. This is illustrated in Figure 4.9, where we randomly split the test dataset into calibration and validation datasets 200 times and evaluated the coverage for each trial. As expected, the coverage is approximately $1 - \alpha$ on average.



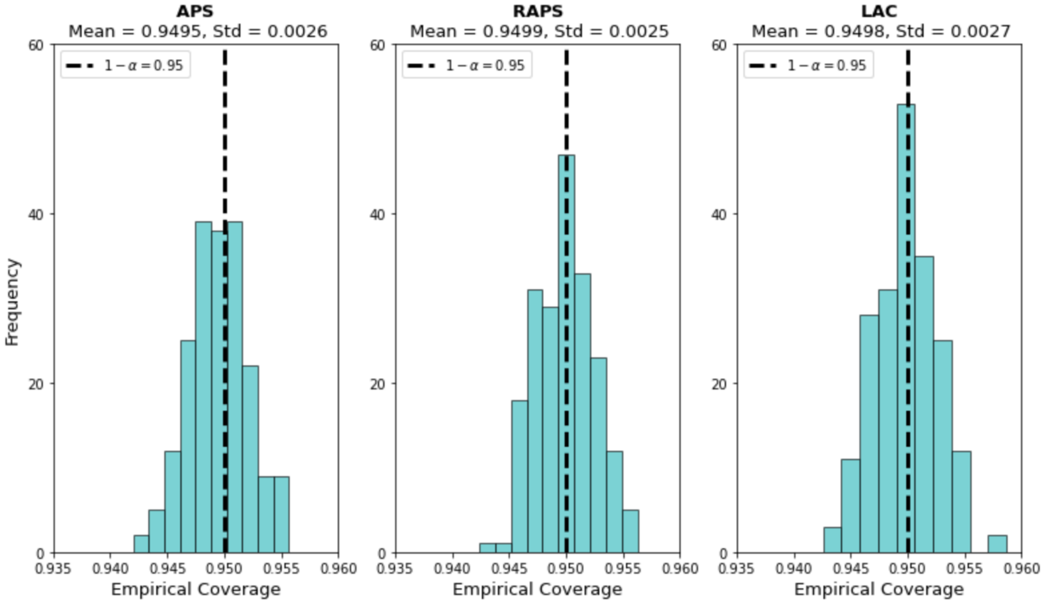


Figure 4.9: (top) Coverage Distribution on Non-Deferred Test Examples for Different D-CP Schemes: Deferral Rate $\beta \approx 0.2$.
(bottom) Coverage Distribution on All Test Examples for Different CP Schemes. For both D-CP and CP, $\alpha = 0.05$, Dataset = CIFAR-10H, Number of Trials = 200, Number of Calibration Points $N = 8000$, Number of Validation Points $N_{val} = 8000$

However, due to the reduced number of finite samples, we would expect a slight increase in the variance of the coverage of the estimator. This is evident in Figure 4.9. [2] show that the standard deviation of the obtained coverage in Equation 4.16 can be expressed as:

$$\text{Std}(C) = \sqrt{\frac{(N+1-l)(N+N_{val}+1)l}{N_{val}(N+1)^2(N+2)}} = \mathcal{O}\left(\frac{1}{\sqrt{\min(N, N_{val})}}\right) \tag{4.17}$$

where N and N_{val} is the size of the calibration and test dataset respectively and $l = \lfloor (N+1)\alpha \rfloor$. Given a deferral rate of β, the effective sizes of N and N_{cal} reduce by a factor of $1-\beta$ for D-CP, increasing the standard deviation of the average coverage by a factor of $\frac{1}{\sqrt{1-\beta}}$. The benefits of smaller predictive sets and human-AI complementarity therefore come at the price of a reduction of statistical efficiency. However, this is not a problem in practice as long as the model doesn't defer a large proportion of examples to an expert. [2] claim that a calibration size of 1000 will be sufficient for most applications employing CP methods. For D-CP, given a model with, say a reasonable 20% deferral rate, the calibration dataset need only be around 25% larger than before to provide empirical coverage with the same variance as conventional CP methods.

4.2.6 Human Subject Evaluation

We now evaluate the benefits of D-CP through human subject evaluations. In particular, we focus on two things:

- Establishing the value of learning to defer, especially in situations where the model may provide misleading examples.
- Establishing the value of smaller set sizes when the error rate α and deferral rate b are held constant

We choose another set of 15 examples from the CIFAR-100 test set for which we generate RAPS prediction sets with error rate $\alpha = 0.1$ and D-RAPS prediction sets with deferral rate 0.2 and error rate $\alpha = 0.1$. We select 12 non-deferred examples at random wherein the D-RAPS predictive set is smaller than the RAPS predictive set, but the ground truth labels are contained in both sets. Lastly, we choose 3 deferred examples where the model is underconfident, i.e. the ground truth label is not in the RAPS set. We ask participants the same questions as in Section 3 and follow a similar recruitment procedure as in Section 3 (except here we have 60 participants total, 2 groups, reward of £10 per hour prorated). We also perform the same statistical tests as in Section 5.1.1 to check for significance. As improved system accuracy can have important implications for real world deployment of Human-AI teams, we additionally display a more comprehensive statistical reportage in Table 4.10. Hereto, we show the minimum sample size needed for statistical significance of accuracy with power $1 - \beta = 0.8$. This is done using the `statsmodels` Python library.

Definition 4.3 (Statistical Power)**.** The statistical power of a binary hypothesis test is the probability that the test correctly rejects the null hypothesis H_0 when the alternative hypothesis H_1 is true.

Metric	D-RAPS	RAPS	p value	Effect Size
Accuracy	0.76 ± 0.08	0.67 ± 0.05	**0.003**	0.832
Reported Utility	7.93 ± 0.39	6.32 ± 0.60	< **0.001**	1.138
Reported Confidence	7.31 ± 0.29	7.28 ± 0.29	0.862	0.046
Reported Trust in Model	8.00 ± 0.45	6.87 ± 0.61	**0.006**	0.754

Table 4.8: D-RAPS vs RAPS: All Examples

Metric	D-RAPS	RAPS	p value	Effect Size
Accuracy	0.88 ± 0.05	0.81 ± 0.04	0.058	0.508
Reported Utility	7.93 ± 0.39	6.19 ± 0.62	< **0.001**	1.211
Reported Confidence	7.78 ± 0.33	7.31 ± 0.34	0.059	0.507

Table 4.9: D-RAPS vs RAPS: Non-Deferred Examples

Metric	RAPS	D-RAPS	N	p-value	Effect Size	$\mathbf{N_{min}}$
Accuracy (All)	0.67	0.76	30	**0.003**	0.87	22
Accuracy (Easy)	0.87	0.83	30	0.310	0.27	218
Accuracy (Difficult)	0.55	0.67	30	< **0.001**	1.04	16

Table 4.10: Accuracy of participants when shown RAPS vs D-RAPS sets on examples stratified by difficulty. N_{min} is the minimum sample size for each group needed for $p \leq 0.05$ with power $1 - \beta = 0.8$ and N is the experimental sample size of each group.

- Tables 4.8 and 4.9 suggest that there is a statistically significant increase in expert accuracy when the D-CP scheme is employed, with borderline significance on non-deferred examples. Even though participants did not perform as well on deferred examples in general (as their overall accuracy is lower than their accuracy on non-deferred examples), we noticed that their accuracy was still higher than when they were shown CP sets, which contained misleading labels.

- Table 4.10 shows a retrospective power analysis on the results of the D-RAPS vs RAPS experiments. We divide the 15 images chosen into 3 difficulty groups - where difficulty is defined as the entropy of the model predictive distribution - and evaluate the statistical significance of the accuracy on the easiest and most difficult groups. It is seen that the accuracy increases the most on examples the model found difficult, which are by definition the most likely to be deferred. On the other hand, there is no increase in accuracy of easy examples.
- Equally interestingly, on examples where both RAPS and D-RAPS sets contain the ground truth label (i.e. the non-deferred examples in Table 4.9), the perceived utility of D-CP sets is higher ($p < 0.001$). As D-RAPS sets are smaller, this shows that, for the same confidence level, smaller set sizes can be more useful to experts and therefore a preferred choice for human-AI teams.
- Table 4.8 also shows a statistically significant difference in reported trust in the model between D-RAPS and RAPS. The improved trust we gain is important – humans strongly prefer work with AI teammates when there is high perceived trust, even when the AI teammate is not state of the art [35].
- Table 4.10 also verifies the choice of sample size. We need at least 22 samples in order to obtain statistically significant results for overall accuracy with sufficiently high power - in our experiments, we initially chose the sample size as 30.

We now provide another interesting finding that warrants caution when deploying models with large CP sets in human-AI teams. [13] established for binary classifiers that model predictions influence expert decisions and that displaying incorrect predictions can cause experts to err in judgement when compared to purely deferring predictions. We report similar findings for set valued predictions in this report.

Definition 4.4 (Bias in CP)**.** We define the *bias* towards incorrect predictions as the proportion of examples where an incorrect prediction made by an expert (who has observed the model's predictive set) is found in the predictive set output by the model averaged across all subjects. That is, given experts h, examples x, the associated label $y(x)$, and the CP set $\Gamma(x)$:

$$\text{Bias} = \mathbb{E}_{h,x}\left[\mathbb{I}_{h(x)\in\Gamma(x)}\mathbb{I}_{h(x)\neq y(x)}\right] \approx \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}_{h(X_i)\in\Gamma(X_i)}\mathbb{I}_{h(X_i)\neq Y_i} \tag{4.18}$$

Metric	**D-RAPS**	**RAPS**	**RAPS**
		Non-Deferred Examples	**Deferred Examples**
Bias	0.063 ± 0.035	0.189 ± 0.046	0.933 ± 0.069

Table 4.11: D-RAPS vs RAPS: Bias towards incorrect or misleading labels

Comparing just the non-deferred examples (where both D-RAPS and RAPS sets contain the true label) we see that experts are much more biased towards incorrect predictions in RAPS sets than in D-RAPS sets. This is a consequence of RAPS sets containing more incorrect labels, which presents more scope for ambiguity. Another interesting observation is that on examples deferred by D-RAPS (whose RAPS sets contain only incorrect labels), expert are much more reliant on RAPS predictions (as evidenced by the high bias of 0.933). These findings warrant caution when deploying models with only CP wrappers in human-AI teams, as large, incoherent sets in critical settings can result in costly mistakes when expert bias their decisions heavily on model predictions.

5 Calibrated Risk Control: A Radicalisation Use-Case

5.1 Introduction

We now discuss an application of D-CP in a real world use-case - predicting the extent of radical activity carried out by extremists. For this, we employ the Profiles of Individual Radicalization in the United States (PIRUS) [1] dataset. This dataset contains anonymised information on the socioeconomic background, personality, childhood, ideology, and radicalization process of 2226 individuals in the USA who have a history of extremist activities, either violent or non-violent in nature, from 1948 to 2018. The aim of this dataset is to enable the development of methods that help understand the process of radicalization from a scientifically rigorous perspective and help mitigate this process in its nascent stages. We summarize the main properties of the dataset in Table 5.1.

Dataset	# Features	# Instances	Labels	Domain
PIRUS	128	2226	Unlabelled	Tabular (i.e. we process each instance to be a 128 dimensional feature vector)

Table 5.1: Summary of the PIRUS dataset

5.2 Data Preprocessing and Methodology

We observed the following characteristics of the PIRUS dataset

- The dataset contains noisy entries with missing features. To deal with this issue, we replaced all instances of missing feature j with the average feature value found across the dataset.

- The dataset is unlabelled, prompting the need to generate synthetic labels for further analysis. We do so using the procedure outlined in Table 5.2. As the labels are generated using select features, in the event that one of the relevant features is missing for an instance, we remove that instance from the dataset. This was done for 20 out of 2226 datapoints.

- For each instance, the dataset contains timestamps reflecting various events such as the approximate date of exposure to radical ideologies, and date of religious conversion (if applicable). We converted dates to UNIX timestamps - which represent seconds passed since 00:00:00 UTC, Jan 1970.

- The dataset contains features of different scales. To counteract this, we performed feature normalization, i.e. for the j^{th} feature of the i^{th} instance, we replaced the feature value by:

$$\hat{x}_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j} \tag{5.1}$$

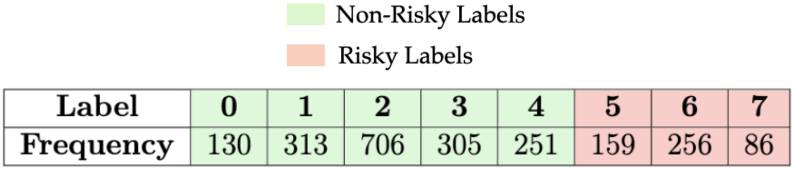
$$\hat{\mu}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij} \tag{5.2}$$

$$\hat{\sigma}_j^2 = \frac{1}{N}\sum_{i=1}^{N}(x_{ij} - \mu_j)^2 \tag{5.3}$$

for all features.

Label	Plot Extent (1-5)	Criminal Severity (0-10)	Anticipated Fatalities (0-3)
7 (Highest Risk of Violence)	Successful execution (5)	Assault with deadly weapon (10)	Greater than 20 (2, 3)
6	Successful execution (5)	Assault with deadly weapon (10)	Greater than 1 (1)
5	Successful execution (5)	Assault with deadly weapon (10)	None
4	Partial or Failed Execution (< 5)	Assault with deadly weapon (10)	Any
3	Any	Felony / Arson (8/9)	Any
2	Any	Threats / Conspiring / Unlawful Possession (5-7)	Any
1	Any	Vandalism / Incitement (3/4)	Any
0 (Lowest Risk of Violence)	Any	None / Illegal Protest / Trespassing (0-2)	Any

Table 5.2: The criterion used for generating labels from 3 features found in the PIRUS dataset. We place high emphasis on severity of crimes already committed as we consider them to be a meaningful indicator of future risk of violence. However, future research could consider alternate methods of label generation.



🟩 Non-Risky Labels

🟥 Risky Labels

Label	0	1	2	3	4	5	6	7
Frequency	130	313	706	305	251	159	256	86

Figure 5.1: PIRUS Dataset Label Distribution

We then remove all features that were used to generate the labels and use the remaining features to predict the labels. We train a standard Multilayer Perceptron (MLP) with an additional deferral class on 50% of the dataset, using the loss function in Equation 3.2. We leave the remaining 50% of instances for calibration and validation. We train for 100 epochs, using the Adam optimizer [20] with learning rate $\eta = 0.005$. We now have several objectives in mind during inference, as discussed next.

5.3 Experiments

In this section, we validate risk control using D-CP on the PIRUS dataset. While the previous section only evaluated D-CP in the sense of controlling for the coverage (binary risk or false negative rate) on non-deferred examples, here we look at controlling broader risks on both deferred and non deferred examples. For this dataset, the problem of classifying dangerous individuals is *cost sensitive*, i.e. the cost of misclassifying a dangerous individual as non-violent is much greater than the converse. We aim to show the following:

- We can generate set predictors that simultaneously control for the marginal false negative rate of dangerous instances (i.e. cases where an extremist assaulted civilians with a deadly weapon) and the misclassification rate of the expert. This follows from Theorem 2 in the Appendix.
- Under the cost sensitive setup, we may also want to control the risk of false positives - i.e. including a dangerous label in the predictive set when the true label isn't dangerous - as it can lead to wasted resources pursuing less dangerous targets. We can therefore accordingly define a risk function that simultaneously controls for any desired tradoff between false positive and false negative rates as well as the misclassification rate of the expert.

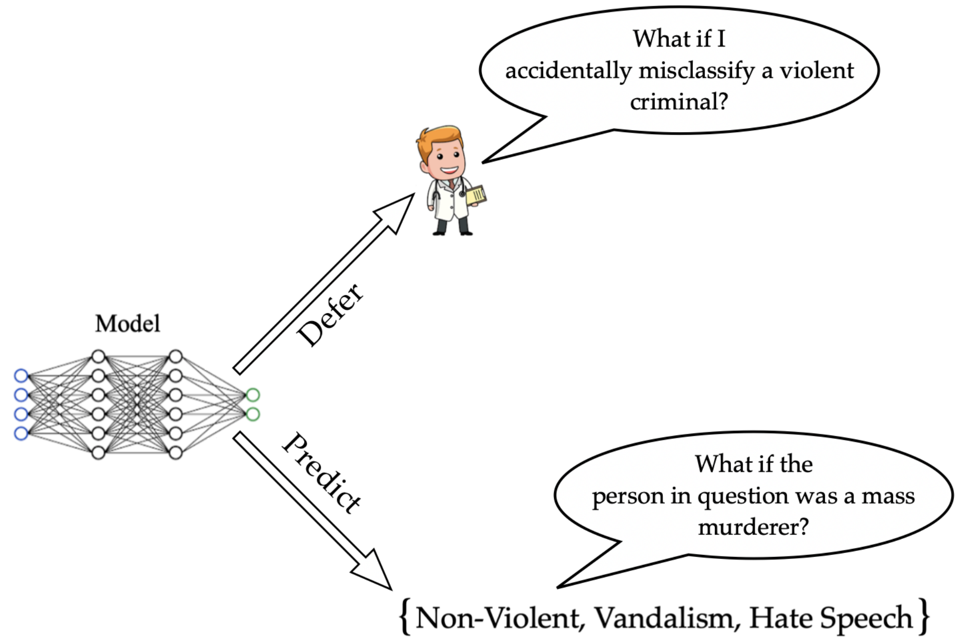

What if I accidentally misclassify a violent criminal?
Model
Defer
Predict
What if the person in question was a mass murderer?
{Non-Violent, Vandalism, Hate Speech}

Figure 5.2: An illustration of the types of risks we want to control with D-CP. We argue that risk control is a multi-faceted problem that needs to be carefully considered from the perspective of both the downstream decision maker and the machine learning model insofar as cost sensitive classification is concerned

We first generate 4 synthetic experts with the following characteristics:

- If the true label is not risky, the expert randomly assigns any label to the example.
- If the true label is risky, i.e. one of $5, 6$, or 7 (where the extremist assaulted civilians with a deadly weapon), the expert can classify the label with the following accuracies:
  - **Expert A:** 95%
  - **Expert B:** 90%
  - **Expert C:** 80%
  - **Expert D:** 70%

We now want a suitable deferral policy such that we defer a non-zero number of examples to such an expert whilst ensuring that the expert is right $1 - \alpha_2$ proportion of the time, on future examples, with high probability. Moreover, whenever the true label is dangerous and we don't defer, we also want the set to contain the label $1 - \alpha_1$ proportion of the time[4]. To this end, given an expert $h(X)$, we define the set predictor

$$\Gamma_\lambda(X) = \begin{cases} \emptyset & \pi(X)_\perp \geq \lambda_1 \\ \{y : \tau(X, y) \geq \lambda_2\} & \text{otherwise} \end{cases}$$

where deferral is equivalent to $\Gamma_\lambda(X) = \emptyset$. Next, we define the risk functions that satisfy the above aforementioned requirements:

$$R_1(\lambda_1) = P(\Gamma_\lambda(X) = \emptyset | h(X) \neq Y, Y \in [5, 6, 7]) \tag{5.4}$$

$$R_2(\lambda_1, \lambda_2) = P(Y \notin \Gamma_\lambda(X) | \Gamma_\lambda(X) \neq \emptyset, Y \in [5, 6, 7]) \tag{5.5}$$

[4] This is reminiscent of the Mondrian paradigm, where we want to control for the error rate on a particular group.

The equivalent empirical risks are:

$$\hat{R}_1(\lambda_1) = \frac{1}{\sum_{j=1}^{N} \mathbb{I}_{r(X_j)\geq\lambda_1, Y_j\in[5,6,7]}} \sum_{i=1}^{N} \mathbb{I}_{h(X_i)\neq Y_i} \mathbb{I}_{r(X_i)\geq\lambda_1, Y_i\in[5,6,7]} \tag{5.6}$$

$$\hat{R}_2(\lambda_1, \lambda_2) = \frac{1}{\sum_{j=1}^{N} \mathbb{I}_{r(X_j)\leq\lambda_1, Y_j\in[5,6,7]}} \sum_{i=1}^{N} \mathbb{I}_{Y_i\notin\Gamma_\lambda(X_i)} \mathbb{I}_{r(X_i)\leq\lambda_1, Y_i\in[5,6,7]} \tag{5.7}$$

Here, λ_1 is the threshold for deferral, i.e. whenever the model's prediction for the deferral class $\pi_\perp(X) \in [0,1] \geq \lambda_1$ we defer. λ_2 is the threshold for including any label in the set. We first tune λ_1 using the RCPS procedure outlined in Section 2.3 to find the smallest λ_1 such that:

$$P(R_1(\lambda_1) \leq \alpha_1) \geq 1 - \delta \tag{5.8}$$

Then, we fix λ_1 and tune λ_2 using RCPS such that

$$P(R_2(\lambda_1, \lambda_2) \leq \alpha_2) \geq 1 - \delta \tag{5.9}$$

We now split the validation set into calibration and test sets over 1000 trials. For each trial, we first perform calibration to determine appropriate λ_1 and λ_2 and then calculate the empirical risks on the test set. The distribution of these empirical risks is then shown in a violin plot in Figure 5.3. We set $\alpha_2 = 0.05$, α_1 equal to the error rate of the expert on risky examples, and $\delta = 0.1$[5].

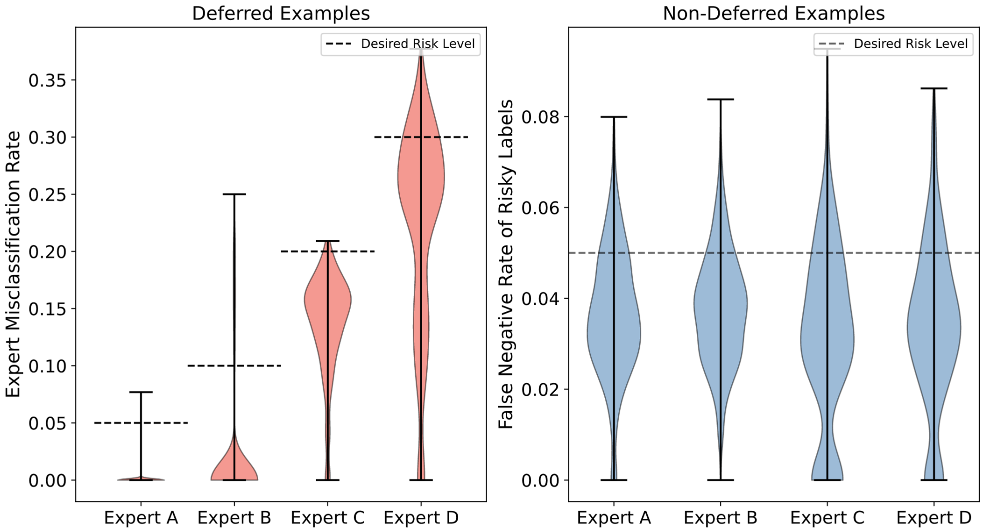


Figure 5.3: Illustration of dual risk control for instances where the extremist assaulted civilians with a deadly weapon over $N = 1000$ calibration-test dataset splits (**left**: Expert Misclassification Rate, **right**: False Negative Rate).

- For each expert, we are able to achieve precise control of the expected misclassification rate. The probability mass above the desired risk level is very small - we found this to be smaller than δ for the values tested, demonstrating that we satisfied the guarantee in Equation 5.8. Note that in general, we were not able to achieve precise control for α_1 when it is much lower than the expert's error rate (for smaller risks, the trivial

[5] Note that while we consider individual control of risks at level δ for these experiments, if we wanted to ensure that we violate *either* of the risks with probability δ, we tune λ_1, λ_2 at a level $\frac{\delta}{2}$ due to the Bonferroni correction procedure (see Proofs)

solution found by the scheme is to defer no examples). This is likely due to inherent limitations of the deferral policy (because it is trained on finite, noisy data, it cannot always accurately gauge which examples an expert will be correct on)[6].

- Furthermore, the deferral policy employed ensured that, with high probability, the expert misclassification rate on risky examples is lower than if we deferred random examples to the expert, illustrating the benefits of leveraging the expert's strengths, as explored in the previous section.
- On the non deferred examples, we obtain precise control of the false negative rate (i.e. the probability the true label is not in the set) for labels $5, 6, 7$. This guarantee is independent of the expert we defer examples to.

Lastly, we can also control for a weighted combination of the false positive rate (FPR) and false negative rate (FNR) for risky labels on non deferred examples. For a given weight β, we define this as:

$$\begin{aligned} R_2(\lambda_1, \lambda_2) = & \beta \underbrace{P\Bigg(\bigcup_{y' \in [5,6,7]} y' \in \Gamma_\lambda(X) | Y \notin [5,6,7], \Gamma_\lambda(X) \neq \emptyset \Bigg)}_{\text{FPR of Risky Labels}} + \\ & (1-\beta) \underbrace{P(Y \notin \Gamma_\lambda(X) | Y \in [5,6,7], \Gamma_\lambda(X) \neq \emptyset)}_{\text{FNR of Risky Labels}} \end{aligned} \tag{5.10}$$

Note that we cannot have small FPR and small FNR simultaneously. This is because, for a given model, smaller set sizes will provide small FPR but larger FNR and vice versa. However, as illustrated in Figure 5.4 (with $\alpha_1 = 0.1$, $\alpha_2 = 0.1$, $\delta = 0.1$, and $\beta = 0.1$) we are able to ensure that the risk in Equation 5.10, which defines the user desired tradeoff between FPR and FNR, is controlled on test examples. Furthermore, changing the value of β also enables us to choose the desired tradeoff between FPR and FNR, which may be useful in many real world applications.

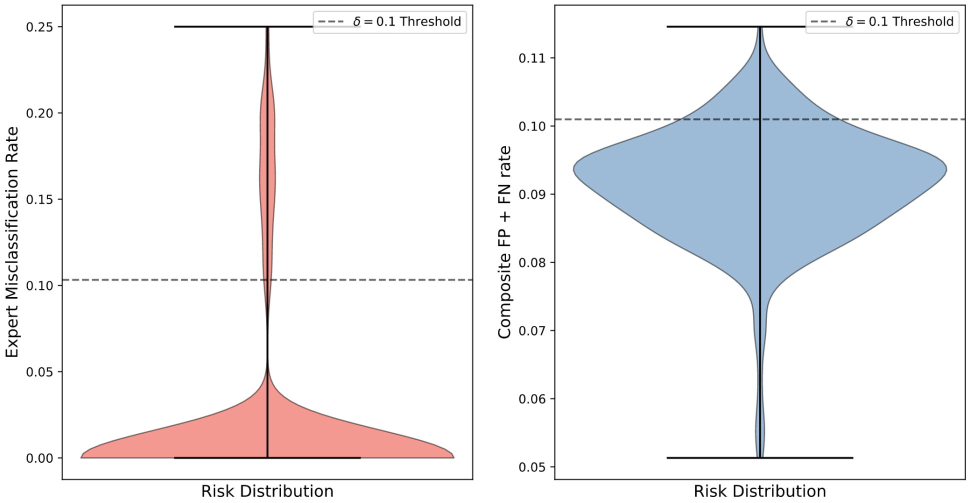


Figure 5.4: Illustration of dual risk control on instances where the extremist assaulted civilians with a deadly weapon over $N = 1000$ calibration-test dataset splits (**left**: Misclassification Rate of Expert B, **right**: Weighted FPR and FNR)

[6] Note that being unable to find a risk controlling set is not unexpected: indeed [3] recommend the model either abstain or provide feedback to the end user that certifiable risk control at the desired level is not possible.

6 Shaping Predictive Sets with Desirable Properties

The construction of D-CP enabled the model to take advantage of the human's strengths and generate smaller predictive sets, obtain better overall performance, and establish dual risk control. However, in many real world applications, it may be impractical to obtain instance-level human labels for large training datasets, i.e. we cannot always defer. Alternatively, we might also want to integrate global, non-instance specific information provided by a human into our predictive set. Thus, we now consider the problem of tackling the relevancy desideratum from the perspective of RCPS. However, the same arguments presented henceforth can easily be applied to CP.
Consider the RCPS setup where we seek to penalise certain labels more (e.g. not predicting `cancer`). In this case, our loss function $L(Y, \Gamma(X))$ associated with the set $\Gamma(X)$ and example Y will be higher for label Y = "cancer". The RCPS setup allows us to control this risk $R(\lambda) = \mathbb{E}_{(X,Y)}(L(Y, \Gamma_\lambda(X))$ to a desired level with high probability. However, we also might want the labels contained in the set to exhibit desirable properties laid out by a human. In this instance, a doctor may want the predictive set to output all lung-related ailments alongside cancer as much as possible - *whilst maintaining the same risk / coverage guarantees.*

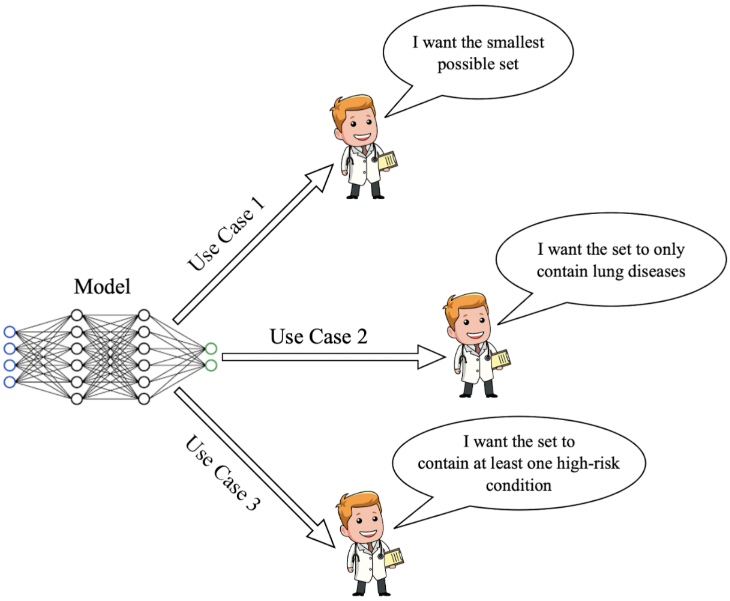

I want the smallest possible set
Use Case 1
Model
Use Case 2
I want the set to only contain lung diseases
Use Case 3
I want the set to contain at least one high-risk condition

Figure 6.1: A set-valued prediction output by a trained model might be used in different scenarios. In order to be useful, we want the model to be able to provide a set that can adapt to any information provided by the stakeholder. As the notion of useful sets is too broad to be captured in this report, we restrict ourselves to one scenario - providing sets that contain as similar labels as possible, where the definition of label similarity is provided by the user.

In the more general setting, we consider a label similarity matrix M provided by a human and we wish to find a set valued predictor Γ that minimizes the cost of dissimilarity whilst being as small as possible and maintaining the desired risk guarantees. Concretely, we might aim to tackle the following problem:

$$\min_{\Gamma} \ \mathbb{E}_{X,Y}[\max_{y,y' \in \Gamma(X)} d(y, y') + \mu|\Gamma(X)|] \tag{6.1}$$

$$s.t. \ \ \hat{R}^{+}(\lambda) \leq \gamma \tag{6.2}$$

where $d(y_i, y_j) = M_{ij}$ and γ is the user defined risk. Note we can also minimise the average dissimilarity, but this might cause the predictor to increase the set size in order to "dilute" the effect of a dissimilar label. This is a hard problem to solve in general because of its non convex nature, so we attempt to make some progress by introducing a heuristic in the next section.

6.1 Proposed Approach

In this section, we propose a heuristic developed independently that can shape the predictive sets in a desirable manner - we resort to this approach as the problem in Equation 6.1 is non-convex and hard to solve analytically. While [37] have recently developed a loss function that can shape CP sets by penalizing the presence of 2 or more labels in the set together, the proposed approach, unlike [37], can be employed in a post-hoc manner and is valid for RCPS as well.

- We modify one key part of the CP / RCPS procedure: construction of sets when given a threshold.
- For multi-class and multi-label classification, the original procedure added all labels whose probabilities were greater than the threshold to the set.
- In our procedure, we construct sets *greedily*: at each stage after adding the most probable label (assuming its probability mass is greater than the threshold) to the set, we redistribute some of the probability mass associated with the remaining labels.
- This redistribution is done according to whether a given label has low dissimilarity with the labels in the set S generated so far. To determine this, we determine the dissimilarity measure of a label y not in the set

  $$d(y, S) = \max_{y' \in S} d(y, y') \tag{6.3}$$

  We ensure that labels which are similar to ones in the set constructed so far will be allocated higher probability mass (at the expense of other remaining labels), so that it is more likely to be included in the set.
- However, we do not want to compromise the predictive power of the model by including low probability labels in the set. The algorithm therefore, requires a parameter μ that determines the relative importance of including similar labels in the set.
  - Higher positive values of μ mean that the set is composed of similar labels, some of which may have had lower probability originally. Consequently, one would expect that the average set size increases.
  - Lower positive values of μ mean that the set may be smaller, but is composed of more dissimilar labels.
  - From Algorithm 5 in the Appendix and the evalauation of the procedure in Figure 6.3, positive values of μ increase the label *similarity* and negative values increase label *dissimilarity.*

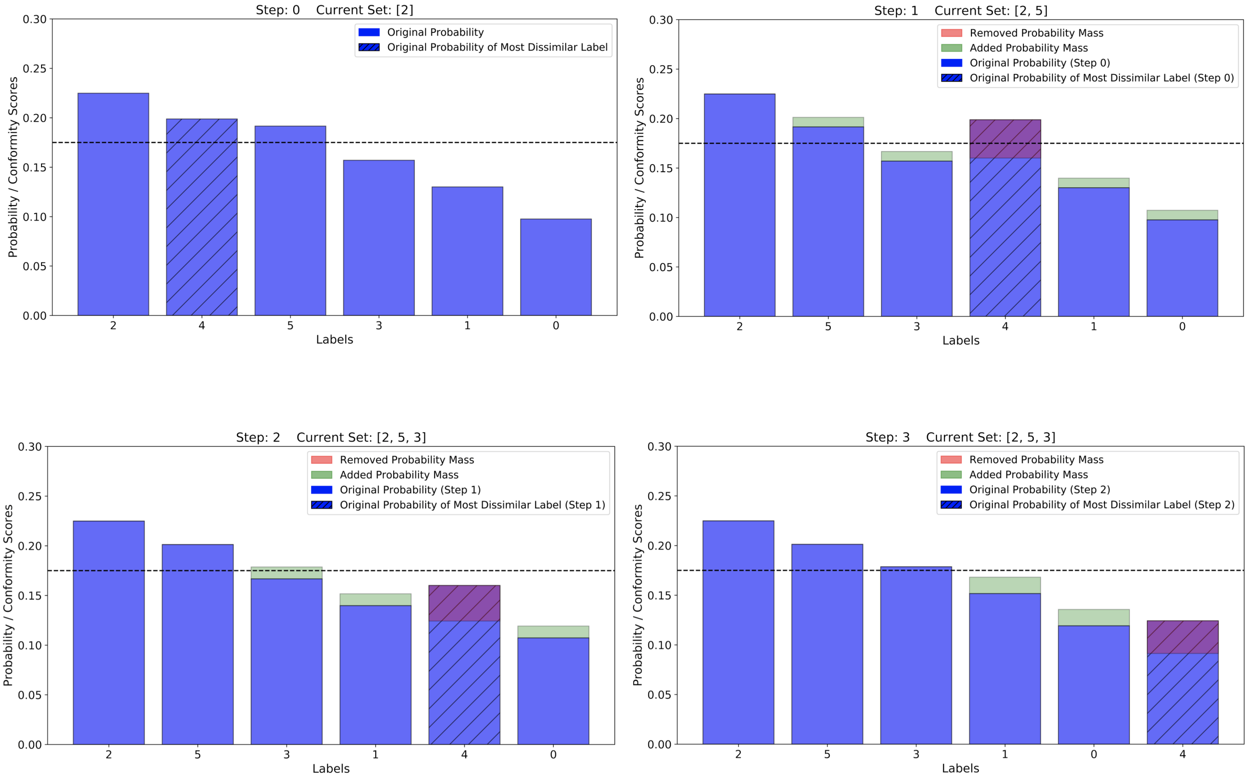


Figure 6.2: Our set construction procedure for a 6 label toy softmax score distribution ($\mu = 0.2$). The dotted black line denotes the threshold. We consider the scenario when $d(4, 2)$ is high, i.e. it is undesirable to have labels 4 and 2 in the same set. Note that the same approach works for any label dissimilarity matrix M provided.

6.2 Experimental Evaluation

We perform experiments with RCPS to investigate the effect of our procedure on the CIFAR-100 dataset. We define the loss and risk as

$$L(Y, \Gamma_{\lambda,M}(X)) = L_Y \mathbb{I}_{Y \notin \Gamma_{\lambda,M}(X)} \tag{6.4}$$

$$R(\lambda, M) = \mathbb{E}[L(Y, \Gamma_{\lambda,M}(X))] \tag{6.5}$$

where each L_Y is sampled from the uniform distribution $\mathcal{U}[0, 1]$, i.e. we assign random, fixed costs to each label. To compute the $1 - \delta$ upper-confidence bound for the risk, we use the Central Limit Theorem concentration inequality (Equation 2.35). For our experiments, we define the label similarity matrix M, where:

$$M_{y,y'} = ||\texttt{emb}(y) - \texttt{emb}(y')||_1 = d(y, y') \tag{6.6}$$

where $\texttt{emb}(y)$ refers to the *word embedding* of label y. A word embedding is a vectorised representation of a word learned by a language model that encodes semantically meaningful information. Here, we employ the BERT word embeddings [15] for our label similarity matrix. This enables us to construct sets that contain *semantically* similar labels. Note that some labels contain multiple words (e.g. `maple tree`). In those cases, we take the element-wise average of the embedding for each individual word as our overall label embedding.

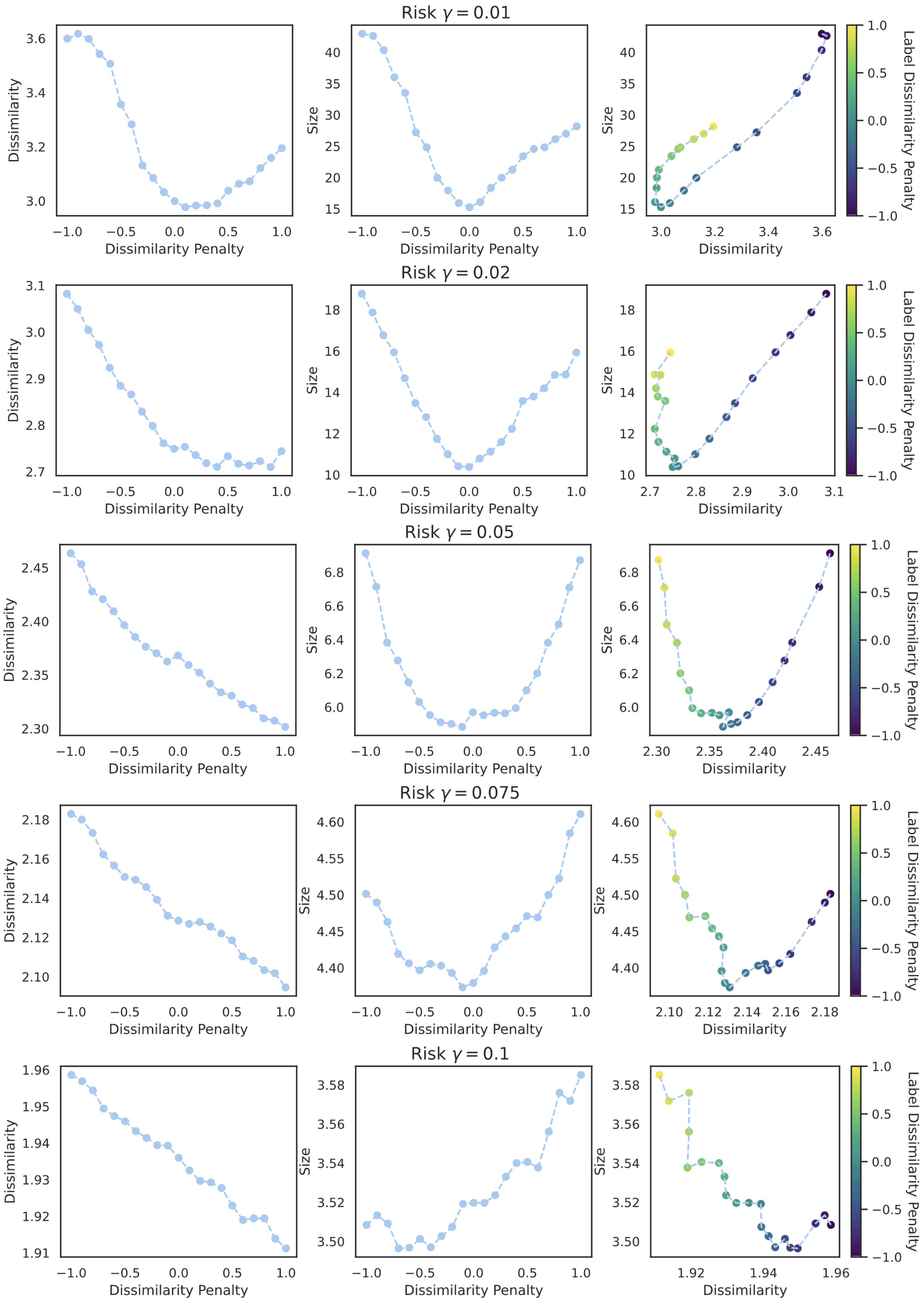


Figure 6.3: For each risk level γ in the RCPS setting, we plot the effect of the dissimilarity penalty μ on the average maximum dissimilarity between labels in the set and the average set size. Dataset - CIFAR-100. Note $\delta = 0.1$ for all γ, μ tested. Similar results were seen for the CIFAR-10 and ImageNet datasets (not shown in this thesis).

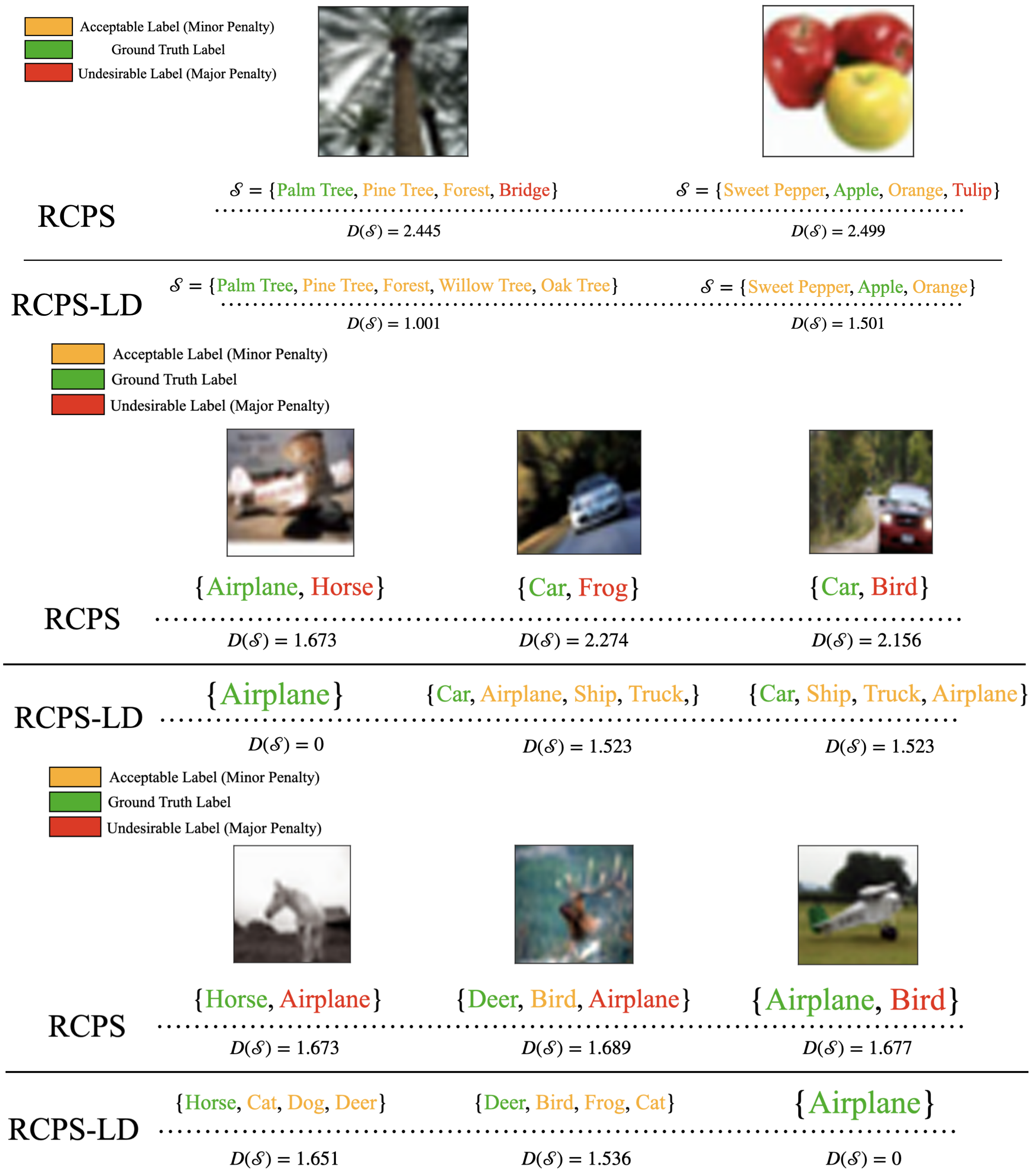


Figure 6.4: Applying our algorithm with the dissimilarity matrix defined by BERT embedding distances gives semantically similar sets on average. However, the set size often increases. The top row shows images from the CIFAR-10 dataset and the bottom 2 rows show images from the CIFAR-100 dataset. Here, $\gamma = 0.1$, $\delta = 0.1$ and, $\mu = 0.5$.

From Figures 6.3 and 6.4, we make the following observations:

- For higher levels of risk, we obtain an approximately monotonic relationship between μ and the label dissimilarity and μ and the set size. Because higher risks imply smaller set sizes, there are a lot of similar labels not yet in the set - the algorithm therefore has scope for reshaping the set drastically by redistributing probability mass towards or away from these labels.

- Furthermore, the minimum set size is found approximately when $\mu = 0$: the average set size starts increasing regardless of whether we penalize label dissimilarity or label similarity. This is likely because the redistribution of probabilities causes a slight degradation in the Top-K performance of the underlying model - the RCPS wrapper compensates for this by increasing the set size. However, the increases are not very significant except for very low levels of risk.

- For lower levels of risk, we notice that the reduction in dissimilarity obtained by the algorithm starts plateauing (and indeed increases even for positive μ for $\gamma = 0.01$). Because controlling for lower levels of risk result in larger set sizes (so that we don't miss out on "risky" labels), it becomes increasingly likely that the incumbent set contains very dissimilar labels. Thus, as our scheme can only choose the least dissimilar label not in the set in the best case, we essentially end up choosing labels $y \notin S$ that have a high $d(y, S)$, thereby increasing $\max_{y,y' \in S} d(y, y')$. This is an inherent limitation of the model - we cannot reduce dissimilarity for large set sizes.

6.3 A Short Note on Coverage Guarantees

To compute the example wise loss / risk, we first first perform Algorithm 5 to generate a set of modified softmax probabilities and the corresponding set of labels. We then find the threshold that provides us the guarantee in Equation 2.26 using the standard RCPS procedure. Thus, with the addition of the label similarity matrix, we retain the same guarantees, as we are making no modification to the calibration procedure - the idea of redistributing probability mass is equivalent to modification of the conformity scores of each label. This is illustrated in Figure 6.5, where the distribution of risk (defined in Equation 6.4) on CIFAR-100 is plotted over 1000 random splits of the test and calibration datasets. The risk distributions look identical, *implying that risk only exceeds the desired level δ proportion of the time, regardless of the level of set dissimilarity desired.*

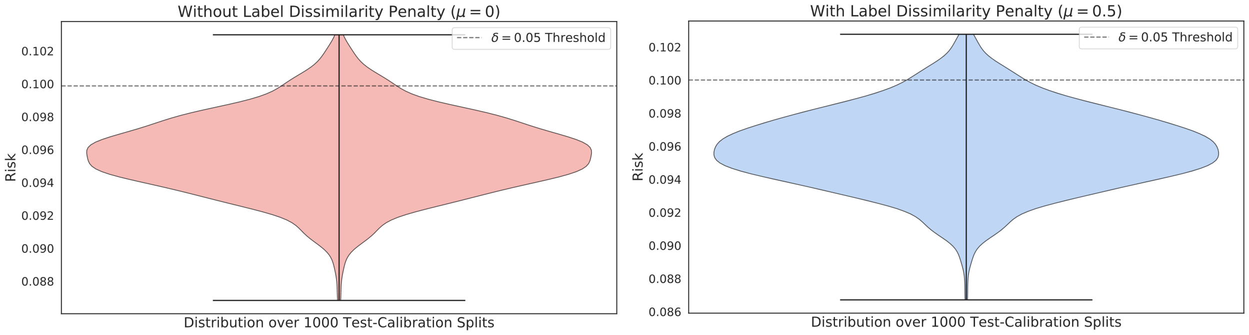


Figure 6.5: The risk distributions without (left) and with (right) application of the label dissimilarity penalty look identical, proving that we can obtain similar sets without compromising risk control. Here, $\gamma = 0.1$, $\delta = 0.05$, and $\mu \in \{0, 0.5\}$.

Note that we can also obtain the same coverage guarantees if we defined an instance wise label similarity matrix. That is, we draw an example, its corresponding label, and a label similarity matrix from the joint distribution $P(X, Y, M)$. Thus, we define the risk as:

$$R(\lambda) = \mathbb{E}_{(X,Y,M)\sim P(X,Y,M)}[L(\Gamma_\lambda(X), Y, M)] \approx \frac{1}{N}\sum_{i=1}^{N} L(\Gamma_\lambda(X_i), Y_i, M_i) \tag{6.7}$$

and the coverage guarantee in Equation 2.26 holds marginally over all examples, labels, and label similarity matrices.

7 Conclusions and Future Work

In this project, we developed techniques for constructing predictive sets that are maximally useful to humans. We first explored main ideas and contemporary techniques in conformal prediction literature that satisfy the size, adaptiveness, and coverage / risk control desiderata [34, 4, 33, 9]. We then briefly introduced a scheme by [26] that introduces a loss function for deferring examples to an expert. We finally combined the above two ideas to develop a new scheme called D-CP that learns to perform CP on some examples and defer others. We formally proved some theoretical properties, performed experimental evaluation on 2 datasets, and conducted human subject experiments to verify how our scheme performs. We then evaluated how D-CP can help control for the misidentification of dangerous, extremist individuals in a real-world use case. Lastly, keeping a broader notion of useful sets in mind, we introduced an alternate, complementary idea of utility wherein we enable humans to shape predictive sets according to their notion of similarity - all whilst maintaining desired risk / coverage guarantees. The resulting post-hoc algorithm was evaluated on CIFAR-100 with varying risk and semantic label similarity parameters. Our conclusions are the following:

- Our preliminary human subject experiments showed a statistically significant increase in perceived utility of CP sets (in particular, RAPS) over Top-1 predictions. Providing CP sets also resulted in humans trusting the model more. We further verified these findings through subsequent experiments with a control that consisted of a Top-1 + random labels prediction, showing that humans prefer principled, calibrated predictions that accurately quantify model uncertainty.

- Experimental evaluation on toy datasets (CIFAR-100 [22] + CIFAR-10 [21]) showed that D-CP leverages the complementary strengths of the model and the human to provide smaller set sizes for the same level of coverage and higher overall system accuracy.

- Our theoretical results prove that D-CP lowers the predictive set size of non-deferred examples with a suitable deferral policy (i.e. as long as it defers examples it is less confident on) that is practical to achive in real world settings. We also proved that we can simultaneously guarantee user-defined coverage on non-deferred examples and control for the misclassification rate by an expert with high probability.

- Our human subject experiments on D-CP also revealed that D-CP sets provide increased team accuracy, utility, and trust on all examples, and reduced bias towards misleading labels.

- We showed that we can apply D-CP to obtain multiple risk control on a real world dataset (PIRUS) where the aim is to identify extremist individuals who have the potential to commit violent crimes. We were able to simultaneously control for the misclassification risk of an expert and achieve a desirable trade-off between false positive and false negative rates on non-deferred examples for instances where the individual committed violent crimes.

- Our algorithm for controlling the label dissimilarity can provide more similar sets, regardless of the definition of label similarity provided by a human. In this case, we showed through some examples on CIFAR-10 and CIFAR-100 that we can achieve sets with greater semantic similarity, which we defined as the distance between BERT embeddings of labels.

- However, we also showed that the benefits of this approach both in terms of the increase in label similarity and a small rise in predictive set size are compromised if we aim to control for small risks - this is a natural consequence of the algorithm being unable to deal with very large set sizes that are necessary to represent low risks.

Our work, however, leaves a lot of scope for future evaluation and experimentation.

7.1 Future Work: Human Subject Experiments

- A preliminary survey (*not discussed in this thesis*) was conducted asking people what properties did they thing a predictive set ought to have. We found that people did not provide useful answers, with the exception of a few participants who requested sets to provide more similar labels. Thus, quantifying and qualifying the exact kind of sets humans may find useful remains an open question.

- We haven't evaluated the effect of inducing label similarity on resulting human-AI team performance. While our scheme is general in that it allows humans to provide any label similarity matrix, we need to establish its benefits in real-world human-AI teams.

- Our human subject experiments for validating whether CP sets are actually useful and for establishing the value D-CP over CP sets were performed with relatively small sample sizes (15 per group for the former experiment and 30 per group for the latter). While these choices of sample sizes gave us statistically significant results on a toy dataset like CIFAR-100, it raises the question of whether our conclusions will be applicable on real-world datasets, where a host of other factors come into play. Some of these include but are not limited to:
  - Potential for bias in responses: for example, unequal importance of some labels over others - as seen in the PIRUS dataset - may bias participant responses. Given this, does controlling for a more involved risk function using RCPS affect trust, fairness, and accuracy of Human-AI teams?
  - Niche datasets where humans are only good on a known subset of examples - does this affect our conclusions? Does the type of deferral policy found in contemporary literature impact the utility of predictive sets as perceived by humans?
  - Does the value of α - the error tolerance parameter - impact the utility of predictive sets? What is the optimal value of α and does this vary in different settings? [36] has recently made some progress in this regard by developing an algorithm that tunes α to find the most suitable one for a human, but this needs to be tested on real participants.
  - We established that set valued predictions and D-CP in general lead to greater perceived trust in human-AI teams. However, we may also want to be cautious of unwarranted trust [18] or over reliance by humans, and appropriately design and test schemes that can control for this.

7.2 Future Work: Algorithm Development

Most of prior work in designing deferral policies has focused on enhancing trust, accuracy, and compatibility [7]. Here, we argue that more focus needs to be given to develop improved deferral policies that are designed with the model's predictive uncertainty in mind. In this thesis, the deferral policy is trained to defer whenever it thinks the human would get the example right - no specific details of the model's predictive uncertainty were learnt. One step in this direction could consider the development of a composite loss function that jointly learns to defer and optimize the CP set size on non-deferred examples, i.e. a joint loss function combining [26] and [37] that is optimized batch-wise.

7.3 Future Work: Statistical Analysis and Theoretical Development

- The RCPS procedure relies on finding a upper confidence bound of the empirical risk function using a concentration inequality. If we can find lower upper concentration bounds to the risk function, we will be able to obtain smaller set sizes that control for the same risk with the same probability $1-\delta$. Can we find a lower upper concentration bound on the risk using human intervention? If a human can provide alternate side information $h(X)$, one way to do this might be constructing a control variate setup wherein the risk is defined as:

$$\hat{R}(\lambda) = \frac{1}{N}\sum_{i=1}^{N} L(Y, \Gamma_\lambda(X)) + c(h(X) - \mathbb{E}[h(X)]) \tag{7.1}$$

  and we tune c to minimize the variance of $\hat{R}(\lambda)$, which should provably lower the UCB for any concentration inequality. However, extensive analysis would need to be conducted to verify the conditions under which this scheme would be practically useful.

- One potential method for formalizing the label similarity problem may be to look at redistribution of softmax probabilities from an optimal transport [30] perspective. In this problem, we want to transfer probability mass from one distribution to another (in our case, the original vs optimal softmax distribution) whilst minimizing the cost of transportation. Hereto, the unit cost of transporting probability mass from y_i to y_j could be the term M_{ij} in the label similarity matrix M. However, in this case, the optimal distribution is unknown, so we would need to simultaneously solve for the optimal redistribution procedure and the optimal distribution, which may present theoretical and practical challenges.

- While coverage for the ground truth label conditional on any given example is unrealistic in practice, can we generalise CP to a setting where we provide conditional coverage guarantees on any label (not necessarily the ground truth)? For instance, given an MRI scan of the brain, what is the probability that the label `cancer` should be in the set? In particular, for any label y, can we bound the probability

$$P(y \notin \Gamma(X)|X) \tag{7.2}$$

  If so, we can output a set that is guaranteed to contain the true label with $1-\alpha$ probability but is additionally certified to contain all labels whose probabilities of being in the set are greater than $1-\beta$. This is useful because a) the presence of dangerous labels in a set could potentially bias the human's predictions and b) there may be borderline candidates that actually belong to the predictive set but are not included because of residual error in the finite sample estimate of τ_{cal}. As a starting point for future research, we consider the uncertainty in the threshold τ_{cal} and provide a bound for Equation 7.2 in the Appendix (Theorem 3). This can also be empirically verified using a bootstrap procedure, where one can use multiple resampled datasets to generate a confidence bound for τ_{cal}.

A Proofs

Theorem 1. *Consider a deferral policy $r(x) : \mathcal{X} \to \{0,1\}$ and a classification model $m_\theta(x) : \mathcal{X} \to \mathcal{Y}$ acting on a dataset $\mathcal{D} = \{(X_1, Y_1), ...(X_n, Y_n)\}$. Define some conformity measure $\tau(x, y)$ such that if $\pi_{\hat{y}}(\hat{X}) \geq \pi_y(X)$ then $\tau(\hat{X}, \hat{y}) \geq \tau(X, Y)$ for any softmax probabilities $\pi_{\hat{y}}(\hat{X}), \pi_y(X)$, labels $\hat{y}, y \in \mathcal{Y}$, and inputs $\hat{X}, X \in \mathcal{X}$. If the expected loss on non-deferred examples is lower than the original loss, i.e. $\mathbb{E}_{(X,Y)|r(X)=0}[\mathcal{L}(Y, m_\theta(X))] \leq \mathbb{E}_{(X,Y)}[\mathcal{L}(Y, m_\theta(X))]$, then the average conformal predictive set of non-deferred examples will contain fewer incorrect labels on average.*

Proof. Because the expected loss on non-deferred examples is lower, we know that:

$$\mathbb{E}_{(X,Y)|r(X)=0}[\pi_Y(X)] \geq \mathbb{E}_{(X,Y)}[\pi_Y(X)] \tag{A.1}$$

From our definition of the conformity measure $\tau(x, y)$ above:

$$\mathbb{E}_{(X,Y)|r(X)=0}[\tau(Y, X)] \geq \mathbb{E}_{(X,Y)}[\tau(Y, X)] \tag{A.2}$$

for any ground truth label Y associated with an example X. Therefore,

$$\mathbb{E}_{(X,Y)|r(X)=0}\left[\sum_{\substack{y'=1\\y'\neq Y}}^{K} \pi_{y'}(X)\right] \leq \mathbb{E}_{(X,Y)}\left[\sum_{\substack{y'=1\\y'\neq Y}}^{K} \pi_{y'}(X)\right]$$
$$\Rightarrow \mathbb{E}_{(X,Y)|r(X)=0}\left[\sum_{\substack{y'=1\\y'\neq Y}}^{K} \tau(y', X)\right] \leq \mathbb{E}_{(X,Y)}\left[\sum_{\substack{y'=1\\y'\neq Y}}^{K} \tau(y', X)\right]$$

Because $\mathbb{E}_{(X,Y)|r(X)=0}[\tau(Y, X)] \geq \mathbb{E}_{(X,Y)}[\tau(Y, X)]$, $\tau'_\alpha = \text{Quantile}(\alpha, \{\tau(Y, X)|(X, Y) \in \mathcal{D}, r(X) = 0\}) \geq \tau_\alpha = \text{Quantile}(\alpha, \{\tau(Y, X)|(X, Y) \in \mathcal{D}\})$ for any user defined error tolerance $\alpha \in [0, 1]$. Thus, by the monotonicity of the indicator function $\mathbb{I}_{\tau(X,Y)\geq u}$ with respect to argument u:

$$\mathbb{E}_{(X,Y)|r(X)=0}\left[\sum_{\substack{y'=1\\y'\neq Y}}^{K} \mathbb{I}_{\tau(y',X)\geq\tau'_\alpha}\right] \leq \mathbb{E}_{(X,Y)|r(X)=0}\left[\sum_{\substack{y'=1\\y'\neq Y}}^{K} \mathbb{I}_{\tau(y',X)\geq\tau_\alpha}\right] \leq \mathbb{E}_{(X,Y)}\left[\sum_{\substack{y'=1\\y'\neq Y}}^{K} \mathbb{I}_{\tau(y',X)\geq\tau_\alpha}\right]$$

This implies:

$$\mathbb{E}_{(X,Y)|r(X)=0}[|\{y'|\tau(X, y') \geq \tau'_\alpha, y' \neq Y\}|] \leq \mathbb{E}_{(X,Y)}[|\{y'|\tau(X, y') \geq \tau_\alpha, y' \neq Y\}|]$$

□

Theorem 2. *Given any deferral policy $r(X)$, classification model $m_\theta(X)$, and an expert $h(X) \in \mathcal{Y}$, we can guarantee $1 - \alpha_1$ coverage of the true label on non-deferred examples*

$$P(Y_{test} \notin \Gamma(X_{test})|\Gamma(X_{test}) \neq \emptyset) \leq \alpha_1 \tag{A.3}$$

and guarantee that the average $0 - 1$ loss on deferred examples is less than α_2

$$P(\Gamma(X_{test}) = \emptyset|h(X_{test}) \neq Y_{test}) \leq \alpha_2 \tag{A.4}$$

for suitably defined α_1, α_2. Violation of any of these risks will happen with probability at most δ, for any suitably defined δ.

Proof. We employ techniques from the Learn Then Test (LTT) procedure in [3] for Out-Of-Distribution (OOD) detection. This is similar to RCPS except that it is designed to control for risks that are not necessarily monotonic with respect to the threshold parameter λ. Here, the equivalent OOD example would be one the expert is correct on, and we would defer this example. Thus, we mark an example where the model defers as OOD. We want to defer some examples while controlling for the risk of the model deferring an example the expert makes a mistake on. This is equivalent to the risk of marking an example as OOD when it is actually in distribution, i.e. a false positive in a sense. Define the risk functions

$$R_1(\lambda_1) = P(\Gamma_\lambda(X) = \emptyset | h(X) \neq Y) \tag{A.5}$$

$$R_2(\lambda_1, \lambda_2) = P(Y \notin \Gamma_\lambda(X) | \Gamma_\lambda(X) \neq \emptyset) \tag{A.6}$$

where deferral is equivalent to outputting an empty set $\emptyset$. λ_1 is the threshold for deferral, i.e. whenever $r(X) \in [0,1] \geq \lambda_1$ we defer and λ_2 is the threshold for including any label in the set (e.g. but not limited to the threshold conformity score τ_{cal}). That is:

$$\Gamma_\lambda(X) = \begin{cases} \emptyset & r(X) \geq \lambda_1 \\ \{y : \tau(X,y) \geq \lambda_2\} & \text{otherwise} \end{cases}$$

Note that - we have access to examples where the expert errs in the calibration and training datasets (and want to control it for future examples) so this is not a fallible procedure. Employing the method in [3] now controls for the risk as follows:

- Select a discrete grid of parameters $\lambda \in \Lambda$. Here, this corresponds to the 2D vector $\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix}$. For each λ, associate the null hypotheses $H_0^1 : R_1(\lambda_1) \leq \alpha_1$ and $H_0^2 : R_2(\lambda_1, \lambda_2) \leq \alpha_2$.
- For a given λ, obtain the p values for each λ_1 and λ_2. This is done by defining the empirical risks

$$\hat{R}_1(\lambda_1) = \frac{1}{\sum_{j=1}^N \mathbb{I}_{r(X_j) \geq \lambda_1}} \sum_{i=1}^N L(Y_i, h(X_i), \Gamma_\lambda(X_i)) \tag{A.7}$$

$$= \frac{1}{\sum_{j=1}^N \mathbb{I}_{r(X_j) \geq \lambda_1}} \sum_{i=1}^N \mathbb{I}_{h(X_i) \neq Y_i} \mathbb{I}_{r(X_i) \geq \lambda_1} \tag{A.8}$$

$$\hat{R}_2(\lambda_1, \lambda_2) = \frac{1}{\sum_{j=1}^N \mathbb{I}_{r(X_j) \leq \lambda_1}} \sum_{i=1}^N L(Y_i, \notin \Gamma_\lambda(X_i)) \tag{A.9}$$

$$= \frac{1}{\sum_{j=1}^N \mathbb{I}_{r(X_j) \leq \lambda_1}} \sum_{i=1}^N \mathbb{I}_{Y_i \notin \Gamma_\lambda(X_i)} \mathbb{I}_{r(X_i) \leq \lambda_1} \tag{A.10}$$

and determining the p-values p_1 and p_2 such that

$$P(\hat{R}_1(\lambda_1) \geq \alpha_1) \leq p_1 \tag{A.11}$$

$$P(\hat{R}_2(\lambda_1, \lambda_2) \geq \alpha_2) \leq p_2 \tag{A.12}$$

This is done by using a concentration inequality in a manner similar to the RCPS procedure (see Section 3.3.2). We can control both risks by ensuring that $p(\lambda) =$

$\max(p_1, p_2) \leq \frac{\delta}{T}$ where $T = |\Gamma|$. This stems from the Bonferroni correction procedure applied in [3]. This ensures that, when T different values of λ are tested, the probability of choosing any λ in the set Λ that rejects any of H_0^1 or H_0^2 is at most δ. By construction $P(R_1(\lambda_1) \geq \alpha_1 \cup R_2(\lambda_1, \lambda_2) \geq \alpha_2) \leq \delta$ for any $\lambda = \begin{bmatrix}\lambda_1 \\ \lambda_2\end{bmatrix} \in \{\lambda : \lambda \in \Lambda, p(\lambda) \leq \frac{\delta}{|\Lambda|}\}$,

The guarantees in Equations A.3 and A.4 follow hence. Note that not all values of α_1, α_2, and δ are controllable - these may depend on the performance of the human / deferral policy as well as the sample size provided. For controllable risks, we may choose a particular λ depending on the deferral budget. For example, if we want to defer 20% of examples, we may choose an appropriate $\lambda \in \Lambda'$ and obtain the same guarantees. □

Theorem 3. *Given a calibration set $D_{cal} = \{X_i, Y_i\}_{i=1}^N$, the probability that any given label $y \in \mathcal{Y}$ is not in the predictive set corresponding to a given example X_{test}, i.e.*

$$P(y \notin \Gamma(X_{test})|X_{test}) \tag{A.13}$$

is bounded by:

$$P(y \notin \Gamma(X_{test})|X_{test}) \leq \sum_{i=0}^{n-1} \binom{N}{N-i} (p_t^+)^{N-i}(1-p_t^+)^i \tag{A.14}$$

where

$$p_t^+ = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{\tau(X_i,Y_i)\geq\tau(X_{test},y_{test})} + \sqrt{\frac{1}{2N}\log\frac{1}{\delta}} \tag{A.15}$$

with probability at least $1-\delta$.

Proof. We can prove this by modelling the uncertainty in the threshold τ_{cal}, which is the $n = N - \lceil(1-\alpha)(N+1)\rceil$ order statistic of the N ground truth conformity scores in the calibration dataset.

- Given X_{test} and a label y', we generate an upper confidence bound on $p_t = \mathbb{E}_{(X,Y)}[\mathbb{I}_{\tau(X,y)\geq\tau(X_{test},y')}]$. We first generate the Monte-Carlo approximation:

$$\hat{p}_t = \frac{1}{N_{val}}\sum_{i=1}^{N_{val}} \mathbb{I}_{\tau(X_i,Y_i)\geq\tau(X_{test},y')} \tag{A.16}$$

- Using this, we can bound $P(p_t \leq \hat{p}_t + \epsilon) \geq 1-\delta$ using Hoeffding's inequality (Equation 2.31). Call this $1-\delta$ UCB of p_t as p_t^+. By construction, p_t^+ has the form in Equation A.15.

- We now require the probability that the true conformity score threshold is greater than t, because if this condition holds, then the label in question y' will not belong to the CP set. This stems from the basic definition of a CP set in Equation 2.8.

- Given N calibration examples, the threshold score is the $n = N - \lceil(1-\alpha)(N+1)\rceil$ order statistic of the calibration score, i.e. the n^{th} largest ground truth conformity score in the calibration dataset.

- For an n^{th} order statistic to be greater than some $\tau(X_{test}, y')$, we need at least $N-n-1$ conformity scores to be greater than this value. This is achieved using the combinatorial expression below

$$P(\tau_{(n)} \geq \tau(X_{test}, y')) = \sum_{i=0}^{n-1} \binom{N}{N-i} p_t^{N-i}(1-p_t)^i \tag{A.17}$$

- But we just upper bounded the probability p_t that a given conformity score is greater than $\tau(X_{test}, y')$. Therefore, we get:

$$P(\tau_{(n)} \geq \tau(X_{test}, y')) = p(y' \notin \Gamma(X_{test})|X_{test}) \leq \sum_{i=0}^{n-1} \binom{N}{N-i} (p_t^+)^{N-i}(1-p_t^+)^i \tag{A.18}$$

□

B Algorithms

Algorithm 2 Inductive Conformal Prediction (adapted from [40])

Input: Trained Classifier $m_\theta(x) : \mathcal{X} \rightarrow \mathcal{Y} \cup \perp$, Calibration Set $\mathcal{D}_{\text{cal}}$, Error Tolerance α, Conformity Score Function $\tau(X, y)$, Test Example X_{test}

```
$\mathcal{S} = \varnothing$
**for** $(X_i, Y_i) \in \mathcal{D}_{\text{cal}}$ **do**
    $\mathcal{S} = \mathcal{S} \cup \tau(X_i, Y_i)$
$\tau_{cal} \rightarrow$ the $N - \lceil (1-\alpha)(N+1) \rceil$ smallest value in $\mathcal{S}$
**return** $\Gamma(X_{test}) = \{y : \tau(X_{test}, y\}$
```

Algorithm 3 Risk Controlling Predictive Sets (adapted from [9])

Input: Trained Classifier $m_\theta(x) : \mathcal{X} \rightarrow \mathcal{Y} \cup \perp$, Calibration Set $\mathcal{D}_{\text{cal}}$, Risk Tolerance γ, Excess Risk Probability δ, Conformity Score Function $\tau(X, y)$, Test Example X_{test}, Grid of parameters Λ

```
$\Lambda' = \varnothing$                                        ▷ Acceptable Parameters
**for** $\lambda \in \Lambda$ **do**
    $\Gamma_\lambda(X_i)) = \{y : \tau(X, y) \geq -\lambda\} \ \forall (X_i, Y_i) \in \mathcal{D}_{cal}$
    $\hat{R}(\lambda) = \frac{1}{N} \sum_{i=1}^{N} L(Y_i, \Gamma_\lambda(X_i))$
    $\hat{R}^+(\lambda) = \hat{R}(\lambda)$ + UCB term        ▷ Apply Hoeffding, CLT, any other conc inequality here to obtain $1-\delta$ UCB
    **if** $\hat{R}^+(\lambda) \leq \gamma$ **then**
        $\Lambda' = \Lambda' \cup \lambda$
$\hat{\lambda} = \inf \Lambda'$
**return** $\Gamma_{\hat{\lambda}} = \{y : \tau(X, y) \geq -\hat{\lambda}\}$
```

Below, we present one instance of the D-CP paradigm that was used for experiments in this thesis. Note that while the exact algorithm depends on the deferral policy being trained (while we employed the approach in [26], it is equally valid to use approaches in [28] or [41]), the main workflow followed is illustrated in Figures 4.3 and 4.4 in the paper.

Algorithm 4 D-CP

Input: Classifier $m_\theta(x) : \mathcal{X} \rightarrow \mathcal{Y} \cup \perp$, Training Set $\mathcal{D}$, Expert $h(x) \in \mathcal{Y}$, Calibration Set $\mathcal{D}_{\text{cal}}$, Validation Set $\mathcal{D}_{\text{val}}$, Error Tolerance α, Number of Epochs N, Learning Rate γ, Test Example X_{test}

```
for $i \in \{1, ...N\}$ do
    for Batch $\mathcal{B} \in \mathcal{D}$ do
        $\theta = \theta - \gamma \mathbb{E}_{(x,y)\in\mathcal{B}}[\nabla_\theta l(m_\theta(x), y, h(x))]$        ▷ Loss function in [26]
$D'_{\text{cal}} = \varnothing$
for $(X, Y) \in \mathcal{D}_{\text{cal}}$ do
    if argmax $m_\theta(X) \neq |\mathcal{Y}| + 1$ then
        $m'_\theta = \text{softmax}(m_\theta)$        ▷ Deferral Policy
        $m'_\theta = \frac{m'_\theta[1:|\mathcal{Y}|]}{1 - m'_\theta[|\mathcal{Y}|+1]}$        ▷ Bayes' Rule
        $D'_{\text{cal}} = D'_{\text{cal}} \cup (X, Y, m'_\theta(X))$
$\tau_{cal} = \alpha$ threshold conformity score learnt from conformal calibration on $\mathcal{D}'_{cal}$
if argmax $m_\theta(X_{test}) \neq |\mathcal{Y}| + 1$ then
    Output predictive set: $\Gamma(X_{test}) = \{y | y \in \mathcal{Y}, \tau(X_{test}, y) \geq \tau_{cal}\}$
else
    Defer to expert $h(X_{test})$
```

Algorithm 5 RCPS with Label Dissimilarity Minimization: Set Construction Procedure

Require: Threshold λ, Trained Model $m_\theta(X)$ with softmax probabilities $\pi_y(X)$, Dissimilarity Penalty $\mu \in [-1, 1]$, Dissimilarity Function $d(y, \Gamma_\lambda(X)) : \mathcal{Y} \times 2^{\mathcal{Y}} \rightarrow \mathcal{R}$ for a label y and set prediction $\Gamma_\lambda(X)$

```
$\Gamma_\lambda(X) \rightarrow \{\}$
$P \rightarrow \{\pi_y(X) \; \forall \; y \in \mathcal{Y}\}$
$R \rightarrow \mathcal{Y}$
while $\max(P) \geq \lambda$ do
    $\Gamma_\lambda(X) \rightarrow \Gamma_\lambda(X) \cup \text{argmax}(P)$
    $P \rightarrow P \setminus \max(P)$
    $R \rightarrow R \setminus \text{argmax}(P)$
    $p_{redistribute} \rightarrow \mu \sum_{p \in P} p$        ▷ Probability mass to redistribute amongst remaining labels
    $D_{inv} \rightarrow \sum_{y \in R} \frac{1}{d(y, \Gamma_\lambda(X))}$
                                                            ▷ Redistribute probabilities based on label similarity between the set and the incumbent label
    for $y \in R$ do
        $\Delta \rightarrow p_{redistribute} \frac{1}{d(y, \Gamma_\lambda(X))} D_{inv} - \frac{p_{redistribute}}{\text{length}(R)}$
                                                            ▷ How good is this label compared to the average remaining?
        $p \rightarrow$ Current Probability in $P$ associated with label $y$
        $p_{new} \rightarrow p + \Delta$
        $P \rightarrow P \cup p_{new} \setminus p$        ▷ Update probability associated with label $y$
return $\Gamma_\lambda(X)$
```

Bibliography

[1] Profiles of Individual Radicalization in the United States (PIRUS), 2018. https://www.start.umd.edu/data-tools/profiles-individual-radicalization-united-states-pirus.

[2] Anastasios N Angelopoulos and Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. *arXiv preprint arXiv:2107.07511*, 2021.

[3] Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control. *CoRR*, abs/2110.01052, 2021.

[4] Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty Sets for Image Classifiers Using Conformal Prediction. In *International Conference on Learning Representations*, 2020.

[5] Amina Asif and Fayyaz A. Minhas. Generalized Learning with Rejection for Classification and Regression Problems. *ArXiv*, abs/1911.00896, 2019.

[6] Varun Babbar, Umang Bhatt, and Adrian Weller. On the Utility of Prediction Sets in Human-AI Teams. *ArXiv*, abs/2205.01411, 2022.

[7] Gagan Bansal, Besmira Nushi, Ece Kamar, Dan Weld, Walter Lasecki, and Eric Horvitz. A Case for Backward Compatibility for Human-AI Teams. *ArXiv*, abs/1906.01148, 6 2019.

[8] Gagan Bansal, Tongshuang Wu, and Joyce Zhou. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. Association for Computing Machinery, 5 2021.

[9] Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-Free, Risk-Controlling Prediction Sets. *J. ACM*, 68(6), sep 2021.

[10] Anthony Bellotti. Constructing Normalized Nonconformity Measures Based on Maximizing Predictive Efficiency. In Alexander Gammerman, Vladimir Vovk, Zhiyuan Luo, Evgueni Smirnov, and Giovanni Cherubin, editors, *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 128 of *Proceedings of Machine Learning Research*, pages 41–54. PMLR, 09–11 Sep 2020.

[11] Anthony Bellotti. Optimized Conformal Classification Using Gradient Descent Approximation. *CoRR*, abs/2105.11255, 2021.

[12] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. *Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty*, page 401–413. Association for Computing Machinery, New York, NY, USA, 2021.

[13] Elizabeth Bondi, Raphael Koster, Hannah Sheahan, Martin Chadwick, Yoram Bachrach, Taylan Cemgil, Ulrich Paquet, and Krishnamurthy Dvijotham. Role of Human-AI Interaction in Selective Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[14] Nontawat Charoenphakdee, Zhenghang Cui, Yivan Zhang, and Masashi Sugiyama. Classification with Rejection Based on Cost-Sensitive Classification. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1507–1517. PMLR, 18–24 Jul 2021.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[16] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017.

[17] Simon Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, 1994.

[18] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *FAccT*, 2021.

[19] Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. *Towards Unbiased and Accurate Deferral to Multiple Experts*, page 154–165. Association for Computing Machinery, New York, NY, USA, 2021.

[20] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[21] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. CIFAR-10 (Canadian Institute for Advanced Research).

[22] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. CIFAR-100 (Canadian Institute for Advanced Research).

[23] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-Free Predictive Inference For Regression. *Journal of the American Statistical Association*, 113:1094–1111, 4 2016.

[24] Charles Lu, Andreanne Lemay, Ken Chang, Katharina Hoebel, and Jayashree Kalpathy-Cramer. Fair Conformal Predictors for Applications in Medical Imaging. *ArXiv*, abs/2109.04392, 2021.

[25] David Madras, Toniann Pitassi, and Richard Zemel. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6150–6160, 2018.

[26] Hussein Mozannar and David Sontag. Consistent Estimators for Learning to Defer to an Expert. In *International Conference on Machine Learning*, pages 7076–7087. PMLR, 2020.

[27] Chenri Ni, Nontawat Charoenphakdee, Junya Honda, and Masashi Sugiyama. *On the Calibration of Multiclass Classification with Rejection*. Curran Associates Inc., NY, USA, 2019.

[28] Nastaran Okati, Abir De, and Manuel Gomez-Rodriguez. Differentiable Learning Under Triage. In *Advances in Neural Information Processing Systems*, 2021.

[29] Joshua Peterson, Ruairidh Battleday, Thomas Griffiths, and Olga Russakovsky. Human Uncertainty Makes Classification More Robust. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9616–9625, 2019.

[30] Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[31] P. Jonathon Phillips, Amy N. Yates, Ying Hu, Carina A. Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G. Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, Jun-Cheng Chen, Carlos D. Castillo, Rama Chellappa, David White, and Alice J. O'Toole. Face Recognition Accuracy of Forensic Examiners, Superrecognizers, and Face Recognition Algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, May 2018.

[32] Ramya Ramakrishnan, Ece Kamar, Besmira Nushi, Debadeepta Dey, Julie Shah, and Eric Horvitz. Overcoming Blind Spots in the Real World: Leveraging Complementary Abilities for Joint Execution. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:6137–6145, July 2019.

[33] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with Valid and Adaptive Coverage. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3581–3591. Curran Associates, Inc., 2020.

[34] Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least Ambiguous Set-Valued Classifiers with Bounded Error Levels. *Journal of the American Statistical Association*, 114:223–234, 9 2016.

[35] Ho Chit Siu, Jaime Daniel Pena, Edenna Chen, Yutai Zhou, Victor Lopez, Kyle Palko, Kimberlee Chestnut Chang, and Ross Emerson Allen. Evaluation of human-AI teams for learned and rule-based agents in Hanabi. In *NeurIPS*, 2021.

[36] Eleni Straitouri, Lequn Wang, Nastaran Okati, and Manuel Gomez Rodriguez. Provably Improving Expert Predictions with Conformal Prediction. *CoRR*, abs/2201.12006, 2022.

[37] David Stutz, Krishnamurthy Dj Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning Optimal Conformal Classifiers. In *ICLR*, 2022.

[38] Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candes, and Aaditya Ramdas. Conformal Prediction Under Covariate Shift. *Advances in Neural Information Processing Systems*, 32, 4 2019.

[39] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 32–42, 2021.

[40] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 01 2005.

[41] Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to Complement Humans. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021.

[42] Joy T. Wu, Ken C. L. Wong, Yaniv Gur, Nadeem Ansari, Alexandros Karargyris, Arjun Sharma, Michael Morris, Babak Saboury, Hassan Ahmad, Orest Boyko, Ali Syed, Ashutosh Jadhav, Hongzhi Wang, Anup Pillai, Satyananda Kashyap, Mehdi Moradi, and Tanveer Syeda-Mahmood. Comparison of Chest Radiograph Interpretations by Artificial Intelligence Algorithm vs Radiology Residents. *JAMA Network Open*, 3(10):e2022779, October 2020.

[43] Sergey Zagoruyko and Nikos Komodakis. Wide Residual Networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016.

[44] Aleš Završnik. Criminal Justice, Artificial Intelligence Systems, and Human Rights. *ERA Forum*, 20(4):567–583, February 2020.

Risk Assessment

- As the project was entirely computational, the only risks were eyestrain from excessive screen-time and repetitive strain injury from excessive typing. These were mitigated by taking regular breaks, maintaining good posture, and through regular exercise. Furthermore, the workstation was set up to conform with the Display Screen Equipment regulations.
- Covid-19 had little to no impact on my ability to work on this project. I was able to meet my supervisor both virtually and in person, as permitted by prevailing government regulations regarding Covid.