

From Credal Learning to Learning from Credence

How to build uncertainty-aware predictors?

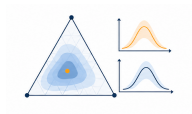
Siu Lun (Alan) Chau

Epistemic Intelligence & Computation Lab

College of Computing and Data Science, Nanyang Technological University

Wednesday, 14 October 2026

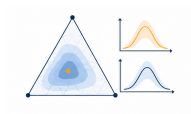
Our goal: AI that reasons about and communicates the limits of its knowledge.



Mathematics

Represent uncertainty

Our goal: AI that reasons about and communicates the limits of its knowledge.



Mathematics

Represent uncertainty

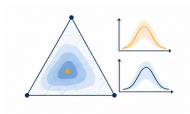


Algorithms

Learn with uncertainty



Our goal: AI that reasons about and communicates the limits of its knowledge.



Mathematics

Represent uncertainty



Algorithms

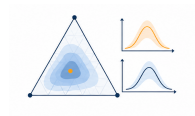
Learn with uncertainty



Applications

Support reliable decisions

Our goal: AI that reasons about and communicates the limits of its knowledge.



Mathematics

Represent uncertainty



Algorithms

Learn with uncertainty



Applications

Support reliable decisions

Our team



**Siu Lun
Chau**
PI



**Kaizheng
Wang**
Postdoc



**Clayton
Chong**
PhD student



**Joshua Adrian
Cahyono**
PhD student



**Yuqi
Zhang**
PhD student
HKU



**Xabier
Gonzalez-Garcia**
PhD visitor
UPNA

Section 1

Predictive uncertainty

The same average can hide different knowledge

Both groups report 50% for class A. Only one group agrees.

Probability of class A	#1: Agreement about ambiguity	
Model 1	0.50	
Model 2	0.50	
Model 3	0.50	
Model 4	0.50	
Average	0.50	

The same average can hide different knowledge

Both groups report 50% for class A. Only one group agrees.

Probability of class A	#1: Agreement about ambiguity	#2: Model Disagreement
Model 1	0.50	0.05
Model 2	0.50	0.05
Model 3	0.50	0.95
Model 4	0.50	0.95
Average	0.50	0.50

The same average can hide different knowledge

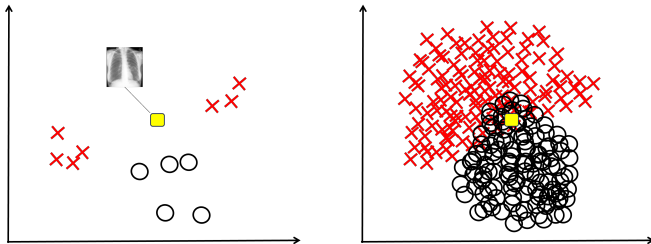
Both groups report 50% for class A. Only one group agrees.

Probability of class A	#1: Agreement about ambiguity	#2: Model Disagreement
Model 1	0.50	0.05
Model 2	0.50	0.05
Model 3	0.50	0.95
Model 4	0.50	0.95
Average	0.50	0.50

The full picture tells us something the average cannot.

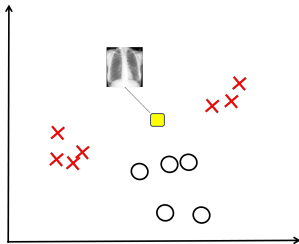
Outcome variability and incomplete knowledge

More data can clarify a relationship while overlapping classes remain hard to distinguish.



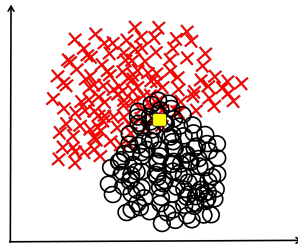
Outcome variability and incomplete knowledge

More data can clarify a relationship while overlapping classes remain hard to distinguish.



Epistemic uncertainty

Incomplete knowledge of the relationship.
Relevant evidence may reduce it.



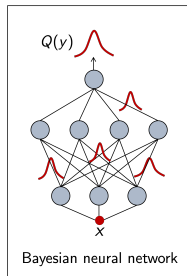
Aleatoric uncertainty

Variability in the outcome, given the task and available measurements.

Bayesian neural networks

Keep models that remain plausible after seeing the data.

- ▶ Express uncertainty about network weights θ through a starting distribution: the **prior**.
- ▶ After observing data D , the **posterior** describes their updated plausibility.



Bayesian neural networks

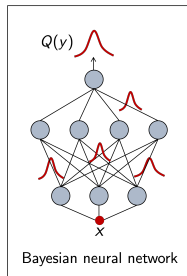
Keep models that remain plausible after seeing the data.

- ▶ Express uncertainty about network weights θ through a starting distribution: the **prior**.
- ▶ After observing data D , the **posterior** describes their updated plausibility.

$$p(\theta \mid D) \propto p(D \mid \theta) p(\theta).$$

The data fit updates the prior assumptions.

Combine predictions across plausible parameter settings.



Bayesian neural networks

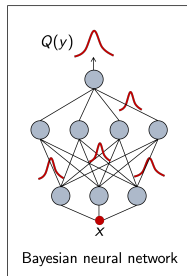
Keep models that remain plausible after seeing the data.

- ▶ Express uncertainty about network weights θ through a starting distribution: the **prior**.
- ▶ After observing data D , the **posterior** describes their updated plausibility.

$$p(\theta \mid D) \propto p(D \mid \theta) p(\theta).$$

The data fit updates the prior assumptions.

Combine predictions across plausible parameter settings.

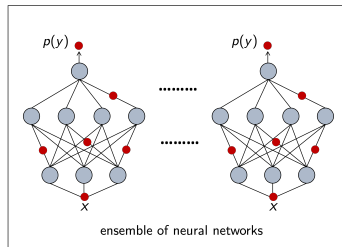


Takeaway. Exact posterior inference is generally intractable for large networks. Practical approximations add computation and design choices.

Deep ensembles

Repeat standard training to obtain several predictions, without fitting a posterior over weights.

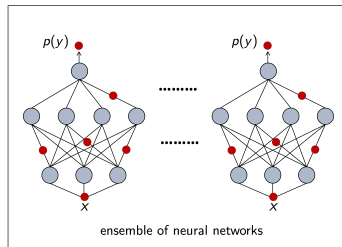
- ▶ Train networks independently from different random initialisations.



Deep ensembles

Repeat standard training to obtain several predictions, without fitting a posterior over weights.

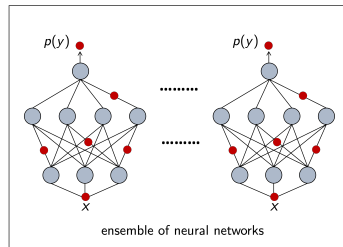
- ▶ Train networks independently from different random initialisations.
- ▶ Average their class probabilities for a single forecast.
- ▶ Retain their disagreement as an uncertainty signal.



Deep ensembles

Repeat standard training to obtain several predictions, without fitting a posterior over weights.

- ▶ Train networks independently from different random initialisations.
- ▶ Average their class probabilities for a single forecast.
- ▶ Retain their disagreement as an uncertainty signal.



Takeaway. Standard training and parallelisation make ensembles attractive. Several networks still cost more than one ordinary network.

Separately fitted members are not automatically Bayesian posterior samples.

The source of disagreement matters

Variation becomes useful when we can explain which uncertainty it explores.

Approach	Meaning of the alternatives
Bayesian neural network	Parameter plausibility under a prior, a model and the observed data
Deep ensemble	Fitted solutions induced by the chosen training procedure

The source of disagreement matters

Variation becomes useful when we can explain which uncertainty it explores.

Approach	Meaning of the alternatives
Bayesian neural network	Parameter plausibility under a prior, a model and the observed data
Deep ensemble	Fitted solutions induced by the chosen training procedure

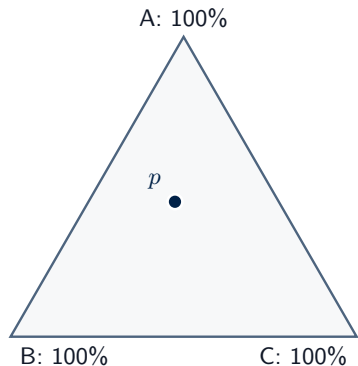
Takeaway. Models can miss possibilities their construction never explores.

A set of plausible probabilities

A point commits to one forecast; a region retains several.

Three nonnegative probabilities that sum to one form a triangle: the **probability simplex**.

One forecast: $p = (0.45, 0.30, 0.25)$.



A set of plausible probabilities

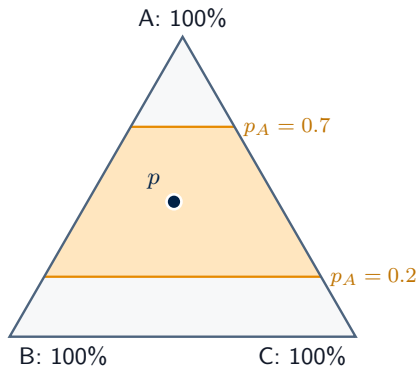
A point commits to one forecast; a region retains several.

Three nonnegative probabilities that sum to one form a triangle: the **probability simplex**.

One forecast: $p = (0.45, 0.30, 0.25)$.

Probability bounds

Class	Lower	Upper
A	0.2	0.7



A set of plausible probabilities

A point commits to one forecast; a region retains several.

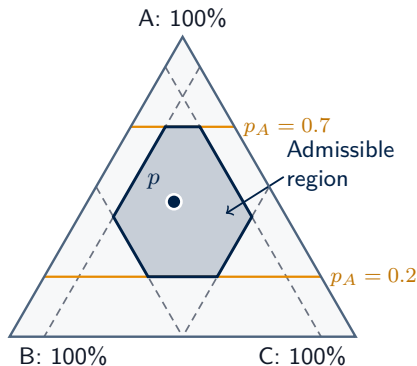
Three nonnegative probabilities that sum to one form a triangle: the **probability simplex**.

One forecast: $p = (0.45, 0.30, 0.25)$.

Probability bounds

Class	Lower	Upper
A	0.2	0.7
B	0.1	0.5
C	0.1	0.5

A **credal set** retains several distributions.
These bounds define a **credal box**.



Disagreement about deployment conditions

What if different models protected against different changes in the data distribution?

- ▶ A new laboratory may encounter different proportions of difficult cases.
- ▶ A training average may underweight cases that matter there.
- ▶ We can vary how strongly training protects against this possibility.

Disagreement about deployment conditions

What if different models protected against different changes in the data distribution?

- ▶ A new laboratory may encounter different proportions of difficult cases.
- ▶ A training average may underweight cases that matter there.
- ▶ We can vary how strongly training protects against this possibility.

Takeaway. CreDRO makes different robustness assumptions produce different predictions.

Section 2

Credal learning

When deployment changes which cases matter

Good average performance can hide failures on cases that become more common.

An illustrative change in case mix

Cases	Training	Deployment
Familiar	80%	50%
Difficult	20%	50%

The same weaknesses now matter more.

When deployment changes which cases matter

Good average performance can hide failures on cases that become more common.

An illustrative change in case mix

Cases	Training	Deployment
Familiar	80%	50%
Difficult	20%	50%

The same weaknesses now matter more.

- ▶ A **loss** is a penalty for a poor prediction: larger means worse.
- ▶ Ordinary training minimizes the average loss on observed cases.
- ▶ Can we train for plausible changes in their importance?

Train for plausible changes in importance

Choose a model that performs well even when the data receive less favourable weights.

Distributionally robust optimisation (DRO) considers a chosen family of alternative data distributions.

- ▶ θ : neural-network parameters; ℓ : the prediction loss.
- ▶ w_n : the importance of training example n .
- ▶ W : the allowed changes in importance: our modelling assumption.

Train for plausible changes in importance

Choose a model that performs well even when the data receive less favourable weights.

Distributionally robust optimisation (DRO) considers a chosen family of alternative data distributions.

- ▶ θ : neural-network parameters; ℓ : the prediction loss.
- ▶ w_n : the importance of training example n .
- ▶ W : the allowed changes in importance: our modelling assumption.

$$\underbrace{\min_{\theta}}_{\text{choose a model}} \quad \underbrace{\max_{w \in W}}_{\text{choose difficult weights}} \quad \frac{1}{N} \sum_{n=1}^N w_n \ell(h_{\theta}(x_n), y_n).$$

A simple knob: emphasize the hardest fraction

A controlled limit on each example's influence makes the inner problem simple. Use $0 < \delta \leq 1$.

$$W_\delta = \left\{ w \geq 0 : \sum_{n=1}^N w_n = N, \quad w_n \leq \frac{1}{\delta} \right\}.$$

- ▶ $\delta = 1$: every example has weight one; ordinary average-loss training.
- ▶ Smaller δ : more emphasis can fall on high-loss examples.

A simple knob: emphasize the hardest fraction

A controlled limit on each example's influence makes the inner problem simple. Use $0 < \delta \leq 1$.

$$W_\delta = \left\{ w \geq 0 : \sum_{n=1}^N w_n = N, \quad w_n \leq \frac{1}{\delta} \right\}.$$

- ▶ $\delta = 1$: every example has weight one; ordinary average-loss training.
- ▶ Smaller δ : more emphasis can fall on high-loss examples.
- ▶ For a batch of B examples: sort losses, retain the top $\lfloor \delta B \rfloor$, and train on their mean.

Averaging the worst fraction is the **conditional value-at-risk** idea.

Use at least one selected example. This batch rule approximates the full-data objective.

The same batch, a different training emphasis

We vary which errors drive learning, using their loss values as the guide.

Example	A	B	C	D	E	F
Loss	0.1	0.2	0.4	0.8	1.1	1.6
All examples, $\delta = 1$	✓	✓	✓	✓	✓	✓
Hardest half, $\delta = 0.5$				✓	✓	✓

The same batch, a different training emphasis

We vary which errors drive learning, using their loss values as the guide.

Example	A	B	C	D	E	F
Loss	0.1	0.2	0.4	0.8	1.1	1.6
All examples, $\delta = 1$	✓	✓	✓	✓	✓	✓
Hardest half, $\delta = 0.5$				✓	✓	✓

Average over all six: 0.70.

Average over the hardest three: 1.17.

The larger number is a different training objective, not a measured performance improvement.

CreDRO keeps several robustness levels

Let disagreement reflect how strongly we relax equal importance in training.

Five models, five choices

$$\delta_i \in \{0.5, 0.625, 0.75, 0.875, 1\}.$$

CreDRO keeps several robustness levels

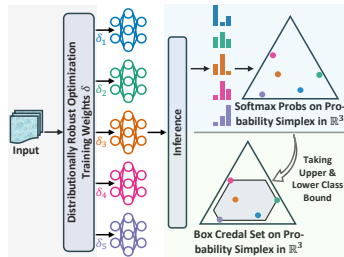
Let disagreement reflect how strongly we relax equal importance in training.

Five models, five choices

$$\delta_i \in \{0.5, 0.625, 0.75, 0.875, 1\}.$$

- ▶ Includes ordinary training.
- ▶ Select losses separately for each model.
- ▶ Keep each prediction at inference.

The range is chosen by the modeller.



Keep compatible probabilities as a set

Averaging gives one answer; probability ranges preserve what the models disagree about.

Prediction	A	B	C
Model 1	0.6	0.3	0.1
Model 2	0.4	0.4	0.2
Model 3	0.5	0.2	0.3
Lower	0.4	0.2	0.1
Upper	0.6	0.4	0.3

Keep compatible probabilities as a set

Averaging gives one answer; probability ranges preserve what the models disagree about.

Prediction	A	B	C
Model 1	0.6	0.3	0.1
Model 2	0.4	0.4	0.2
Model 3	0.5	0.2	0.3
Lower	0.4	0.2	0.1
Upper	0.6	0.4	0.3

A credal set is a set of probabilities.

- ▶ Take each class's smallest and largest predicted probability.
- ▶ Keep all vectors within those bounds that sum to one.
- ▶ The bounds need not occur together in one prediction.

Keep compatible probabilities as a set

Averaging gives one answer; probability ranges preserve what the models disagree about.

Prediction	A	B	C
Model 1	0.6	0.3	0.1
Model 2	0.4	0.4	0.2
Model 3	0.5	0.2	0.3
Lower	0.4	0.2	0.1
Upper	0.6	0.4	0.3

A credal set is a set of probabilities.

- ▶ Take each class's smallest and largest predicted probability.
- ▶ Keep all vectors within those bounds that sum to one.
- ▶ The bounds need not occur together in one prediction.

Here $(0.6, 0.2, 0.2)$ is also compatible with the bounds.

Do unfamiliar inputs receive higher uncertainty?

DRO training helps here. Credal representation and scoring improve the ranking further.

OOD detection: rank inputs from a different data source above familiar inputs.

AUROC: a ranking score; 50% is chance and 100% is perfect.

Do unfamiliar inputs receive higher uncertainty?

DRO training helps here. Credal representation and scoring improve the ranking further.

OOD detection: rank inputs from a different data source above familiar inputs.

AUROC: a ranking score; 50% is chance and 100% is perfect.

Method	SVHN	Places	CIFAR100	FMNIST	ImageNet
DE	94.8 ± 0.3	90.0 ± 0.2	90.6 ± 0.0	92.9 ± 0.3	88.9 ± 0.1
EN-DRO	95.7 ± 0.0	91.1 ± 0.1	91.6 ± 0.1	94.0 ± 0.1	90.0 ± 0.1
CreDRO	97.4 ± 0.1	92.7 ± 0.1	92.5 ± 0.1	96.4 ± 0.0	91.1 ± 0.1

CIFAR10 (in-distribution); ResNet18; 20 members. Mean $\pm$ reported variability over 3 runs.

- ▶ **EN-DRO:** the same DRO training, with ordinary ensemble uncertainty.
- ▶ **CreDRO:** retain probability bounds and use a credal uncertainty score.

Can uncertainty help decide when to defer?

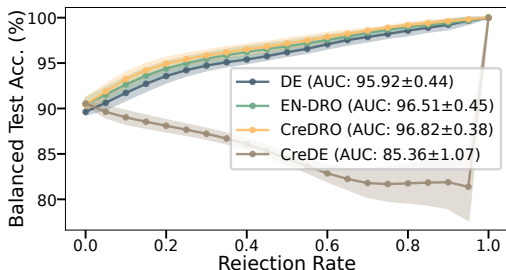
Under a change of hospital and scanner, set aside the cases with highest uncertainty.

Camelyon17: tumour detection

- ▶ Train and test on different clinic/scanner groups.
- ▶ Horizontal: fraction deferred.
- ▶ Vertical: balanced accuracy on retained cases.

Can uncertainty help decide when to defer?

Under a change of hospital and scanner, set aside the cases with highest uncertainty.



Camelyon17: tumour detection

- ▶ Train and test on different clinic/scanner groups.
- ▶ Horizontal: fraction deferred.
- ▶ Vertical: balanced accuracy on retained cases.

Area under the curve:

DE 95.92; EN-DRO 96.51;

CreDRO 96.82.

Here uncertainty is the width of the binary probability interval.

The 100%-deferral endpoint is assigned 100% accuracy by convention.

ResNet34; 5 ensemble members; averages over 5 runs.

When uncertainty comes from the annotator

CreDRO makes a modelling doubt explicit: which training cases should matter more?

- ▶ **Contribution:** train across a chosen range of robustness levels and retain the predictive disagreement.
- ▶ **Evidence:** useful uncertainty rankings in the reported detection and deferral tasks.
- ▶ **Limits:** the reweighting family is chosen; the output intervals have no proved guarantee of containing the true probabilities.

When uncertainty comes from the annotator

CreDRO makes a modelling doubt explicit: which training cases should matter more?

- ▶ **Contribution:** train across a chosen range of robustness levels and retain the predictive disagreement.
- ▶ **Evidence:** useful uncertainty rankings in the reported detection and deferral tasks.
- ▶ **Limits:** the reweighting family is chosen; the output intervals have no proved guarantee of containing the true probabilities.

The next source of information

What if the annotator can tell us which labels or probabilities remain plausible?

Section 3

Learning from credence

Non-hard labels as information

Many non-hard labels hint at which labelling distributions are plausible.

Source

What information it provides

CIFAR-10H

Human votes

Each person chooses one class. Pooled votes describe how people distribute their support across classes.

Non-hard labels as information

Many non-hard labels hint at which labelling distributions are plausible.

Source	What information it provides
CIFAR-10H Human votes	Each person chooses one class. Pooled votes describe how people distribute their support across classes.
CIFAR-10S Individual judgments	People report class probabilities and ruled-out classes, expressing support for alternatives and exclusions.

Non-hard labels as information

Many non-hard labels hint at which labelling distributions are plausible.

Source	What information it provides
CIFAR-10H Human votes	Each person chooses one class. Pooled votes describe how people distribute their support across classes.
CIFAR-10S Individual judgments	People report class probabilities and ruled-out classes, expressing support for alternatives and exclusions.
Model distillation Teacher predictions	A teacher's class probabilities give the student information about relative support across classes.

Takeaway. Non-hard labels can inform constraints on plausible labelling distributions.

Representing labelling information

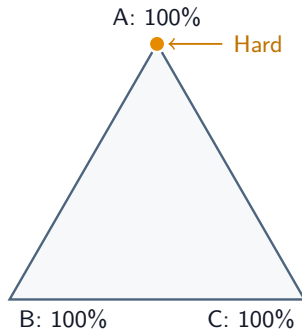
We can represent labelling information as one distribution or a set of plausible distributions.

Credence: evidence-supported degrees of belief.

Ten annotations: classes **A** / **B** / **C**

Evidence	Votes	Majority
Unanimous	10/0/0	A
Divided	7/2/1	A

► **Hard label:** one vertex, $(1, 0, 0)$.



Representing labelling information

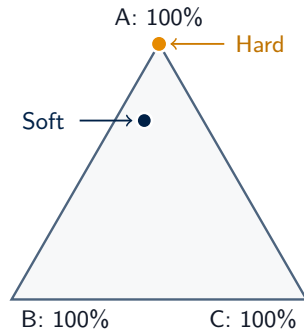
We can represent labelling information as one distribution or a set of plausible distributions.

Credence: evidence-supported degrees of belief.

Ten annotations: classes A / B / C

Evidence	Votes	Majority
Unanimous	10/0/0	A
Divided	7/2/1	A

- ▶ **Hard label:** one vertex, $(1, 0, 0)$.
- ▶ **Soft label:** one point, $(0.7, 0.2, 0.1)$.



Representing labelling information

We can represent labelling information as one distribution or a set of plausible distributions.

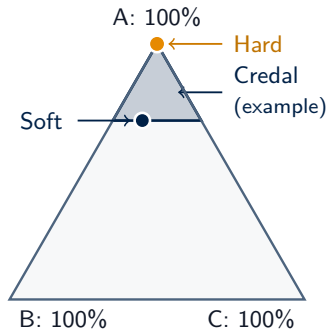
Credence: evidence-supported degrees of belief.

Ten annotations: classes A / B / C

Evidence	Votes	Majority
Unanimous	10/0/0	A
Divided	7/2/1	A

- ▶ **Hard label:** one vertex, $(1, 0, 0)$.
- ▶ **Soft label:** one point, $(0.7, 0.2, 0.1)$.
- ▶ **Credal label:** a set of probability vectors.

A credal label makes the remaining possibilities explicit.



Retain a constraint, relax the rest

One modelling choice: trust a lower bound for class A and leave the rest unresolved.

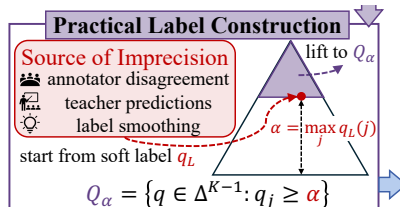
Start from $q_L = (0.7, 0.2, 0.1)$. Interpret its largest entry as a lower bound:

Retain a constraint, relax the rest

One modelling choice: trust a lower bound for class A and leave the rest unresolved.

Start from $q_L = (0.7, 0.2, 0.1)$. Interpret its largest entry as a lower bound:

$$Q_{0.7} = \{q : q_A \geq 0.7, q_A + q_B + q_C = 1, q_k \geq 0\}.$$



Every point in the triangle is a class-probability vector.

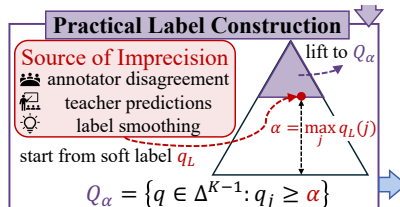
Retain a constraint, relax the rest

One modelling choice: trust a lower bound for class A and leave the rest unresolved.

Start from $q_L = (0.7, 0.2, 0.1)$. Interpret its largest entry as a lower bound:

$$Q_{0.7} = \{q : q_A \geq 0.7, q_A + q_B + q_C = 1, q_k \geq 0\}.$$

- ▶ Admits $(0.7, 0.2, 0.1)$, $(0.7, 0.3, 0)$ and $(1, 0, 0)$.
- ▶ Reserve 0.7 for A; leave 0.3 free.
- ▶ Contains the reference label, not automatically the truth.



Every point in the triangle is a class-probability vector.

One prediction, a range of possible losses

Unresolved targets create a best case and a worst case for the same prediction.

- ▶ A loss measures the cost of a prediction; smaller is better.
- ▶ **Cross-entropy** penalises assigning little probability to classes supported by the target.
- ▶ Different admissible targets $q \in Q(x)$ can give different losses.

One prediction, a range of possible losses

Unresolved targets create a best case and a worst case for the same prediction.

- ▶ A loss measures the cost of a prediction; smaller is better.
- ▶ **Cross-entropy** penalises assigning little probability to classes supported by the target.
- ▶ Different admissible targets $q \in Q(x)$ can give different losses.

$$\ell(p, q) = - \sum_k q_k \log p_k, \quad \left[\underbrace{\min_{q \in Q(x)} \ell(p, q)}_{\text{best case}}, \underbrace{\max_{q \in Q(x)} \ell(p, q)}_{\text{worst case}} \right].$$

One prediction, a range of possible losses

Unresolved targets create a best case and a worst case for the same prediction.

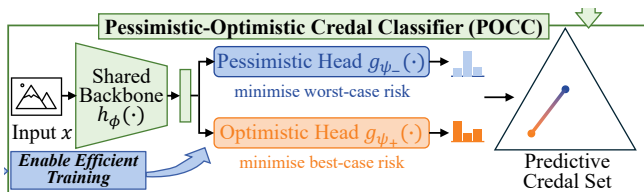
- ▶ A loss measures the cost of a prediction; smaller is better.
- ▶ **Cross-entropy** penalises assigning little probability to classes supported by the target.
- ▶ Different admissible targets $q \in Q(x)$ can give different losses.

$$\ell(p, q) = - \sum_k q_k \log p_k, \quad \left[\underbrace{\min_{q \in Q(x)} \ell(p, q)}_{\text{best case}}, \underbrace{\max_{q \in Q(x)} \ell(p, q)}_{\text{worst case}} \right].$$

Takeaway. Keep both criteria: they reveal different consequences of the evidence.

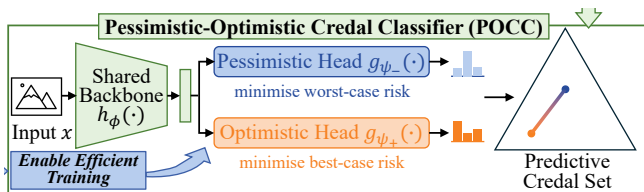
Two heads learn complementary responses

One shared representation supports a cautious prediction and a favourable interpretation.



Two heads learn complementary responses

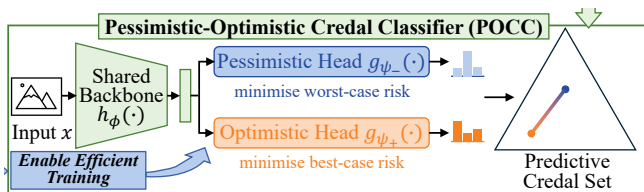
One shared representation supports a cautious prediction and a favourable interpretation.



- **Pessimistic head:** minimise the worst admissible loss.
- **Optimistic head:** minimise the best admissible loss.

Two heads learn complementary responses

One shared representation supports a cautious prediction and a favourable interpretation.



- **Pessimistic head:** minimise the worst admissible loss.
- **Optimistic head:** minimise the best admissible loss.

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N \left[\max_{q \in Q_i} \ell(p_{-}(x_i), q) + \min_{q \in Q_i} \ell(p_{+}(x_i), q) \right].$$

The inner problem is just moving mass

For cross-entropy, choosing an extreme target requires only finding one class.

Pessimistic head

Prediction: $p_- = (0.6, 0.2, 0.2)$.

- ▶ Reserve 0.7 for A.
- ▶ Put the free 0.3 on a least probable class, here B or C.

$$q_-^* = (0.7, 0.3, 0).$$

The inner problem is just moving mass

For cross-entropy, choosing an extreme target requires only finding one class.

Pessimistic head

Prediction: $p_- = (0.6, 0.2, 0.2)$.

- ▶ Reserve 0.7 for A.
- ▶ Put the free 0.3 on a least probable class, here B or C.

$$q_-^* = (0.7, 0.3, 0).$$

Optimistic head

Prediction: $p_+ = (0.9, 0.05, 0.05)$.

- ▶ Reserve 0.7 for A.
- ▶ Put the free 0.3 on the most probable class, here A.

$$q_+^* = (1, 0, 0).$$

The inner problem is just moving mass

For cross-entropy, choosing an extreme target requires only finding one class.

Pessimistic head

Prediction: $p_- = (0.6, 0.2, 0.2)$.

- ▶ Reserve 0.7 for A.
- ▶ Put the free 0.3 on a least probable class, here B or C.

$$q_-^* = (0.7, 0.3, 0).$$

Optimistic head

Prediction: $p_+ = (0.9, 0.05, 0.05)$.

- ▶ Reserve 0.7 for A.
- ▶ Put the free 0.3 on the most probable class, here A.

$$q_+^* = (1, 0, 0).$$

Takeaway. One scan over the classes; no iterative inner optimisation.

Keep the predictions and their disagreement

Use the midpoint for decisions and head separation for uncertainty.

Prediction	Class A	Class B	Class C
Pessimistic p_-	.60	.20	.20
Midpoint $\bar{p}$	.75	.125	.125
Optimistic p_+	.90	.05	.05

Keep the predictions and their disagreement

Use the midpoint for decisions and head separation for uncertainty.

Prediction	Class A	Class B	Class C
Pessimistic p_-	.60	.20	.20
Midpoint $\bar{p}$	.75	.125	.125
Optimistic p_+	.90	.05	.05

- ▶ Every mixture between the two heads belongs to the predictive set.
- ▶ **Total variation (TV):** the probability mass moved between the two vectors.

Keep the predictions and their disagreement

Use the midpoint for decisions and head separation for uncertainty.

Prediction	Class A	Class B	Class C
Pessimistic p_-	.60	.20	.20
Midpoint $\bar{p}$	.75	.125	.125
Optimistic p_+	.90	.05	.05

- ▶ Every mixture between the two heads belongs to the predictive set.
- ▶ **Total variation (TV):** the probability mass moved between the two vectors.

$$\text{TV}(p_-, p_+) = \frac{1}{2} \sum_k |p_{-,k} - p_{+,k}| = 0.3.$$

Valid constraints support useful prediction

Valid annotation constraints can support a useful midpoint prediction.

- ▶ **Validity:** the true conditional distribution belongs to $Q(x)$; here $p_{0,A}(x) \geq \alpha(x)$.
- ▶ Fixed supervision rule, independent inputs, bounded losses and approximate training optimum.
- ▶ Larger α narrows the set, but requires a stronger valid lower bound.

Excess loss compares with the best predictor representable by identical heads.

Valid constraints support useful prediction

Valid annotation constraints can support a useful midpoint prediction.

- ▶ **Validity:** the true conditional distribution belongs to $\mathcal{Q}(x)$; here $p_{0,A}(x) \geq \alpha(x)$.
- ▶ Fixed supervision rule, independent inputs, bounded losses and approximate training optimum.
- ▶ Larger α narrows the set, but requires a stronger valid lower bound.

Excess loss compares with the best predictor representable by identical heads.

$$\begin{aligned} \text{Excess midpoint loss} &\leq \text{model capacity} + \text{sampling error} \\ &\quad + \text{optimisation error} + \text{supervision imprecision.} \end{aligned}$$

Valid constraints support useful prediction

Valid annotation constraints can support a useful midpoint prediction.

- ▶ **Validity:** the true conditional distribution belongs to $\mathcal{Q}(x)$; here $p_{0,A}(x) \geq \alpha(x)$.
- ▶ Fixed supervision rule, independent inputs, bounded losses and approximate training optimum.
- ▶ Larger α narrows the set, but requires a stronger valid lower bound.

Excess loss compares with the best predictor representable by identical heads.

$$\begin{aligned} \text{Excess midpoint loss} &\leq \text{model capacity} + \text{sampling error} \\ &\quad + \text{optimisation error} + \text{supervision imprecision.} \end{aligned}$$

Takeaway. The guarantee concerns midpoint risk. Predictive-set coverage needs separate analysis.

Human disagreement: a useful balance

On CIFAR-10H, POCC combines competitive accuracy, calibration and error ranking.

- ▶ **Calibration error (ECE):** sample-weighted absolute confidence–accuracy gaps across bins.
- ▶ **Deferral score (AUARC):** area under accuracy as uncertain cases are rejected.

Human disagreement: a useful balance

On CIFAR-10H, POCC combines competitive accuracy, calibration and error ranking.

- ▶ **Calibration error (ECE):** sample-weighted absolute confidence–accuracy gaps across bins.
- ▶ **Deferral score (AUARC):** area under accuracy as uncertain cases are rejected.

Method	Accuracy $\uparrow$	Calibration error $\downarrow$	Deferral score $\uparrow$
POCC	87.0 ± 0.4	2.9 ± 0.3	97.1 ± 0.2
Deep ensemble (5)	87.0 ± 0.3	4.3 ± 0.2	96.9 ± 0.1
Dropout (5 passes)	85.6 ± 0.7	0.9 ± 0.1	96.6 ± 0.3
DAPPr	86.6 ± 0.3	6.9 ± 0.2	97.2 ± 0.1

Percentages; mean $\pm$ SD; 10 runs.

- ▶ Other methods remain better on individual metrics.

A teacher can supply imprecise supervision

Relaxing a teacher's probability assignments can also help the student learn.

- ▶ CIFAR-100; ResNet-20 teacher; WRN-28-4 student.
- ▶ Teacher probabilities softened with temperature $T = 2.5$.
- ▶ The same reference targets feed all methods; POCC relaxes them into sets.

A teacher can supply imprecise supervision

Relaxing a teacher's probability assignments can also help the student learn.

- ▶ CIFAR-100; ResNet-20 teacher; WRN-28-4 student.
- ▶ Teacher probabilities softened with temperature $T = 2.5$.
- ▶ The same reference targets feed all methods; POCC relaxes them into sets.

Method	Accuracy $\uparrow$	ECE $\downarrow$	AUARC $\uparrow$
POCC	77.2 ± 0.2	10.6 ± 0.4	90.6 ± 0.2
Deep ensemble (5)	72.7 ± 0.3	29.9 ± 0.3	86.5 ± 0.3
Dropout (5 passes)	71.1 ± 0.3	28.2 ± 0.3	83.2 ± 0.7

Percentages; mean $\pm$ SD; 10 runs.

Learning from credence opens the next questions

The credibility of supervision deserves the same attention as its imprecision.

- ▶ **Richer evidence:** the simple construction keeps only the leading class and its probability.
- ▶ **Valid constraints:** finite annotators and imperfect teachers need not supply a true lower bound.
- ▶ **Richer predictions:** two heads give a segment, not a guarantee that every plausible prediction is recovered.

Learning from credence opens the next questions

The credibility of supervision deserves the same attention as its imprecision.

- ▶ **Richer evidence:** the simple construction keeps only the leading class and its probability.
- ▶ **Valid constraints:** finite annotators and imperfect teachers need not supply a true lower bound.
- ▶ **Richer predictions:** two heads give a segment, not a guarantee that every plausible prediction is recovered.
- ▶ **Next step:** model annotation variability and test what predicted sets actually cover.

Takeaway. Useful uncertainty scores need to be accompanied by defensible supervision constraints.

Where uncertainty enters learning

The alternatives we preserve should reflect the evidence and assumptions we actually have.

Approach	Source of alternatives	What the prediction retains
Model variation	Plausible or differently fitted models	Variation across model predictions
CreDRO	Different robustness assumptions	A probability box built from an ensemble
POCC	Admissible annotation distributions	Two learned responses and their mixtures

Where uncertainty enters learning

The alternatives we preserve should reflect the evidence and assumptions we actually have.

Approach	Source of alternatives	What the prediction retains
Model variation	Plausible or differently fitted models	Variation across model predictions
CreDRO	Different robustness assumptions	A probability box built from an ensemble
POCC	Admissible annotation distributions	Two learned responses and their mixtures

Takeaway. Connect predictive uncertainty to the uncertainty in our evidence.

Questions and discussion

What does your evidence leave unresolved?

10 minutes

Joint work with Kaizheng Wang, Ghifari Adam Faza, Fabio Cuzzolin,
David Moens and Hans Hallez.

Tutorial material with Eyke Hüllermeier and Yusuf Sale.

Backup: a disagreement score

An average can be uncertain because its members are uncertain, or because they disagree.

Entropy measures how spread out a class distribution is: $H(p) = -\sum_k p_k \log_2 p_k$.

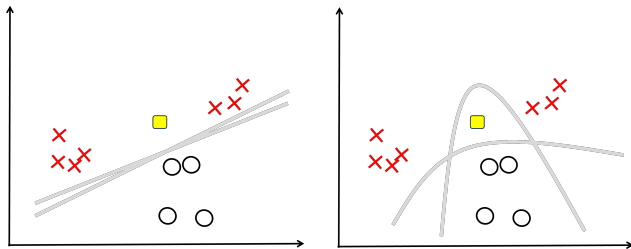
$$\underbrace{H(\bar{p})}_{\text{uncertainty of the average}} = \underbrace{\frac{1}{M} \sum_m H(p_m)}_{\text{average within-model uncertainty}} + \underbrace{\left[H(\bar{p}) - \frac{1}{M} \sum_m H(p_m) \right]}_{\text{disagreement score}}, \quad \bar{p} = \frac{1}{M} \sum_m p_m.$$

Opening example	$H(\bar{p})$	Mean $H(p_m)$	Disagreement
All models predict (.5, .5)	1 bit	1 bit	0 bits
Half (.05, .95), half (.95, .05)	1 bit	0.286 bits	0.714 bits

The identity describes the chosen collection. Its interpretation depends on its construction.

Backup: assumptions change uncertainty

More data is only one route to reducing uncertainty. Model restrictions also affect what remains plausible.



A narrow model family can agree while excluding a relevant possibility.

Backup: exact sample reweighting

The exact empirical objective can assign a fractional weight at the cutoff.

Sort losses in descending order: $\ell_{(1)} \geq \dots \geq \ell_{(N)}$. Let $a = \delta N$, $k = \lfloor a \rfloor$, and $r = a - k$.

$$\max_{\substack{w \geq 0, \\ \sum_i w_i = N \\ w_i \leq 1/\delta}} \frac{1}{N} \sum_i w_i \ell_i = \frac{\sum_{i=1}^k \ell_{(i)} + r \ell_{(k+1)}}{\delta N}.$$

For $r = 0$, omit the last term. This includes $\delta = 1$, which gives the ordinary mean.

CreDRO's implemented minibatch rule averages the top $\lfloor \delta B \rfloor$ losses. It approximates the full empirical objective.

Backup: a set needs a choice of score

Different summaries measure different aspects of the retained probabilities.

CreDRO multiclass experiments

Entropy is the amount of predictive randomness. The score measures its possible range:

$$\max_{p \in \mathcal{C}(x)} H(p) - \min_{p \in \mathcal{C}(x)} H(p).$$

The medical binary experiment instead uses probability-interval width.

Useful uncertainty ranking is an empirical question. A score alone provides no probability-coverage guarantee.

POCC

Total variation is the probability mass moved between the two head predictions:

$$\text{TV}(p_-, p_+) = \frac{1}{2} \sum_k |p_{-,k} - p_{+,k}|.$$

For the two-head segment, this is the paper's maximum mean imprecision score.

Backup: what POCC's bound establishes

Valid supervision leaves an explicit cost for its remaining imprecision.

R_0 is true expected cross-entropy, $\hat{f}_n$ the learned midpoint. With probability at least $1 - \delta$,

$$R_0(\hat{f}_n) - \inf_{f \in \mathcal{F}_\Delta} R_0(f) \leq \underbrace{4\mathfrak{R}_n(\mathcal{J}_Q)}_{\text{model capacity}} + \underbrace{2B\sqrt{\frac{\log(2/\delta)}{2n}}}_{\text{finite sample}} + \underbrace{\eta_n}_{\text{optimisation error}} + \underbrace{L\mathbb{E}[d_Q(X)]}_{\text{supervision imprecision}}.$$

- ▶ $\mathcal{F}_\Delta$: predictors representable by identical heads. $\mathfrak{R}_n$ measures ability to fit random noise, called Rademacher complexity.
- ▶ Fixed valid sets: $p_0(x) \in Q(x)$ almost surely. Independent inputs. Nonempty comparator class.
- ▶ Bounded cross-entropy B and log-loss range L . d_Q is the largest TV distance within Q . Measurability as in the theorem.

Backup: validity and precision are separate

A smaller annotation set is more informative only when its constraints remain valid.

For the chosen class j , let q range over probability vectors:

$$Q_\alpha = \{q : q_j \geq \alpha\}, \quad \text{diam}_{\text{TV}}(Q_\alpha) = 1 - \alpha.$$

- ▶ Increasing α narrows the set and reduces the imprecision term.
- ▶ Validity requires the true conditional probability $p_{0,j}(x) \geq \alpha(x)$.
- ▶ Containing the reference soft label q_L does not establish that requirement.

If all model logits have magnitude at most A , the theorem's imprecision term becomes $2A\mathbb{E}[1 - \alpha(X)]$.

References: core material

Background

Alan Chau, Eyke Hüllermeier and Yusuf Sale.

SIPTA Summer School: Machine Learning and Uncertainty Quantification. Munich, 31 July 2026.

Credal learning through robust optimisation

Kaizheng Wang, Ghifari Adam Faza, Fabio Cuzzolin, Siu Lun Chau, David Moens and Hans Hallez.

Learning Credal Ensembles via Distributionally Robust Optimization. ICML 2026. arXiv version 3.

<https://arxiv.org/abs/2602.08470>

Learning from credal supervision

Kaizheng Wang and Siu Lun Chau.

Epistemic Learning from Imprecise Annotation. 2026 preprint. arXiv version 1.

<https://arxiv.org/abs/2609.34285>

References: neural-network uncertainty

Eyke Hüllermeier and Willem Waegeman.

Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods.

Machine Learning, 110:457–506, 2021.

<https://arxiv.org/abs/1910.09457>

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu and Daan Wierstra.

Weight Uncertainty in Neural Networks. ICML, 2015.

<https://arxiv.org/abs/1505.05424>

Balaji Lakshminarayanan, Alexander Pritzel and Charles Blundell.

Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. NeurIPS, 2017.

<https://arxiv.org/abs/1612.01474>