Consistent Estimators for Learning to Defer to an Expert

Hussein Mozannar (MIT)

David Sontag (MIT)

ICML 2020

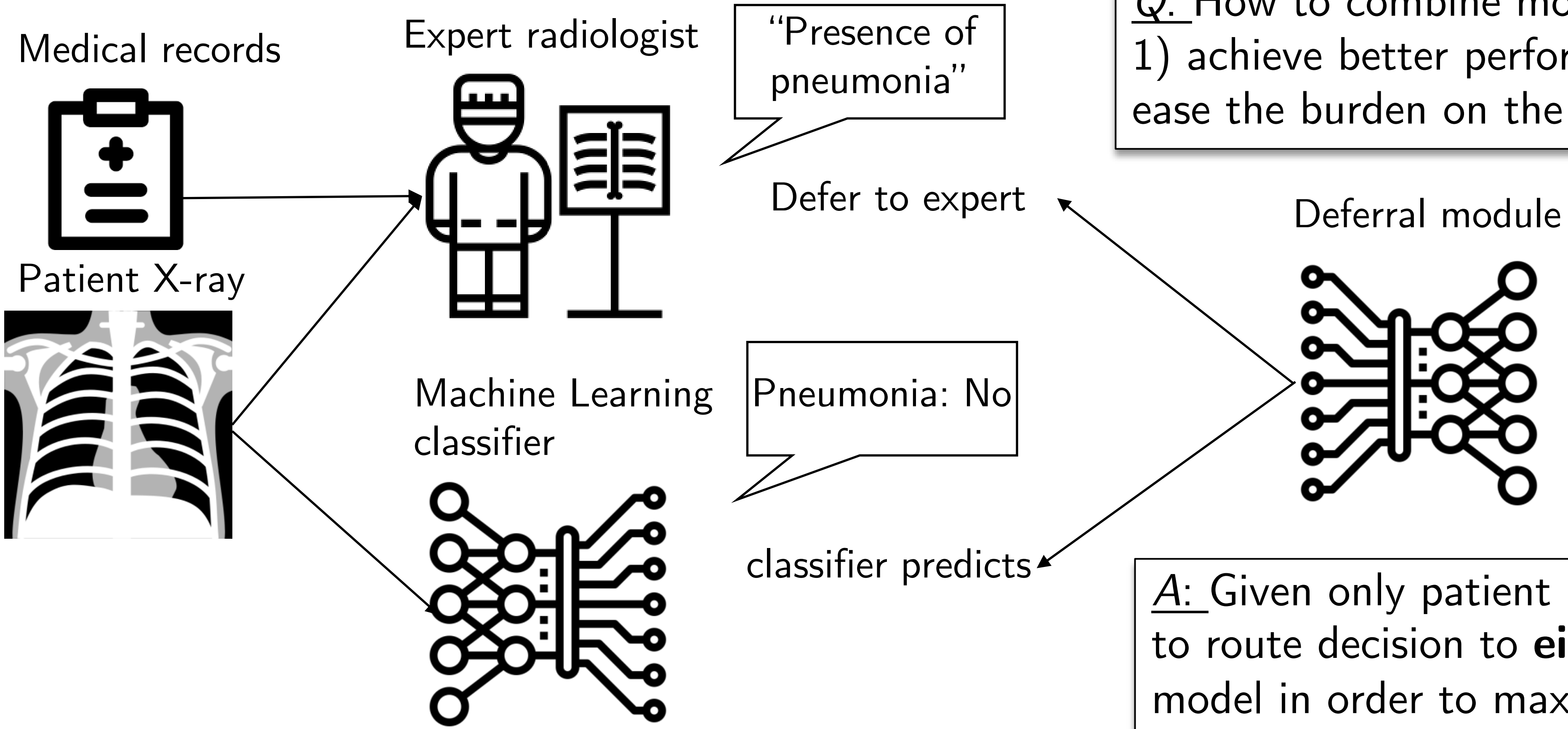
MIT INSTITUTE FOR DATA, SYSTEMS, AND SOCIETY

CSAIL

Learning to Defer

- Example task: Chest X-ray diagnosis of pneumonia

Q: How to combine model and expert to 1) achieve better performance and 2) ease the burden on the radiologist?



A: Given only patient input, learn model to route decision to **either** expert or model in order to maximize system performance

Multiple applications in healthcare and content moderation can (or already) utilize such modules.

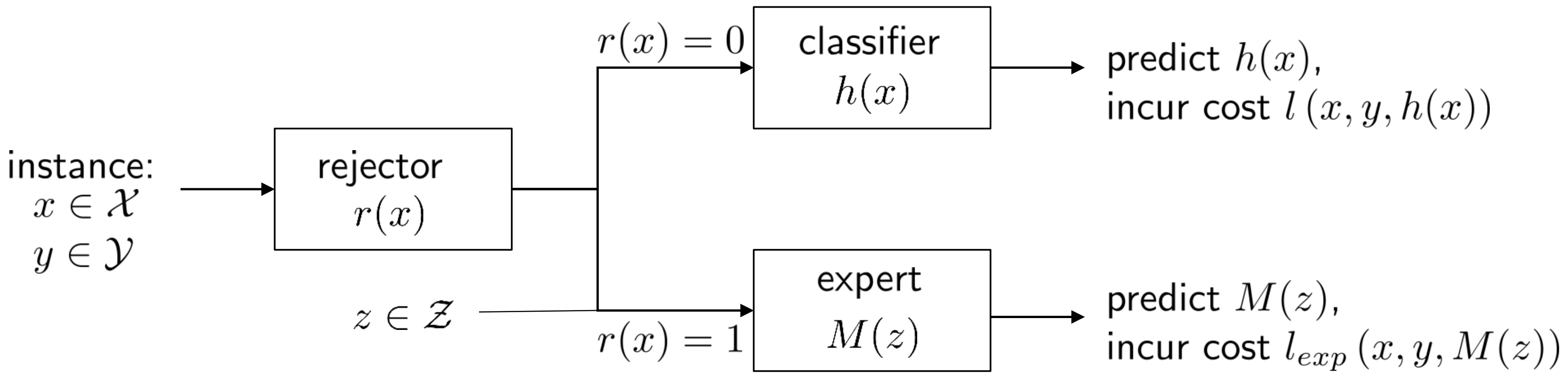
Our Contributions

- We formalize the learning to defer setting and propose a novel convex consistent surrogate loss, this loss is motivated by a reduction to cost sensitive learning. This settles an open problem by [Ni et al., NeurIPS 2019] for a consistent surrogate for rejection learning.
- We analyze previous approaches in the literature from a consistency point of view and give a generalization bound for minimizing the empirical objective.
- We provide a detailed experimental evaluation of our method on various tasks.

Related Work

- Madras et al. (NeurIPS 2018) proposes a mixture of experts loss, resulting loss is not consistent and fails empirically.
- Raghu et al. (2019) propose a confidence score method that compares expert and algorithm confidence. However, classifier cannot adapt to expert.
- Det al. (AAAI 2020) gives an approximate algorithm for ridge regression, Wilder et al. (IJCAI 2020) combines mixtures of experts loss and confidence score comparison.
- Related problems: selective classification (Geifman & El-Yaniv, NeurIPS 2017), learning with a reject option (Ni et al., NeurIPS 2019)

Learning to Defer: Problem Formulation

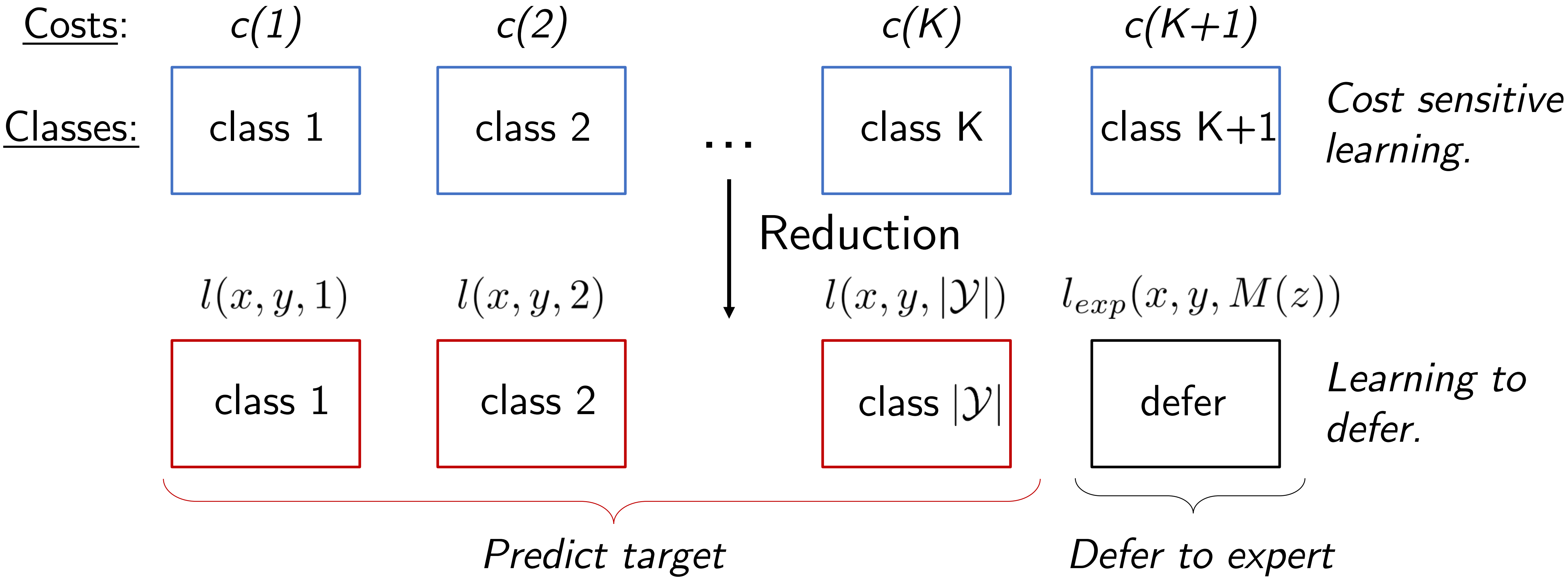


- **Jointly** learn a classifier *h(x)* and rejector *r(x)* to minimize system loss:

$$L(h, r) = \mathbb{E}_{(x,y)\sim\mathbf{P}, m\sim M|(x,y)} \left[\underbrace{l(x, y, \hat{Y}(x))}_{\text{classifier cost}} \overbrace{\mathbf{I}_{r(x)=0}}^{\text{predict}} + \underbrace{l_{\exp}(x, y, m)}_{\text{expert cost}} \overbrace{\mathbf{I}_{r(x)=1}}^{\text{defer}} \right]$$

Reduction to cost sensitive learning

- Cost sensitive learning: given covariate *x pick class in [K+1] that has minimal cost:*



Surrogate loss for cost sensitive learning

- We propose a natural extension of the cross-entropy loss,

  let $g_i : \mathcal{X} \to \mathbb{R}$ for $i \in [K+1]$ and $h(x) = \arg\max_i g_i$, define

$$\tilde{L}_{CE}(g_1, \cdots, g_{K+1}, x, c(1), \cdots, c(K+1))$$
$$= -\sum_{i=1}^{K+1} (\max_{j \in [K+1]} c(j) - c(i)) \log \left(\frac{\exp(g_i(x))}{\sum_k \exp(g_k(x))} \right)$$

Proposition. $\tilde{L}_{CE}$ is a consistent loss function:

let $\tilde{\boldsymbol{g}} = \arg\inf_{\mathbf{g}} \mathbb{E}\left[\tilde{L}_{CE}(\mathbf{g}, \mathbf{c}) | X = x\right]$, then:
$$\arg\max_{i \in [K+1]} \tilde{\boldsymbol{g}}_i = \arg\min_{i \in [K+1]} \mathbb{E}[c(i) | X = x]$$

Minimizing 0-1 error of deferral system

- **Data**: $S = \{(x_i, y_i, m_i)\}_{i=1}^n$ where $\{(x_i, y_i)\}_{i=1}^n$ are the targets and covariates and m_i is the expert prediction
- System loss for misclassification errors:

$$L_{0-1}(h, r) = \mathbb{E}\left[\, \mathbf{I}_{h(x)\neq y}\mathbf{I}_{r(x)=0} + \mathbf{I}_{m\neq y}\mathbf{I}_{r(x)=1} \,\right]$$

- Let $g_y : \mathcal{X} \to \mathbb{R}$ for $y \in \mathcal{Y}$, $h(x) = \arg\max_{y\in\mathcal{Y}} g_y(x)$, similarly let $g_d : \mathcal{X} \to \mathbb{R}$ and define $r(x) = \mathbf{I}_{g_d(x)\geq \max_{y\in\mathcal{Y}} g_y(x)}$, our surrogate becomes:

$$L_{CE} = -\log\left(\frac{\exp(g_y(x))}{\sum_{y'\in\mathcal{Y}\cup d}\exp(g_{y'}(x))}\right) - \mathbf{I}_{m=y}\log\left(\frac{\exp(g_d(x))}{\sum_{y'\in\mathcal{Y}\cup d}\exp(g_{y'}(x))}\right)$$

Consistent surrogate loss and heuristic for adapting to expert

Theorem. The loss L_{CE} is convex in $\mathbf{g}$, upper bounds L_{0-1} and produces consistent estimator for L_{0-1}.

- Heuristic with $\alpha \in \mathbb{R}^+$:

$$L_{CE}(g,x,y,m) = -\log\left(\frac{\exp(g_y(x))}{\sum_{y' \in \mathcal{Y} \cup d} \exp(g_{y'}(x))}\right)$$

$$L^{\alpha}_{CE}(g,x,y,m) = -\overbrace{(\alpha \cdot \mathbf{I}_{m=y} + \mathbf{I}_{m \neq y})}^{\text{re-weight if expert is correct}} \log\left(\frac{\exp(g_y(x))}{\sum_{y' \in \mathcal{Y} \cup d} \exp(g_{y'}(x))}\right)$$

$$- \mathbf{I}_{m=y} \log\left(\frac{\exp(g_d(x))}{\sum_{y' \in \mathcal{Y} \cup d} \exp(g_{y'}(x))}\right)$$

Generalization Bound for Learning

Theorem. For any expert M and $\mathbf{P}$ over $\mathcal{X} \times \mathcal{Y}$, let $0 < \delta < \frac{1}{2}$, then w.p. at least $1-\delta$, the the empirical minimizers $(\hat{h}^*, \hat{r}^*)$ satisfy:

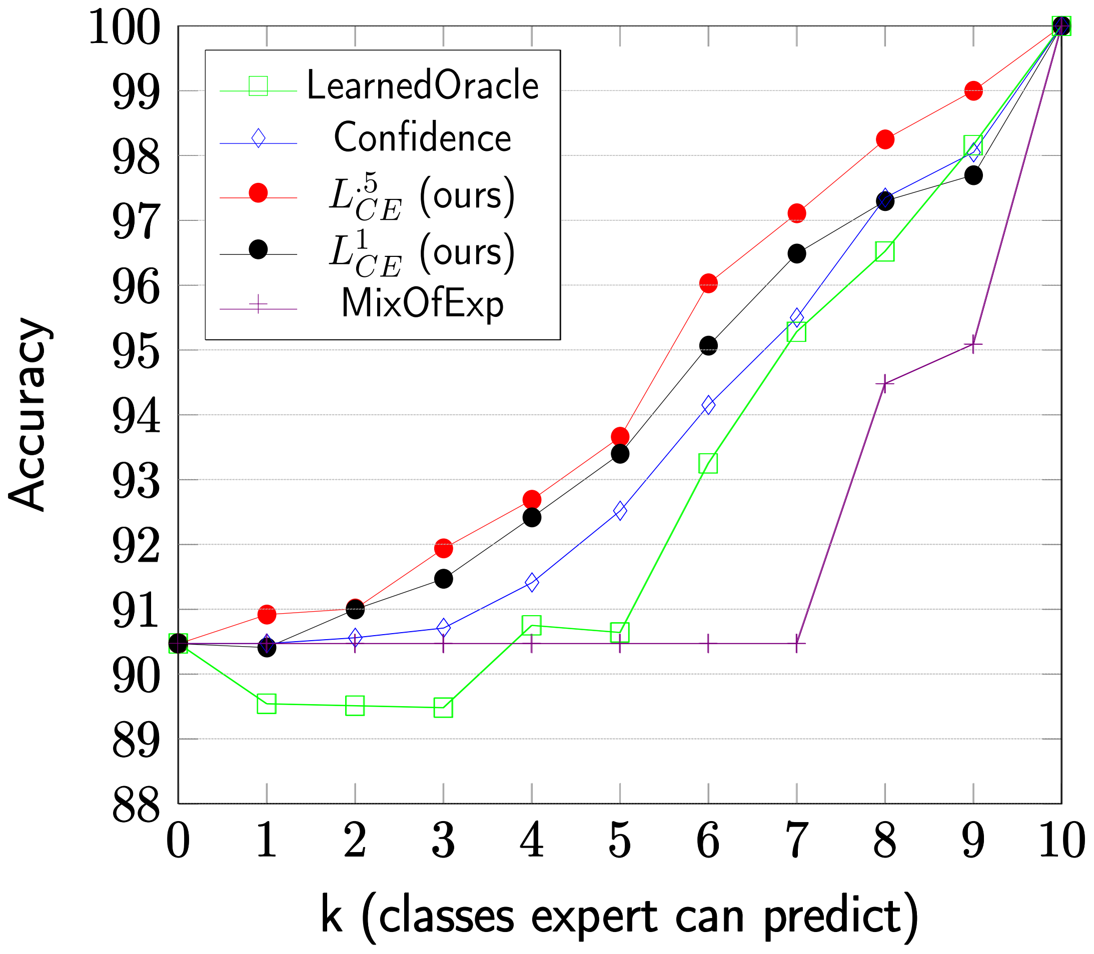
$$L_{0-1}(\hat{h}^*, \hat{r}^*) \leq L_{0-1}(h^*, r^*) + \mathfrak{R}_n(\mathcal{H}) + \mathfrak{R}_n(\mathcal{R}) + \mathfrak{R}_{n\mathbb{P}(M \neq Y)/2}(\mathcal{R}) + 2\sqrt{\frac{\log \frac{2}{\delta}}{2n}} + \frac{\mathbb{P}(M \neq Y)}{2} \exp\left(-\frac{n\mathbb{P}(M \neq Y)}{8}\right)$$

Takeaway: Sample complexity depends on the expert error, complexity of model class of classifier and rejector

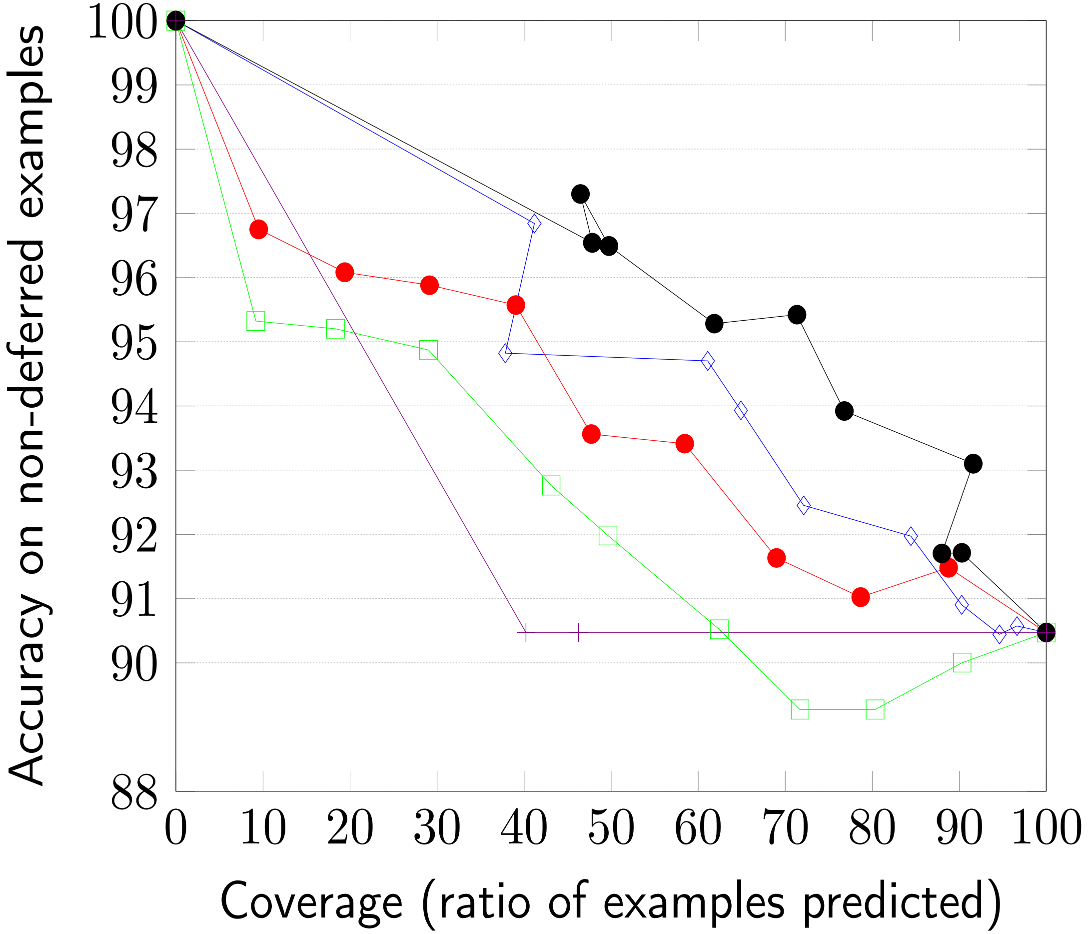
Experiments: CIFAR-10 setup

- **CIFAR 10:** image classification over 10 classes, parameterize model $\boldsymbol{g}$ as a WideResNet with 11 output layers, no data augmentations were used.
- **Synthetic expert:** let $1 \leq\ k \leq 10$, then if the image belongs to the first k classes the expert predicts perfectly, otherwise the expert predicts uniformly at random.
- **Baselines:** 1) MixOfExp (Madras et al. 2018), 2) Confidence (Raghu et al. 2019), 3) LearnedOracle: build model to predict if image is in first k classes and defer accordingly.

Experiments: CIFAR-10 results



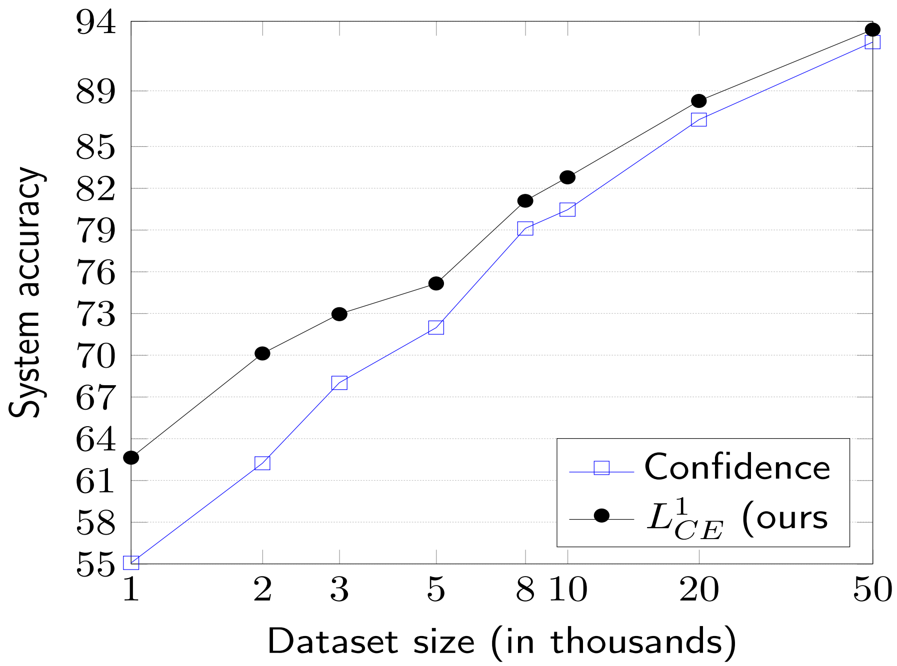
Accuracy of combined system for each expert k



Accuracy of classifier and it's coverage for each expert k

Why do we outperform the baselines?

1. **Sample Complexity:** as we restrict training data, gains over Confidence increase
2. **Considering classifier's confidence:** LearnedOracle baseline does not look at confidence of classifier and hence suffers.
3. **Consistency:** MixOfExp baseline is not consistent, there is a mismatch between the loss and actual misclassification error



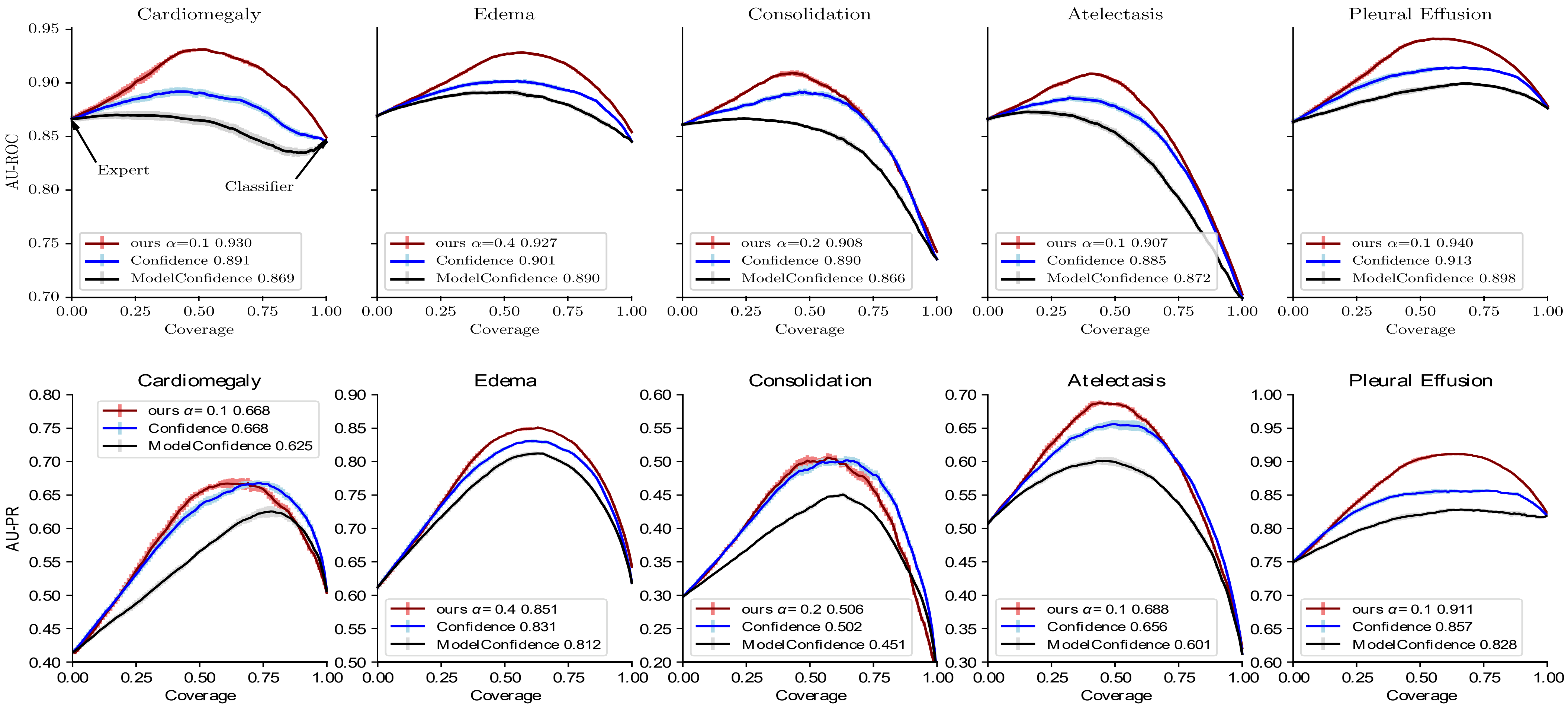
Restricting training data size and showing system accuracy

CheXpert Experimental Setup

- **CheXpert:** large chest X-ray dataset with over 224k automatically labeled images for the presence of 14 observations (Irvin et al., 2019)
- **Synthetic expert:** if patient has supporting device, expert is correct with probability $\boldsymbol{p}$, otherwise expert is correct with probability $\boldsymbol{q}$
- **Baselines:** 1) Confidence (Raghu et al., 2019), 2) ModelConfidence: defer based on confidence of model
- **Task:** We constrain our method and the baselines to achieve c% coverage and measure AU-ROC & AU-PR of the system.

Chest X-ray of patient with Cardiomegaly

CheXpert Results



Plot of AU-ROC of the ROC curve (a) for each level of coverage and of the AU-PR (AP) (b) for each of the 5 tasks comparing our method with the baselines on the training derived test set for the toy expert with q=0.7, p=1.

Future Work

- Ongoing work evaluating with real radiologist data
- Integrating (fairness) constraints for deferral with a theoretical basis
- Deferring to multiple experts.