Knowing what you don't know: Learning to abstain and beyond

Harikrishna Narasimhan

Google Research

Contributors

Wittawat
Jitkrittum

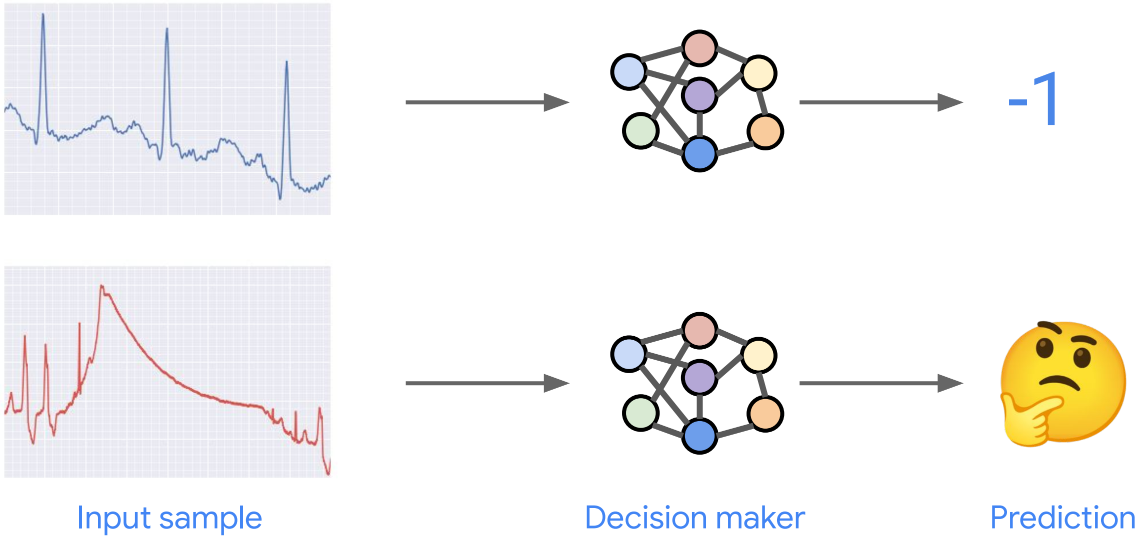
Ankit S.
Rawat

Aditya K.
Menon

Sanjiv
Kumar

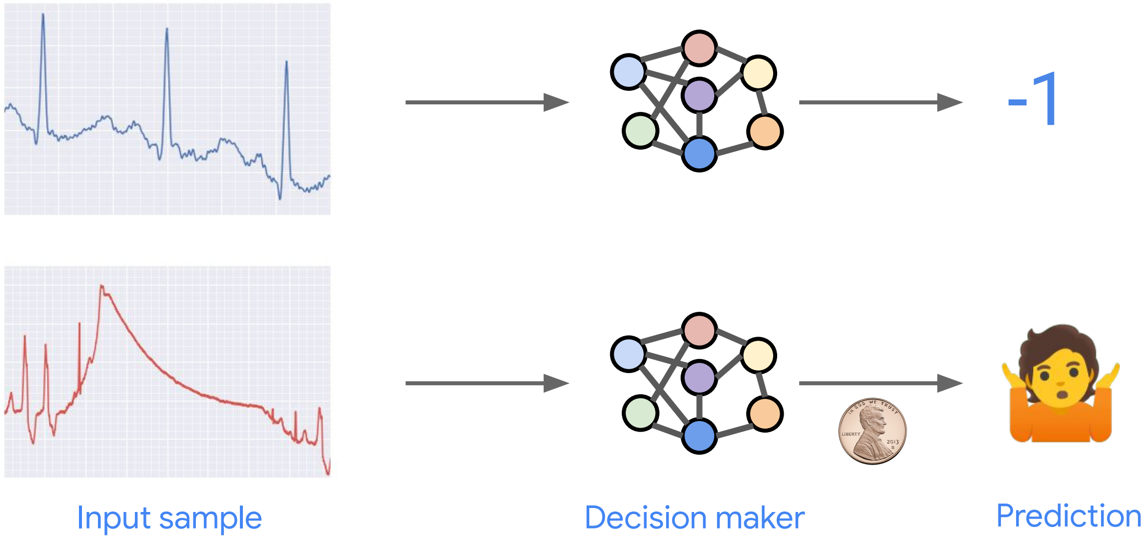
Standard classification paradigm

- Standard classification → **single model** for all samples
- However, it may be challenging to model the entire input space



Learning to reject

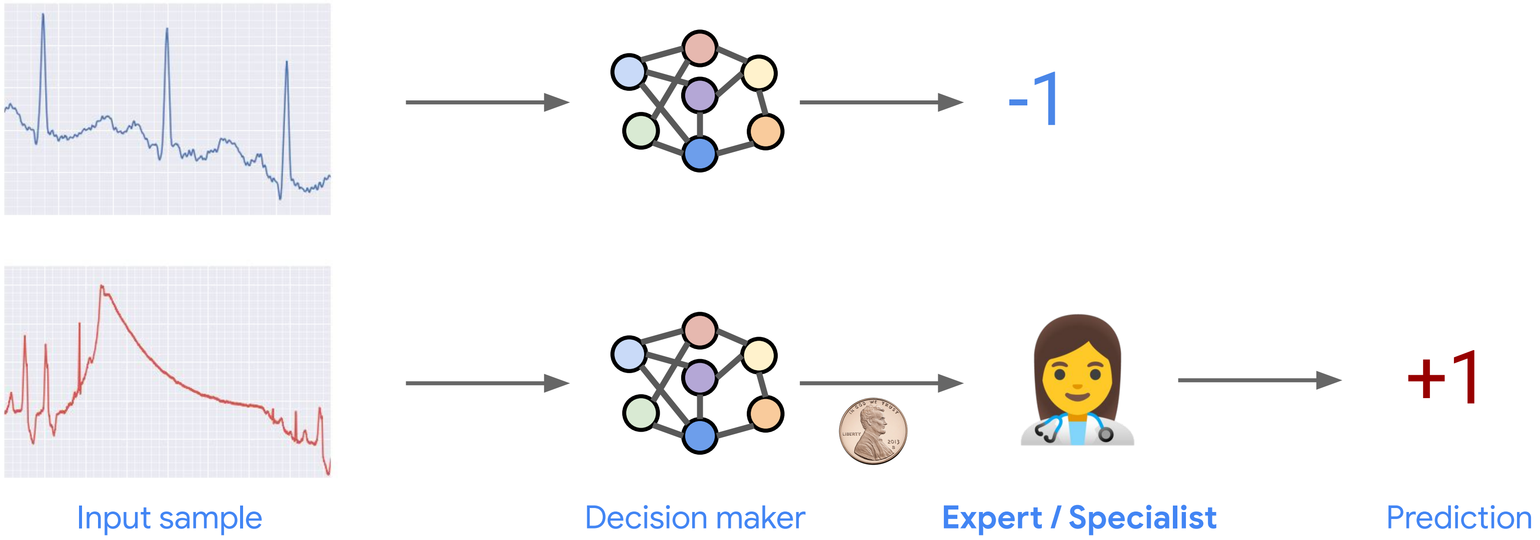
- Model can **give up** on a sample, incurring some **cost**



Chow (1970); Cortes et al. (2016); Ramaswamy et al. (2018)

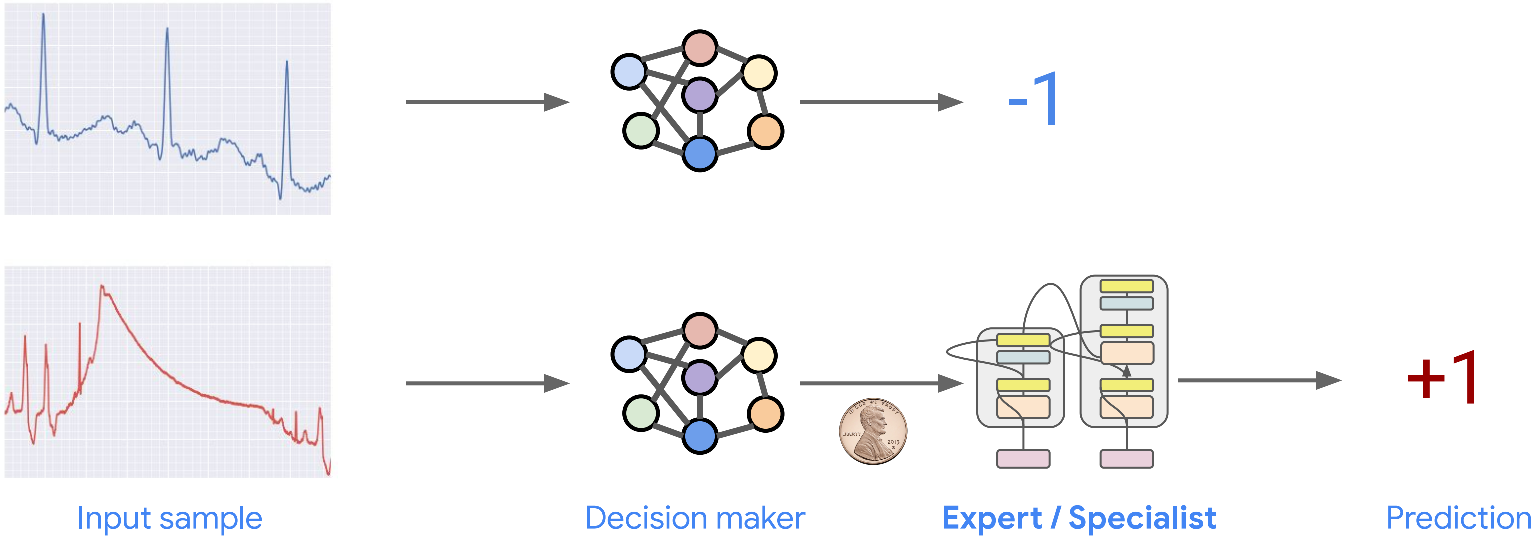
Learning to defer to an expert

- Model can **defer** to an **expert**, incurring some **cost**
  - e.g., human expert



Learning to defer to an expert

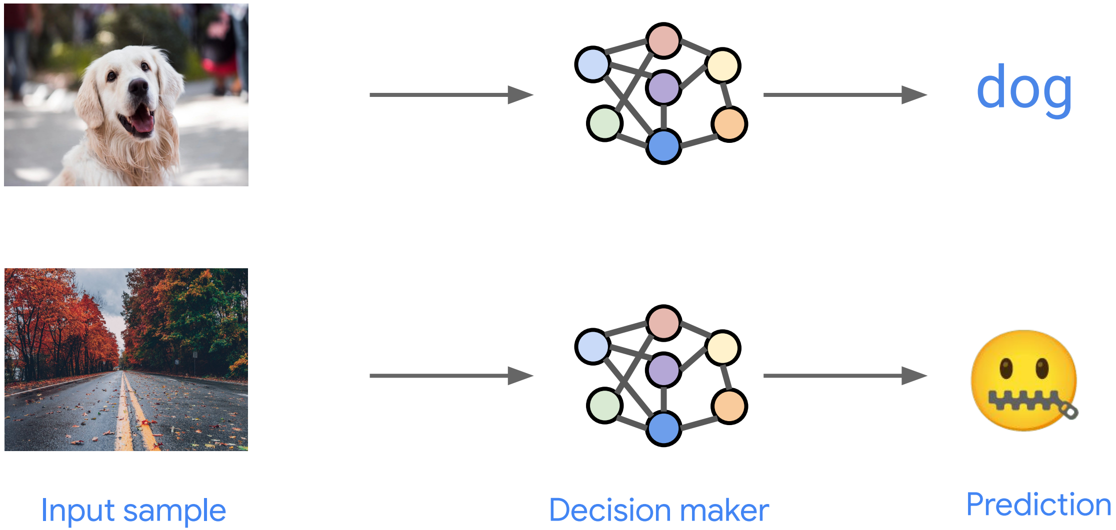
- Model can **defer** to an **expert**, incurring some **cost**
  - e.g., human expert, powerful learning model



Madras et al. (2018)

Learning to abstain on outliers

- Model can abstain on samples it deems to be **out-of-distribution (OOD)**



Goal: learn the base classifier, **and** the abstention rule

Cost of abstention: *classical version*

We will denote a joint classifier $h: X \to [n] \cup \{$🤷$\}$. In the simplest case, one may associate a constant cost c to abstaining on a sample

$$\mathbb{1}(\hat{y} \neq y,\ \hat{y} \neq \text{🤷}) + c \cdot \mathbb{1}(\hat{y} = \text{🤷})$$

Usual error when not abstaining

Constant cost when abstaining

Chow's rule: a surprisingly competitive baseline

C. Chow. On optimum recognition error and reject tradeoff.
IEEE Transactions on Information Theory, 16(1):41–46, 1970.

Bayes-optimal rejection rule: abstain on a sample when

$$\max_y \mathbb{P}(y \mid x) < 1 - c$$

Chow's rule: a surprisingly competitive baseline

C. Chow. On optimum recognition error and reject tradeoff.
IEEE Transactions on Information Theory, 16(1):41–46, 1970.

Bayes-optimal rejection rule: abstain on a sample when

$$\max_y \widehat{\mathbb{P}}(y \mid x) < 1 - c$$

In practice: max softmax probability
from a standard classifier

When Chow’s rule fails and ways to remedy it!

- Learning to reject
  - classical Chow’s rule is very competitive
- Learning to defer to an expert
  - remedy: expert-aware Chow’s rule
- Learning to abstain on outliers
  - remedy: outlier-aware Chow’s rule

Cost of abstention: *when deferring to an expert*

In the learning to defer paradigm, the cost of invoking the expert:

$$\mathbb{1}\left(\hat{y} \neq y,\, \hat{y} \neq \text{🤷}\right) \;+\; c_{\text{exp}}(x, y) \cdot \mathbb{1}\left(\hat{y} = \text{🤷}\right)$$

Usual error when not abstaining

Cost of invoking the expert

Expert cost: fixed cost + expert's error rate

A natural candidate for the expert cost would include both a fixed cost and the penalty when the expert makes a mistake

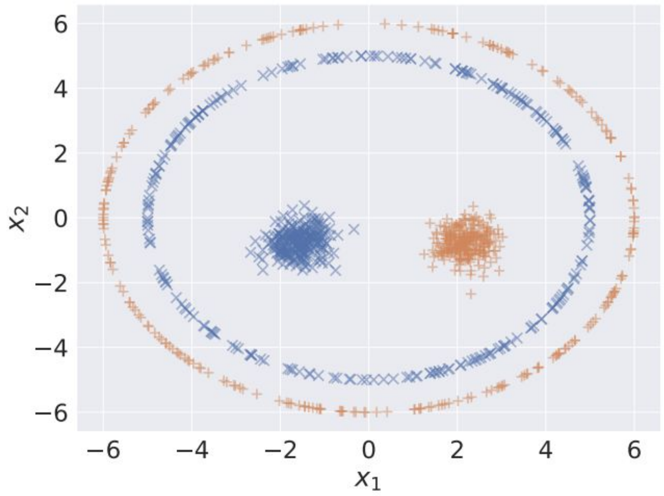
$$c_{\text{exp}}(x, y) = c_0 + \mathbf{1}(y \neq h_{\text{exp}}(x))$$

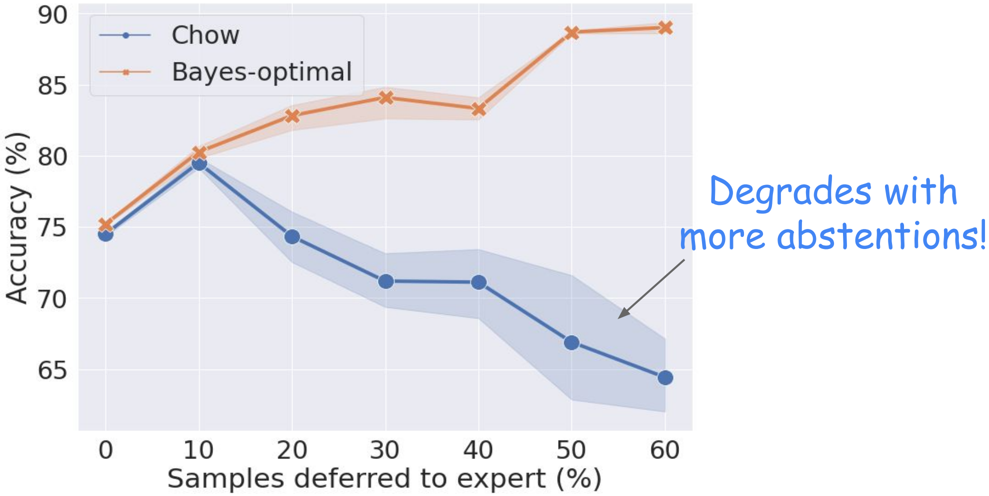
Fixed cost

Expert prediction


(e.g. monetary cost)

Chow’s rule can be sub-optimal for this setting





Synthetic dataset
Base model: linear features
Expert model: quadratic features

Expert-aware Chow's rule

Bayes-optimal rule: defer on a sample when

$$\max_y \mathbb{P}(y \mid x) < \mathbb{E}_{y|x}[\mathbf{1}(y = h_{\exp}(x))] - c_0$$

~ Base classifier's confidence

~ Expert's confidence

Expert-aware Chow's rule

When the expert's confidence is highly non-uniform, this is substantially different from Chow's rule

Bayes-optimal rule: defer on a sample when

$$\max_y \mathbb{P}(y \mid x) < \boxed{\mathbb{E}_{y|x}[\mathbf{1}(y = h_{\exp}(x))]} - c_0$$

~ Base classifier's confidence

~ Expert's confidence

Expert-aware Chow's rule

Bayes-optimal rule: defer on a sample when

$$\max_{y} \widehat{\mathbb{P}}(y \mid x) \; < \; \widehat{\mathbb{E}}_{y|x}[\mathbf{1}(y = h_{\mathrm{exp}}(x))] \; - \; c_0$$

~ Base classifier's confidence

~ Expert's confidence

Unlike classical Chow, we need to estimate the expert's confidence

Separate model for expert's confidence

Raghu et al. '19 suggest training separate model to estimate expert's confidence (using a sample annotated with the expert's predictions)

$$\max_{y} \widehat{\mathbb{P}}(y \mid x) < \widehat{\mathbb{E}}_{y|x}[\mathbf{1}(y = h_{\exp}(x))] - c_0$$

~Softmax probabilities from base classifier

~ Separate model to estimate expert's confidence

Separate model for expert's confidence

- This approach has appealing properties:
  - ✓ Simple to compute
  - ✓ Approximates the Bayes deferral rule
  - ! Separate models to estimate base and expert confidence

Cost-sensitive softmax cross-entropy (CSS)

- Mozannar & Sontag ‘20 suggest training a joint model with an additional label $\perp$

Google

Cost-sensitive softmax cross-entropy (CSS)

- Mozannar & Sontag '20 suggest training a joint model with an additional label $\perp$
- Minimize a **cost-sensitive** softmax cross-entropy (**CSS**) loss

$$\ell_{\text{css}}(x, y, \bar{f}(x)) = \underbrace{-\log\left(\bar{p}_y(x)\right)}_{\text{Classification loss to train base model}} - \underbrace{\mathbf{1}(y = h_{\text{exp}}(x)) \cdot \log\left(\bar{p}_\perp(x)\right)}_{\text{Loss to estimate expert's confidence}} - \underbrace{c_0 \cdot \sum_{y'} \log(\bar{p}_{y'}(x))}_{\text{Takes into account fixed cost } c_0}$$

*The original loss of Mozannar and Sontag uses a slightly tighter formulation; see our paper for details

The case for CSS

- The CSS loss has a number of appealing characteristics:

  ✓ **Joint model** for both base classifier and expert's confidence

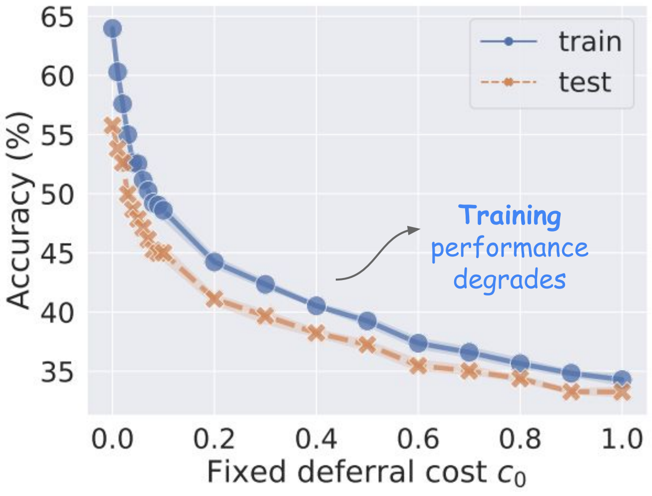
  ✓ Optimal solution matches the **Bayes-optimal classifier**

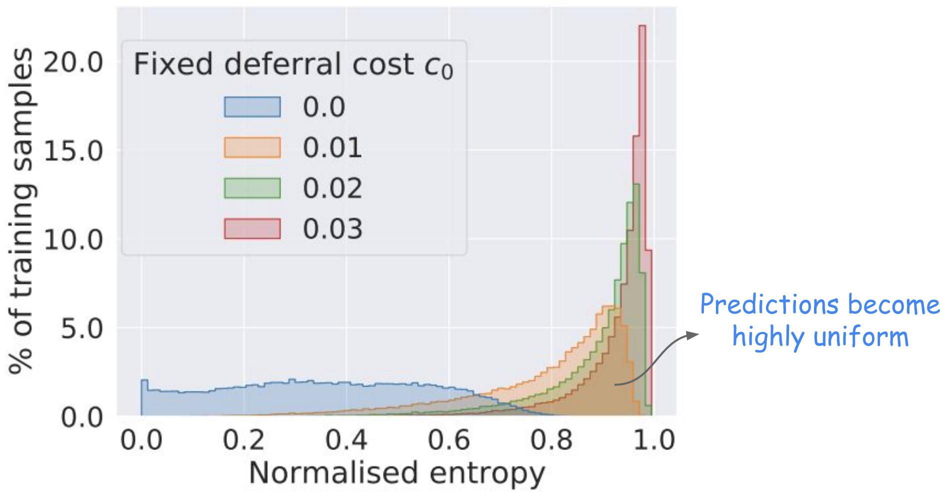
  ✓ **Empirically effective** on several benchmarks

  ! when fixed cost $\mathbf{c_0 = 0}$...

The case against CSS?

- CSS strongly **underfits** when there is non-zero fixed deferral cost c_0!





CIFAR 100
ResNet8 base
ResNet32 expert

A label smoothing perspective

- CSS equivalently applies high level of **label smoothing**:

$$\ell_{\text{css}}(x, y, \bar{f}(x)) = -\log\left(\bar{p}_y(x)\right) - \mathbf{1}(y = h_{\text{exp}}(x)) \cdot \log\left(\bar{p}_\perp(x)\right) - \boxed{c_0 \cdot \sum_{y'} \log(\bar{p}_{y'}(x))}$$

Treat **all** labels as candidate positive

  - Encourages predictions to become highly uniform
  - Low separation between true label and competing labels

A label smoothing perspective

- CSS equivalently applies high level of **label smoothing**:

$$\ell_{\text{css}}(x, y, \bar{f}(x)) = -\log\left(\bar{p}_y(x)\right) - \mathbf{1}(y = h_{\text{exp}}(x)) \cdot \log\left(\bar{p}_\perp(x)\right) - \boxed{c_0 \cdot \sum_{y'} \log(\bar{p}_{y'}(x))}$$

Treat **all** labels as candidate positive

  - Encourages predictions to become highly uniform
  - Low separation between true label and competing labels

- Not apparent when $c_0 = 0$ (as in prior work)!
  - $c_0 > 0$ is crucial in practical settings (e.g. when the expert is a larger model)

Solution: Set c_0 = 0 during training; include it in a post-hoc step

- Train base model with $c_0 = 0$, i.e., by minimizing:

$$\ell_{\text{css}}(x, y, \bar{f}(x)) = -\log\left(\bar{p}_y(x)\right) - \mathbf{1}(y = h_{\text{exp}}(x)) \cdot \log\left(\bar{p}_\perp(x)\right) - \cancel{c_0 \cdot \sum_{y'} \log(\bar{p}_{y'}(x))}$$

$\overline{p}_1$	$\overline{p}_2$	...	$\overline{p}_L$

Class probabilities

$\overline{p}_\perp$

Probability that the expert is correct

Solution: Set c_0 = 0 during training; include it in a post-hoc step

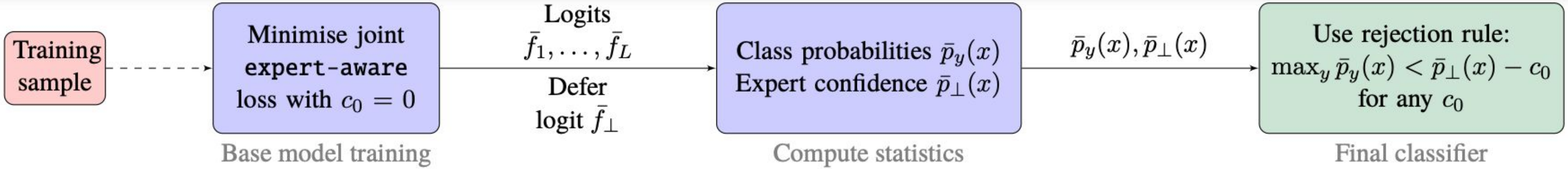
Construct a **post-hoc rejector** to include c_0 (that mimics the Bayes-optimal rule):

$$\max_y \bar{p}_y(x) < \bar{p}_\perp(x) - c_0$$

Probability that the expert is correct

Deferral cost

Proposal: two-step plug-in approach [Narasimhan et al ‘22]



Experimental setup

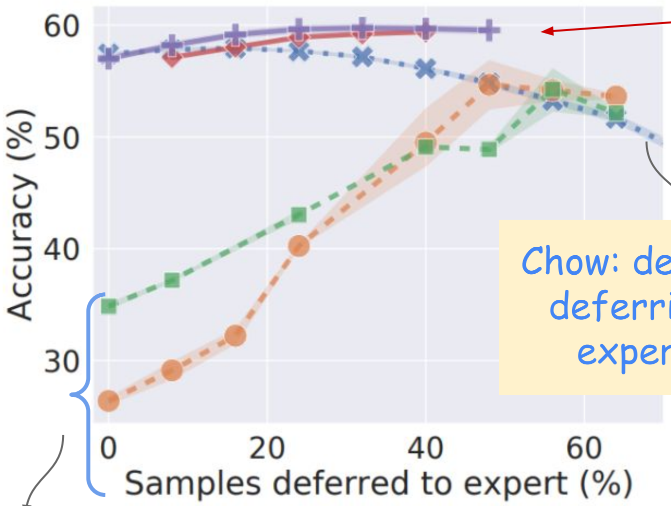
- **Specialist** expert
  - Model allowed to defer to a "specialist" expert trained on a subset of labels
- Baselines
  - **Chow**: confidence thresholding based only on the deferral cost c_0
  - **CSS**: in-training loss of Mozannar & Sontag (2020) with c_0 included
  - **OvA**: in-training loss of Verma and Nalisnick (2022) with c_0 included

Ignores expert error

Underfits when c_0 is large

Experimental results: expert-aware abstention

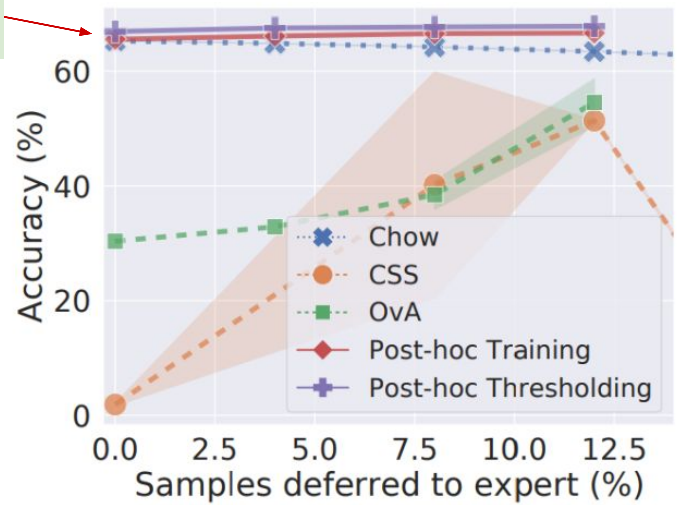
Ours



Chow: degrades when deferring more to expert (low c_0)

Joint losses: degrade when deferring less to expert (high c_0)

CIFAR 100
ResNet8 base
ResNet56 expert
(expert trained on first 10 coarse labels)



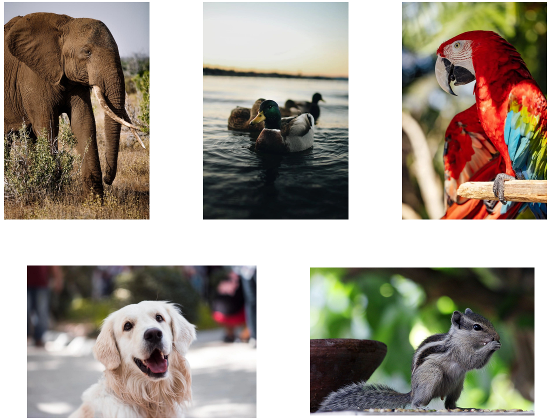
ImageNet
MobileNet-v2 base
EfficientNet-B0 expert
(expert trained on “dog” synset)

When Chow’s rule fails and ways to remedy it!

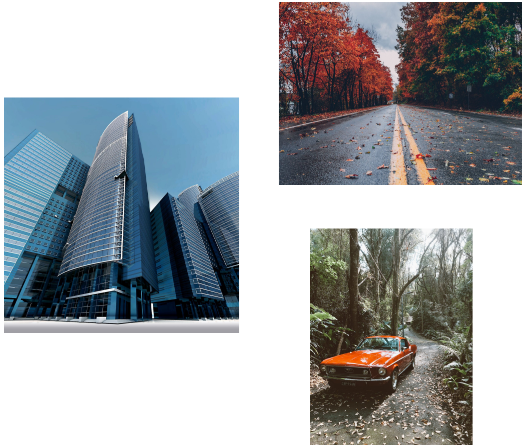
- Learning to reject
  - classical Chow’s rule is very competitive
- Learning to defer to an expert
  - remedy: expert-aware Chow’s rule
- Learning to abstain on outliers
  - remedy: outlier-aware Chow’s rule

Learning to abstain on outliers

Abstain on "out-of-distribution" samples that come from distribution different from the one used for training



Inlier samples



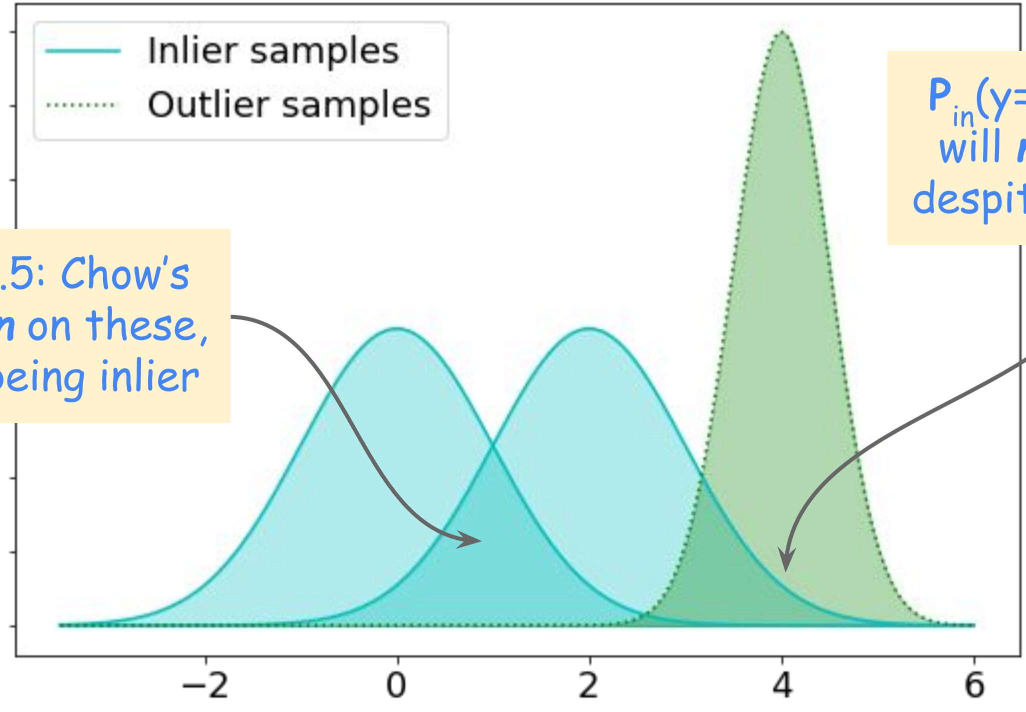
Outlier samples

Chow's rule (or the MSP scorer) is a popular baseline!

Thresholding the maximum softmax probability (MSP) from a standard classifier is a common baseline in this literature [Hendrycks et al. '17; Vaze et al. '22].

$$\max_{y \in [L]} \widehat{\mathbb{P}}_{\text{in}}(y \mid x) \; < \; t$$

Chow's rule can fail for outlier detection



$P_{in}(y=1|x)$ ~= 0.5: Chow's rule will ***abstain*** on these, despite them being inlier

$P_{in}(y=1|x)$ ~= 1: Chow's rule will ***not abstain*** on these, despite them being outliers.

Cost of abstention: *when abstaining on outliers*

We need to account for both inlier **and** outlier abstentions.

$$\underbrace{\mathbb{P}_{\text{in}}\left(\hat{y} \neq y,\, \hat{y} \neq 🤐\right)}_{\text{Error on inlier samples (when not abstaining)}} + \underbrace{\alpha}_{\text{Cost of abstaining on inlier samples}} \cdot \mathbb{P}_{\text{in}}\left(\hat{y} = 🤐\right) + \underbrace{\beta}_{\text{Cost of } \textit{not} \text{ abstaining on outlier samples}} \cdot \mathbb{P}_{\text{out}}\left(\hat{y} \neq 🤐\right)$$

Outlier-aware Chow's rule

Bayes-optimal rule: abstain on a sample when [Narasimhan et al. '23]

$$\max_y \mathbb{P}_{\text{in}}(y \mid x) \; < \; 1 - \alpha + \beta \cdot \frac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{in}}(x)}$$

Inlier class probabilities

Outlier-to-inlier density ratio

Outlier-aware Chow's rule

Bayes-optimal rule: abstain on a sample when [Narasimhan et al. '23]

$$\max_y \widehat{\mathbb{P}}_{\text{in}}(y \mid x) < 1 - \alpha + \beta \cdot \widehat{\frac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{in}}(x)}}$$

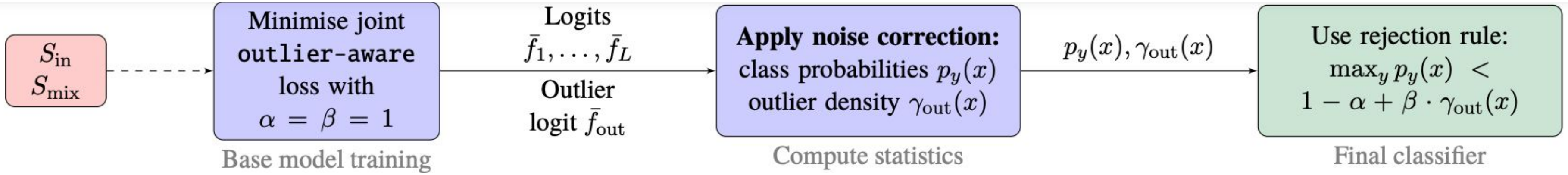
Inlier class probabilities

Outlier-to-inlier density ratio

We need to estimate both the inlier probabilities and the density ratio

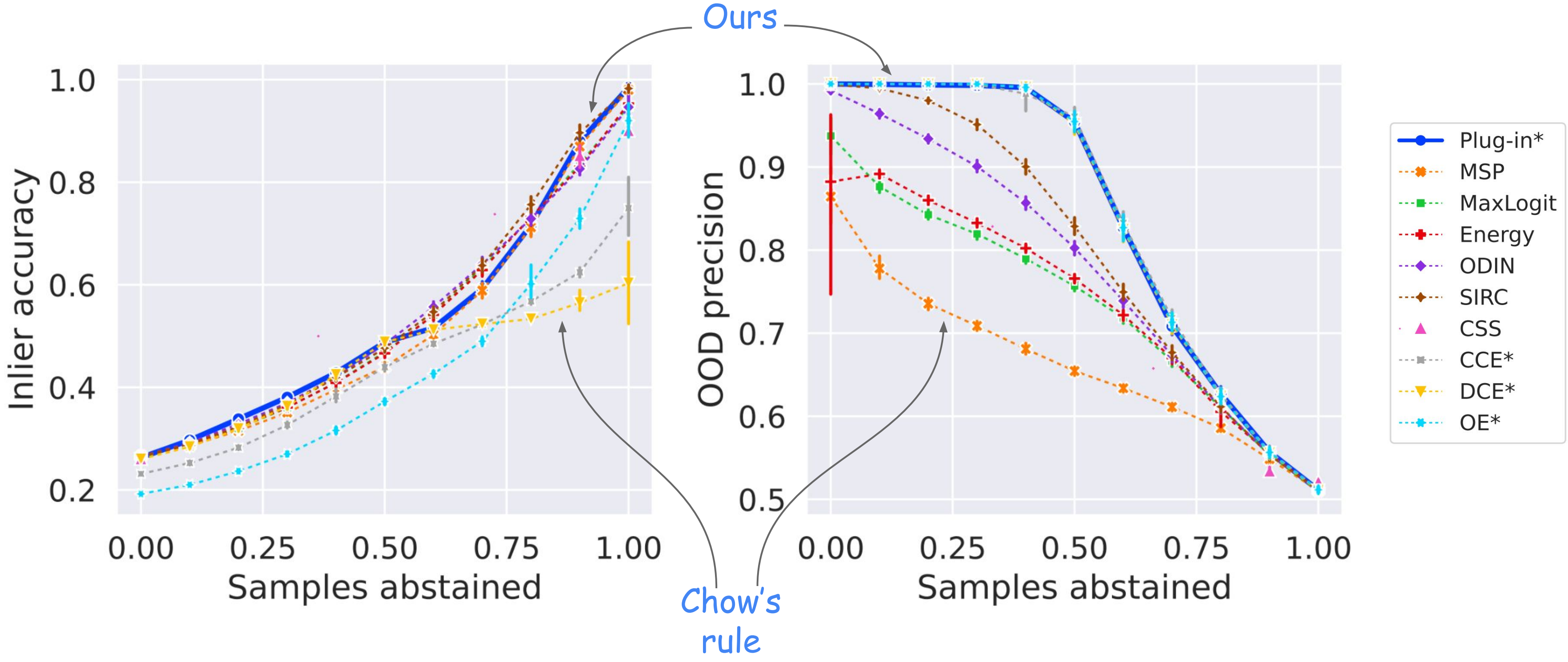
Our proposal: Two-step plug-in approach [Narasimhan et al '23]

Given: labeled inlier sample S_{in}, and an unlabeled mix of inlier and outlier samples S_{mix}



Experimental results: outlier-aware abstention

Inlier: CIFAR 100
Outlier: OpenImages
ResNet56 model



When Chow's rule fails and ways to remedy it!

- Learning to reject (L2R)
    - Classical Chow's rule is very competitive
- Learning to defer to an expert (L2D)
    - Chow may fail; use *expert-aware* Chow
- Learning to abstain on outliers (OOD)
    - Chow may fail; use *outlier-aware* Chow

L2R	$\max_y \mathbb{P}(y \mid x)$	$<$	$1 - c$
L2D	$\max_y \mathbb{P}(y \mid x)$	$<$	$\mathbb{E}_{y \mid x}[\mathbf{1}(y = h_{\text{exp}}(x))] - c_0$
OOD	$\max_y \mathbb{P}_{\text{in}}(y \mid x)$	$<$	$1 - \alpha + \beta \cdot \dfrac{\mathbb{P}_{\text{out}}(x)}{\mathbb{P}_{\text{in}}(x)}$

Narasimhan et al. "Post-hoc Estimators for Learning to Defer to an Expert". NeurIPS 2022.

Narasimhan et al. "Learning to Reject Meets OOD Detection: Are All Abstentions Created Equal?". Manuscript, 2023. [arXiv:2301.12386]

Thank you!