Google DeepMind

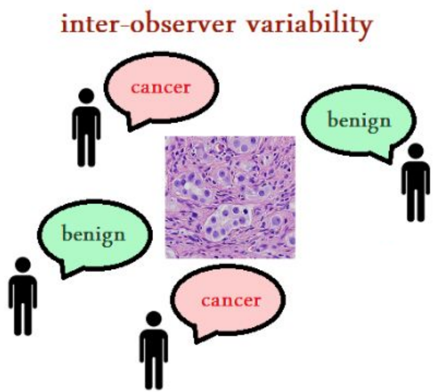
Conformal training and conformal prediction under ambiguous ground truth

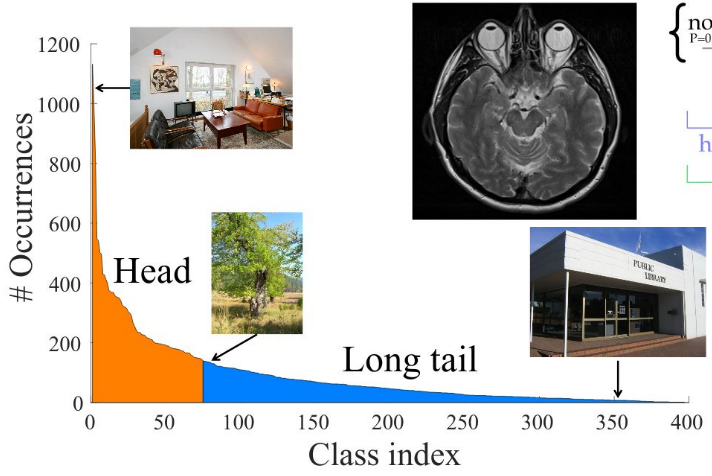
David Stutz

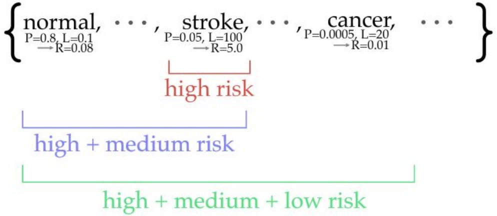
September 8th 2023

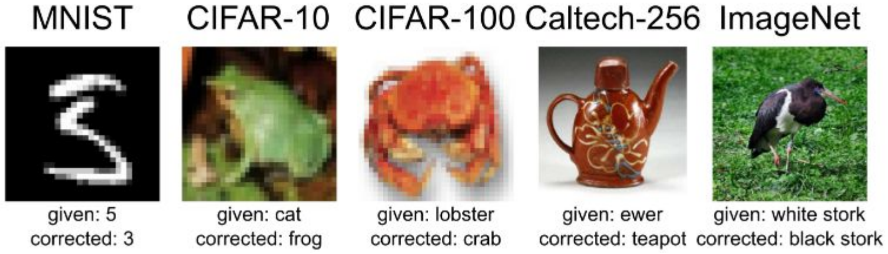
Motivation: Ambiguity in Classification

- High-stakes and security-critical applications
- Rich structure of (hierarchical) classes
- Rare classes or long-tailed class distribution
- True ground truth unknown or uncertain









Wang et al. Learning to Model the Tail, 2017; Karimi et al., Deep learning with noisy labels: exploring techniques and remedies in medical image analysis, 2020; Bates et al., Distribution-Free, Risk-Controlling Prediction Sets, 2021; Northcutt et al., Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks, 2021.

Talk Outline

Conformal prediction:

- Notation and background

Conformal training:

- How to better integrate conformal prediction with deep learning?
- Improve “efficiency” or application-specific losses

Paper: arxiv.org/abs/2110.09192

Ambiguous ground truth:

- How to deal with ambiguous/uncertain ground truth?
- For example, when annotators disagree

Paper: arxiv.org/abs/2307.09302

Conformal Prediction

For model $\pi_{\theta,y} \approx p(y|x)$ construct confidence sets $C_\theta(x) \subseteq [K] = \{1, \ldots, K\}$ such that

$$P(y \in C_\theta(x)) \geq 1 - \alpha$$

- confidence level α user-specified

Conformal Prediction

For model $\pi_{\theta,y} \approx p(y|x)$ construct confidence sets $C_\theta(x) \subseteq [K] = \{1, \ldots, K\}$ such that

$$P(y \in C_\theta(x)) \geq 1 - \alpha$$

- confidence level α user-specified
- *inefficiency* = average confidence set size $|C_\theta(x)|$
- requires only exchangeability (weaker than i.i.d.)




	{airplane}	{cat}	{deer,horse,dog}	{cat,frog}	true class
coverage/inefficiency	yes/1	yes/1	no/3	yes/2	

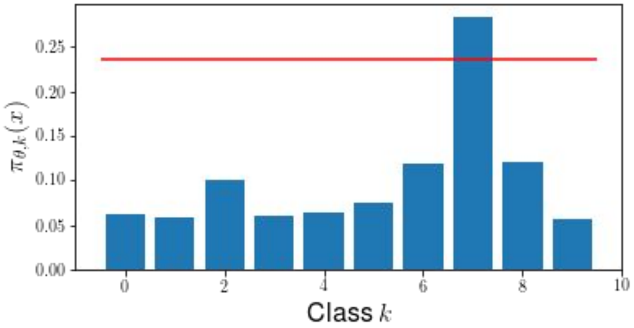
Split Conformal Prediction

Split conformal prediction with two steps: prediction and calibration:

1. Prediction (test time): define how confidence sets are constructed

$$C_\theta(x) := \{k \in [K] : E(x,k) := \pi_{\theta,k}(x) \geq \tau\}$$

with $E(x,k) := \pi_{\theta,k}(x)$ called conformity scores.



Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. Journal of the American Statistical Association (JASA), 114(525):223–234, 2019.

Split Conformal Prediction

Split conformal prediction with two steps: prediction and calibration:

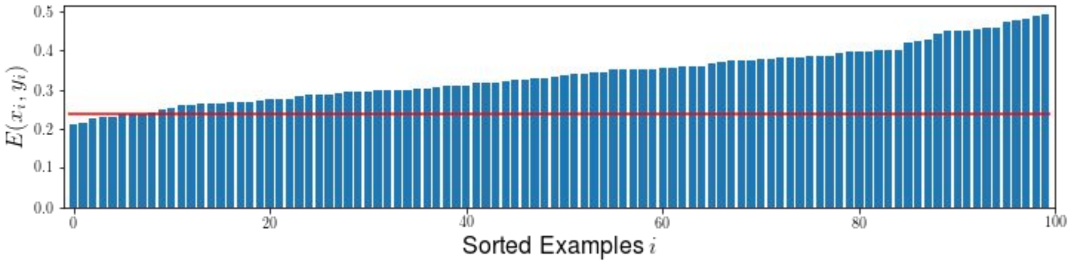
1. Prediction (test time): define how confidence sets are constructed

$$C_\theta(x) := \{k \in [K] : E(x,k) := \pi_{\theta,k}(x) \geq \tau\}$$

   with $E(x,k) := \pi_{\theta,k}(x)$ called conformity scores.

2. Calibration: define threshold τ on N held-out calibration examples as

$$\frac{\lfloor \alpha(N+1) \rfloor}{N}\text{-quantile of } \{E(x_i, y_i)\}_{i \in [N]}$$



Example Results

Inefficiency ↓ for different methods (82% base accuracy):

Dataset, α	Thr	APS	RAPS
CIFAR10, 0.05	**1.64**	2.06	1.74
CIFAR10, 0.01	**2.93**	3.30	3.06

Different conformity scores

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candes. Classification with valid and adaptive coverage. In Advances in Neural Information Processing Systems (NIPS), 2020.
Anastasios Nikolas Angelopoulos, Stephen Bates, Michael I. Jordan, Jitendra Malik: Uncertainty Sets for Image Classifiers using Conformal Prediction. ICLR 2021

Conformal Training

Conformal training:

- Notation and background

Conformal training:

- How to better integrate conformal prediction with deep learning?
- Improve “efficiency” or application-specific losses

Paper: arxiv.org/abs/2110.09192

Ambiguous ground truth:

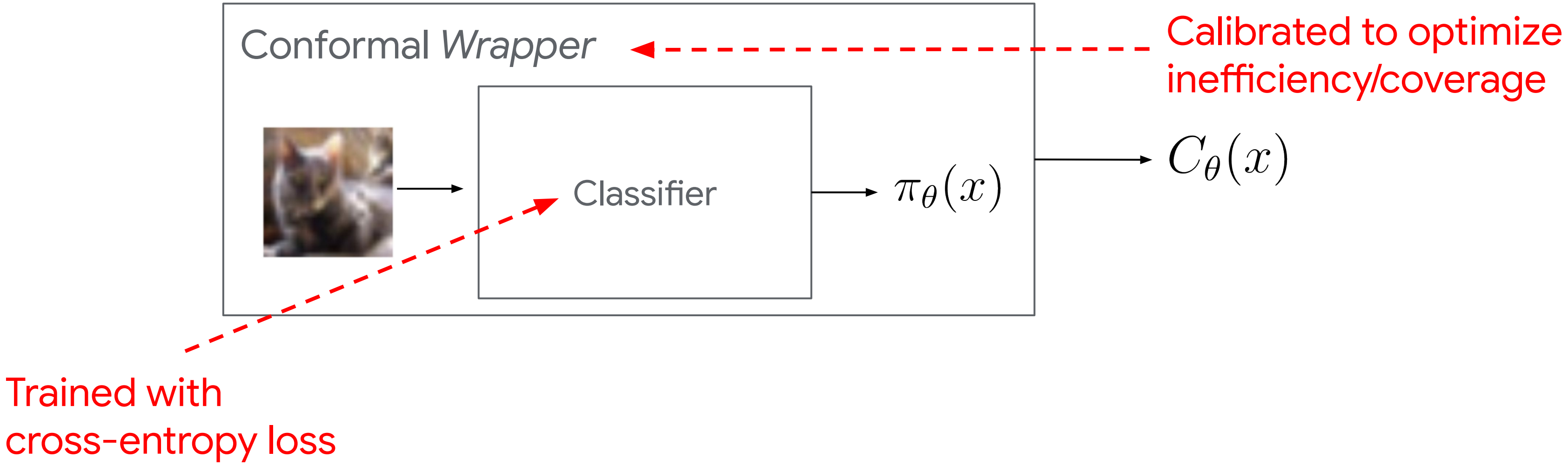
- How to deal with ambiguous/uncertain ground truth?
- For example, when annotators disagree

Paper: arxiv.org/abs/2307.09302

Conformal Training

Conformal prediction is typically applied *after* training:

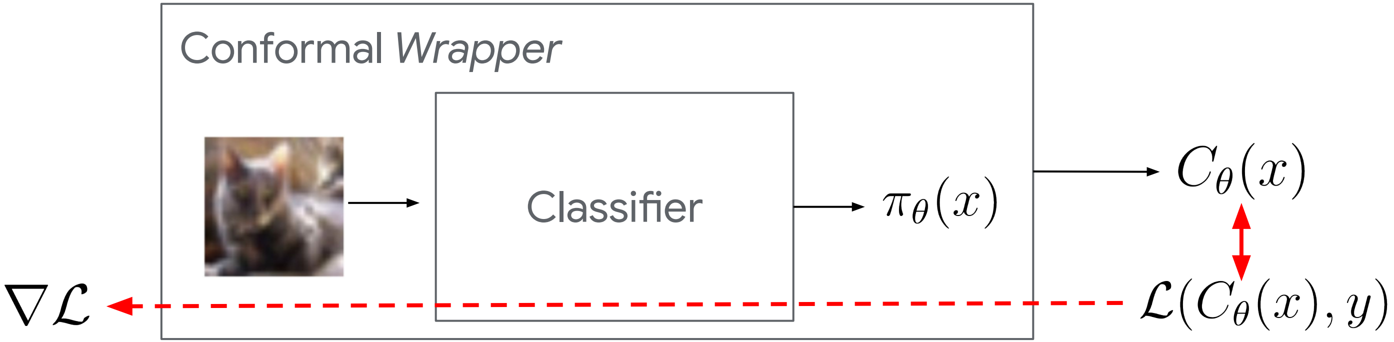
- Training loss and calibration objectives are not aligned!



Conformal Training

Conformal prediction is typically applied *after* training:

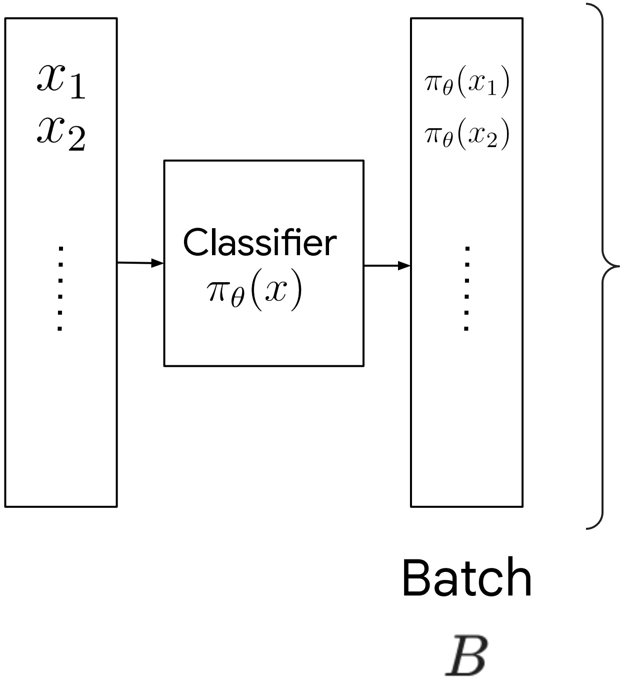
- Training loss and calibration objectives are not aligned!



➔ Preserve coverage guarantee

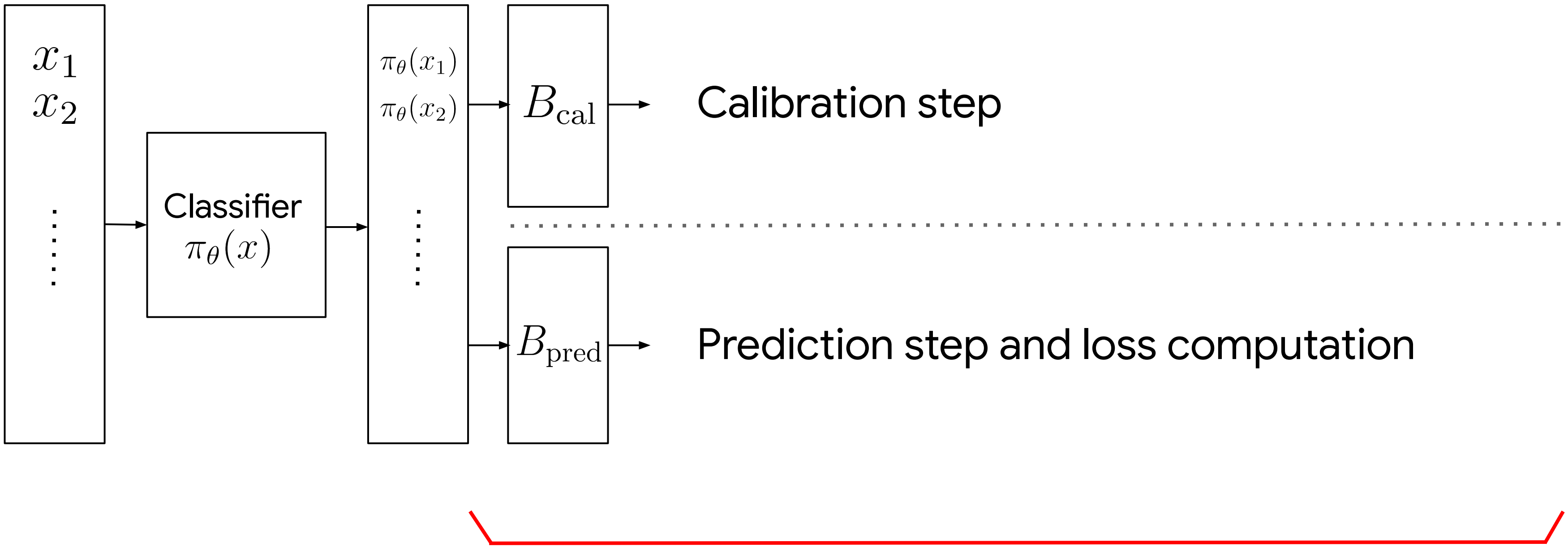
➔ Independent of conformal predictor used at test time

Conformal Training in Detail



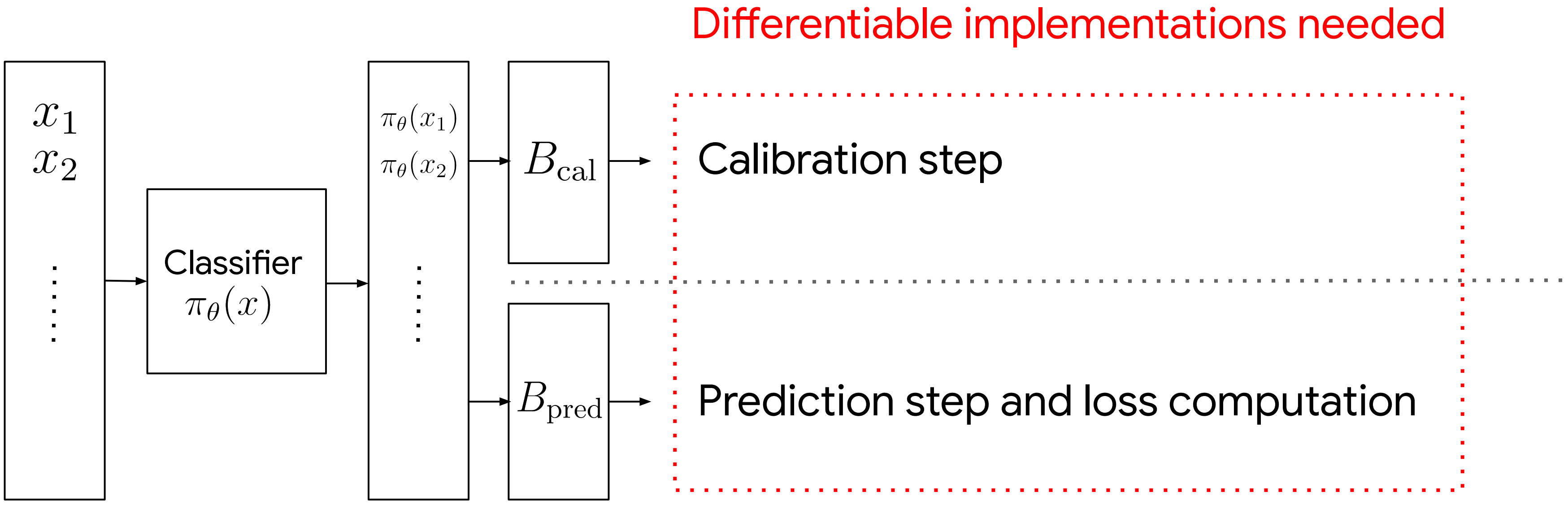
Standard mini-batch SGD-based training loop

Conformal Training in Detail

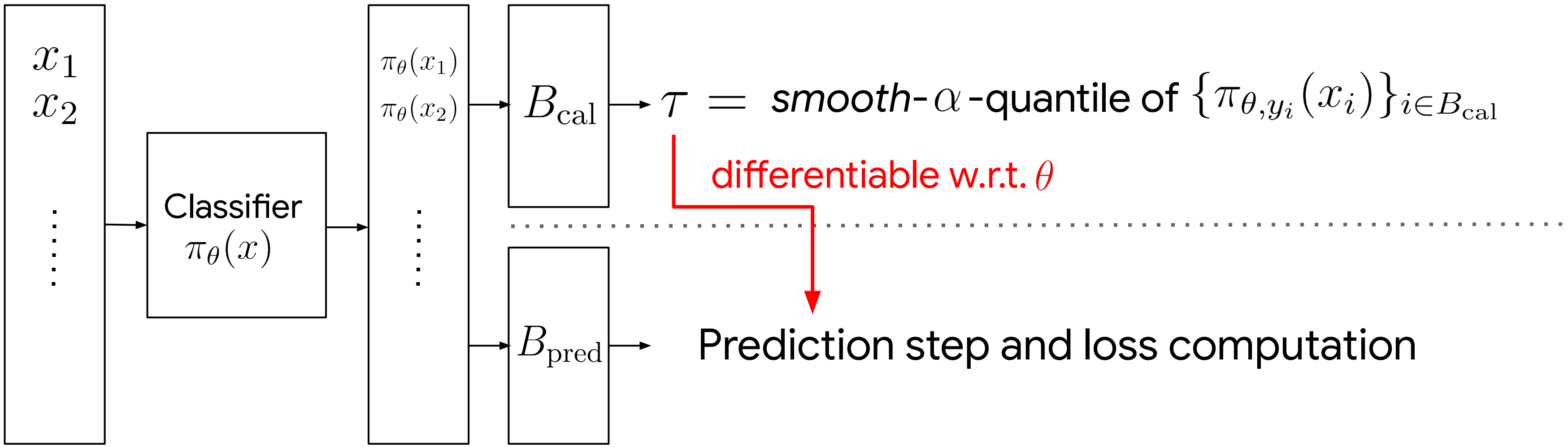


“Simulate” conformal prediction on each mini-batch

Conformal Training in Detail

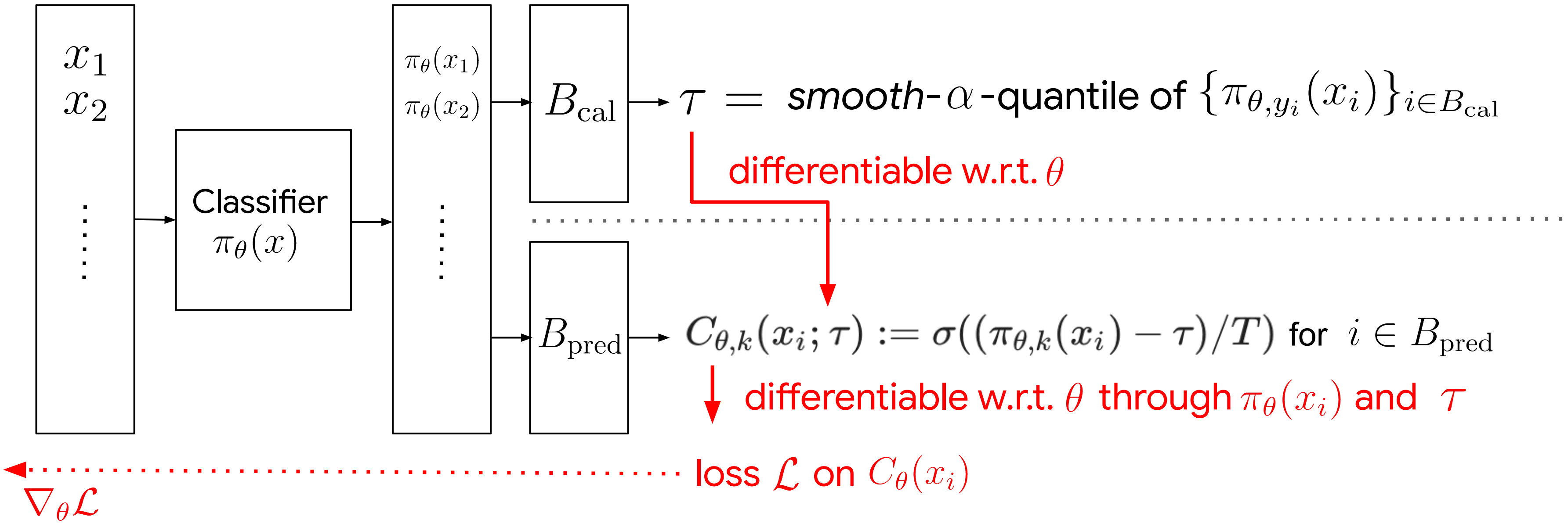


Conformal Training in Detail

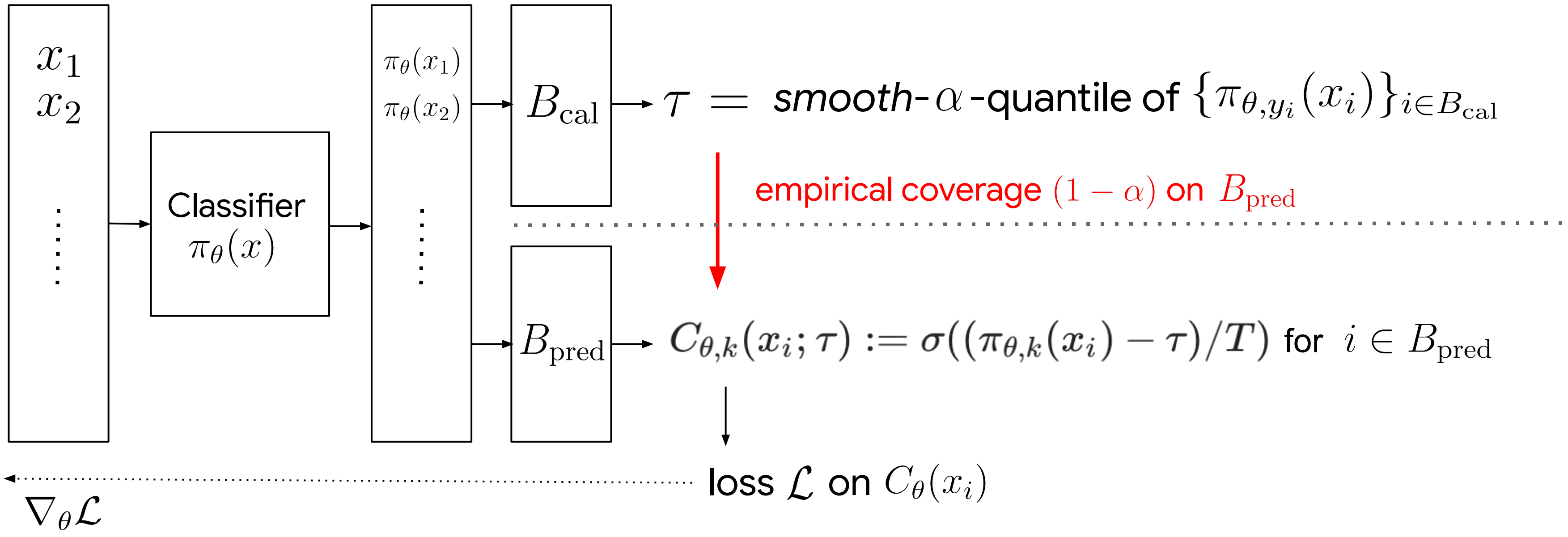


Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking. In Proc. of the International Conference on Machine Learning (ICML), 2020.
Marco Cuturi, Olivier Teboul, and Jean-Philippe Vert. Differentiable ranking and sorting using optimal transport. In Advances in Neural Information Processing Systems (NeurIPS), 2019.
John H Williamson. Differentiable parallel approximate sorting networks, 2020

Conformal Training in Detail



Conformal Training in Detail



➔ Re-calibrate at test time to obtain coverage guarantee!

Conformal Training in *More* Detail

- Differentiable sorters usually come with a "smoothness" parameter ϵ :
  $\epsilon, T \to \infty$ recovers "hard" split conformal prediction
- Batch size needs to fit confidence level
- Can use different conformity scores during training and test time (we use the model's softmax as conformity score during training)
- Training from scratch vs. fine-tuning:
  - Training deeper networks from scratch difficult
  - Fine-tuning often limits the benefits we get from conformal training
- Conformal training independent of architecture, optimization algorithm, regularizers, etc.

Training Objectives

A Reducing inefficiency:

- Reduce overall uncertainty
- Reduce *class-conditional* uncertainty

Why Reduce Inefficiency?

Remember:

- Coverage is guaranteed
- Inefficiency reflects uncertainty

{frog, dog, deer, horse}

{frog, dog, deer}

{frog, deer}

{deer}

reduced inefficiency = lower uncertainty translates to better resource/time usage to users

Optimizing Inefficiency

Train to directly reduce inefficiency:

$$\Omega(C_\theta(x)) = \sum_{k=1}^{K} C_{\theta,k}(x)$$

- $C_{\theta,k}(x) \in [0, 1]$ interpreted as "soft assignments"
- can be seen as smooth approximation of $\mathbb{E}[|C_\theta(x)|]$
- no loss on true label y as empirical coverage close to $(1 - \alpha)$

Optimizing Inefficiency

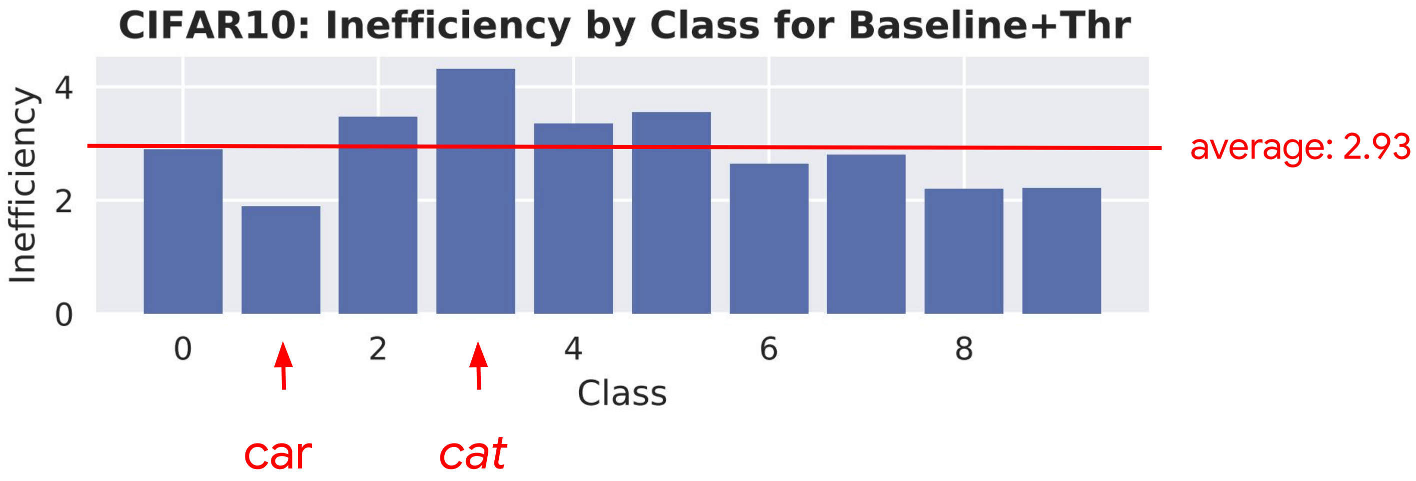
Inefficiency ↓ for = 0.01:			
CP at test time:	Thr		
Dataset	Cross-entropy baseline	ConfTr (ours)	
MNIST	2.23	**2.11** (-5.4%)	
F-MNIST	2.05	**1.67** (-18.5%)	
EMNIST (K = 52)	2.66	**2.49** (-6.4%)	
CIFAR10	2.93	**2.84** (-3.1%)	
CIFAR100	10.63	**10.44** (-1.8%)	

Optimizing Inefficiency

Inefficiency ↓ for = 0.01:				
CP at test time:	Thr		APS	
Dataset	Cross-entropy baseline	ConfTr (ours)	Cross-entropy baseline	ConfTr (ours)
MNIST	2.23	**2.11** (-5.4%)	2.50	**2.14** (-14.14%)
F-MNIST	2.05	**1.67** (-18.5%)	2.36	**1.72** (-27.1%)
EMNIST (K = 52)	2.66	**2.49** (-6.4%)	4.23	**2.87** (-32.2%)
CIFAR10	2.93	**2.84** (-3.1%)	3.30	**2.93** (-11.1%)
CIFAR100	10.63	**10.44** (-1.8%)	16.62	**12.73** (-23.4%)

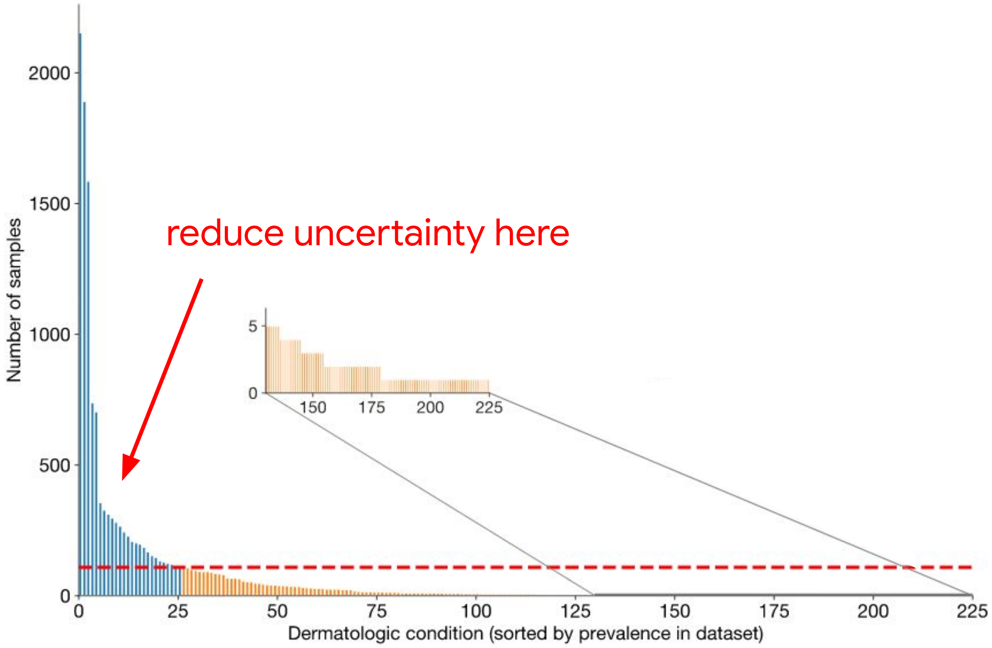
Inefficiency Distribution

Inefficiency ↓ distributed very differently across classes:



Reducing Class-Conditional Inefficiency

- Reduce inefficiency for “easy” / low-risk classes



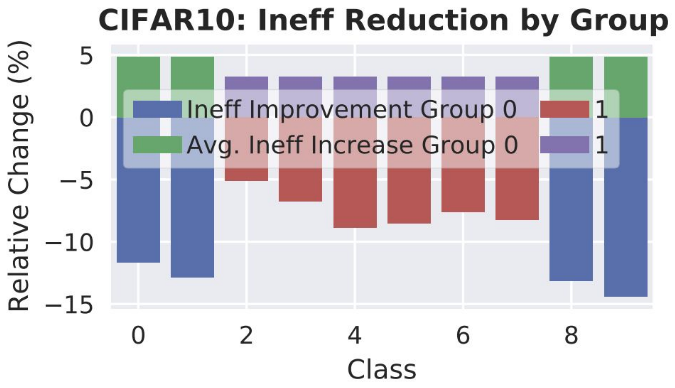
Roy et al. Does your dermatology classifier know what it doesn't know? Detecting the long-tail of unseen conditions. Medical Image Anal., 2022.

Reducing Class-Conditional Inefficiency

- Possible inefficiency improvement per class (in %)
- Cost in terms of average inefficiency increase across classes (in %)

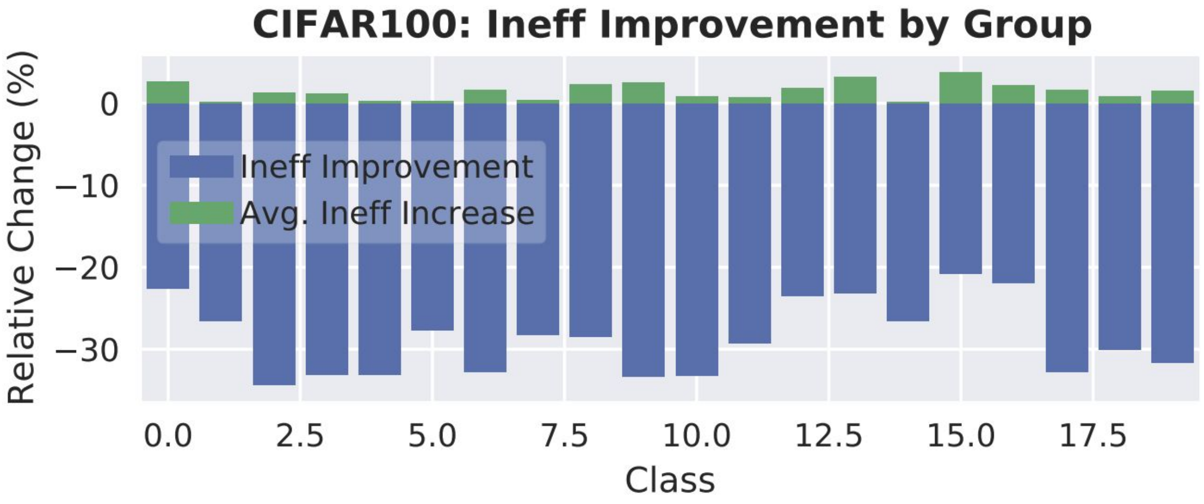
Results: CIFAR10

- Possible inefficiency improvement per class (in %)
- Cost in terms of average inefficiency increase across classes (in %)



More on Class-Conditional Inefficiency

- Possible inefficiency improvement per class (in %)
- Cost in terms of average inefficiency increase across classes (in %)



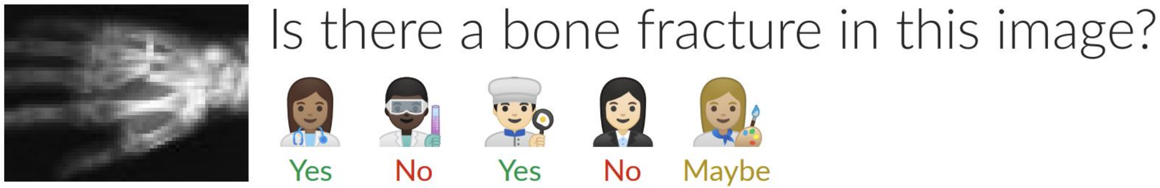
Training Objectives

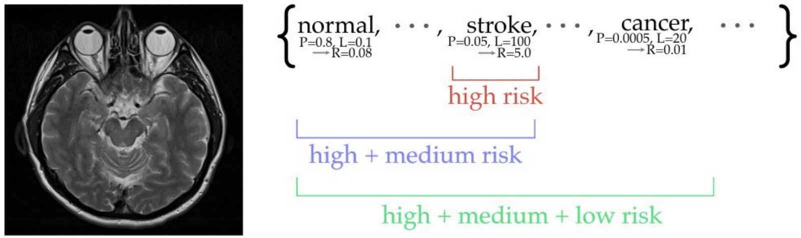
B Influencing the composition of confidence sets:

- Avoiding coverage confusion
- Reducing mis-coverage

Beyond Reducing Inefficiency

- Shape composition of confidence sets:
  - Avoid confusion of specific, easily confused classes
  - Avoid mixing classes of different categories





Platanios et al. Learning from Imperfect Annotations. ArXiv, 2020.

Beyond Reducing Inefficiency

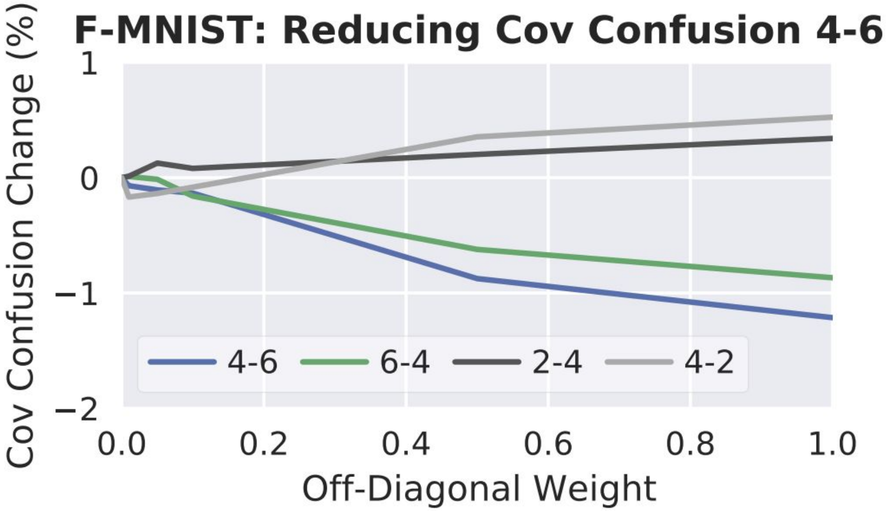
- Which classes are actually included in $C_\theta(x)$?

$$\underbrace{\Omega(C_\theta(x))}_{\text{Ineff loss}} + \sum_{k=1}^{K} L_{y,k} \underbrace{[(1 - C_{\theta,k}(x))\delta[y = k]}_{\text{True class included}} + \underbrace{C_{\theta,k}(x)\delta[y \neq k]]}_{\text{Other classes } \textit{not} \text{ included}}$$

- “just” enforces coverage with $L = I_K$
- use $L_{y,k} > 0$ to penalize class k occurring in confidence sets of class y

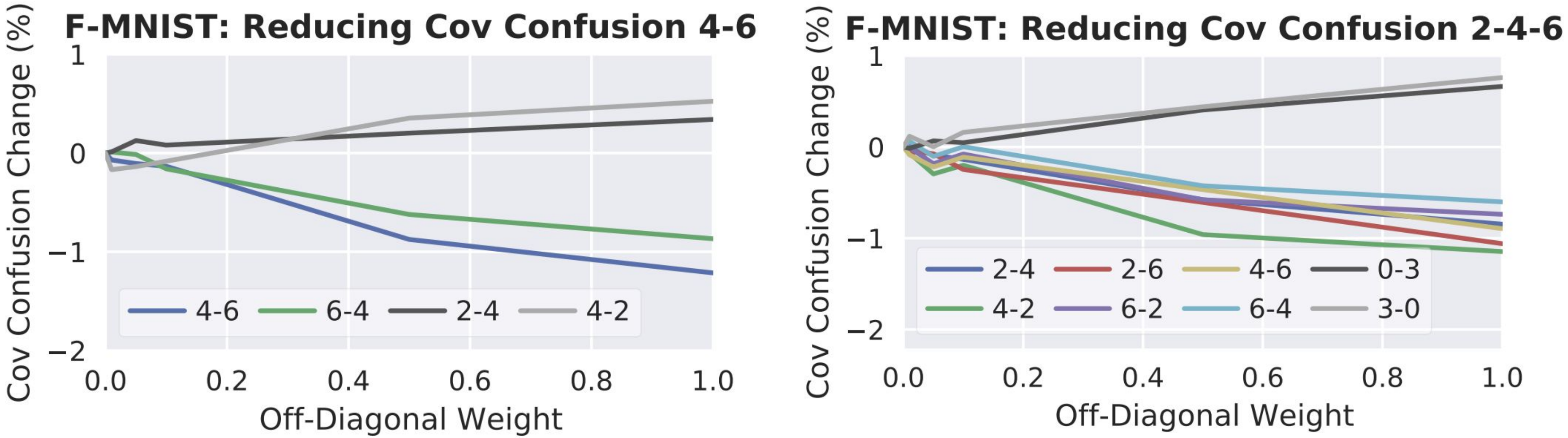
Example: Reduce Coverage Confusion

Reduce confusion between 4 (coat) and 6 (shirt) in confidence sets:



Example: Reduce Coverage Confusion

Reduce confusion between 2 (pullover), 4 and 6 in confidence sets:



Example: Reduce Mis-Coverage

Avoid natural and human-made classes in the same confidence sets:

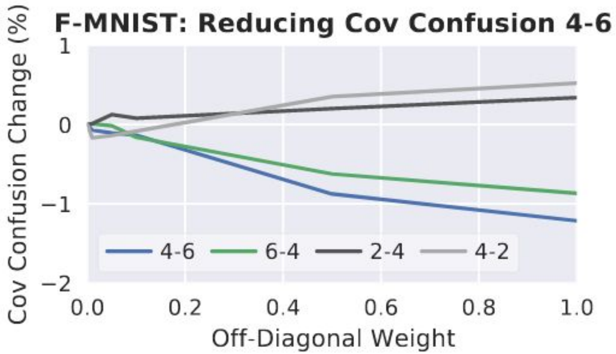
CIFAR100	Inefficiency	% *natural* classes in human-made confidence sets	% human-made classes in *natural* confidence sets
ConfTr	**10.44**	40.09	29.60
$L_{\text{human-made,natural}} > 0$	16.50	**15.77**	70.26
$L_{\text{natural,human-made}} > 0$	11.35	45.37	**17.56**

Conclusion for Conformal Training

= end-to-end training of classifier and conformal wrapper.

- retains coverage guarantee
- reduces inefficiency
- allows arbitrary, application-specific losses

Paper: arxiv.org/abs/2110.09192 | github.com/deepmind/conformal_training



Future Work Ideas for Conformal Training

- Extend conformal training beyond split conformal prediction to cross-conformal or transductive settings (better sample efficiency)
- "Conformal risk training": apply conformal training to arbitrary risks
- Semi-supervised conformal training (labels not needed on full batch)
- Integrate approaches for conditional coverage with conformal training
- Scale conformal training to larger models (training from scratch?)
- …

More research ideas:

davidstutz.de/some-research-ideas-for-conformal-training

Ambiguous Ground Truth

Conformal training:

- Notation and background

Conformal training:

- How to better integrate conformal prediction with deep learning?
- Improve “efficiency” or application-specific losses

Paper: arxiv.org/abs/2110.09192

Ambiguous ground truth:

- How to deal with ambiguous/uncertain ground truth?
- For example, when annotators disagree

Paper: arxiv.org/abs/2307.09302

The Hidden Assumption

Conformal prediction requires exchangeable $x, y \sim p(x, y) = p(y|x)p(x)$

The Hidden Assumption

Conformal prediction requires exchangeable $x, y \sim p(x, y) = p(y|x)p(x)$

Unknown true label

?

Observation

True label covered?

Confidence set

{bird, dog}

The Hidden Assumption

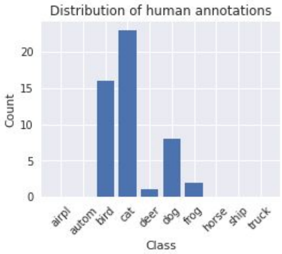
Conformal prediction requires exchangeable $x, y \sim p(x, y) = p(y|x)p(x)$

Unknown true label

?

Observation

Annotations



Majority vote

“cat”

No!

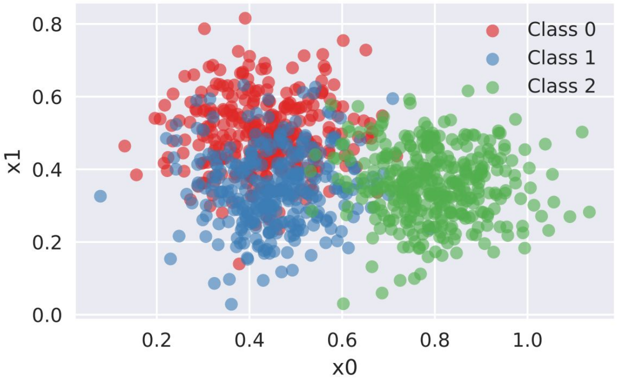
Confidence set

{bird, dog}

- We have access to labels $y_{\text{vote}} \sim p_{\text{vote}}(y|x)$
- But does ”$p_{\text{vote}} = p$” hold so we can guarantee coverage w.r.t. p?

An Intuitive Example

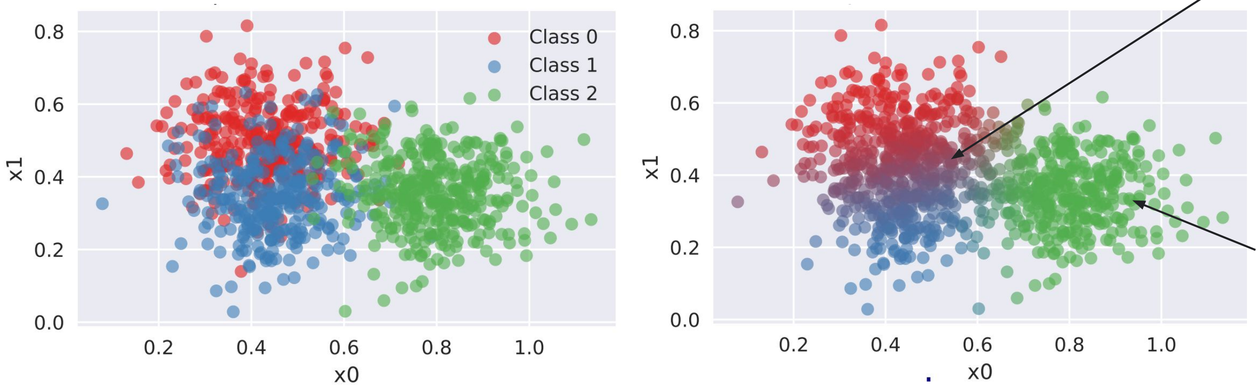
$$x, y \sim p(x, y)$$



- In practice, we never observe these true labels
(we cannot calibrate against them or obtain coverage against them)

An Intuitive Example

$x, y \sim p(x, y)$ $\qquad$ $p(y|x)$



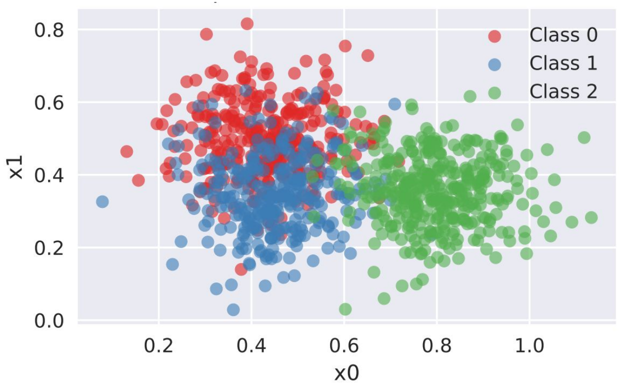
ambiguous example

“crisp” example

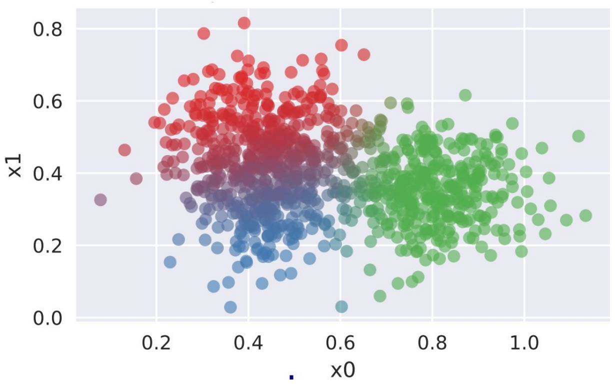
- Ambiguity is captured in the true posteriors $p(y|x)$
- In practice, we usually do not observe the true posteriors either

An Intuitive Example

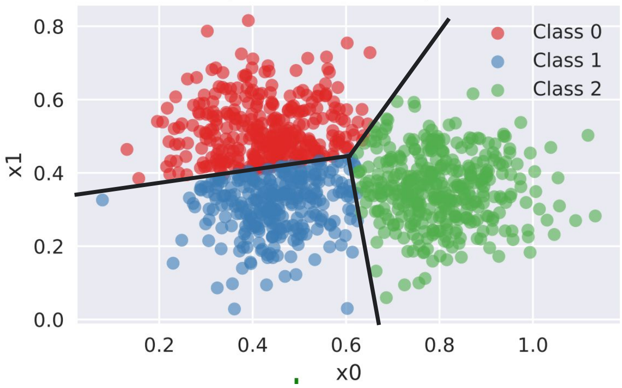
$x, y \sim p(x, y)$



$p(y|x)$



$x, y \sim p_{\text{vote}}(x, y)$



- The "majority voted" (i.e., top-1) label $y_{\text{vote}} \sim p_{\text{vote}}(y|x)$ ignores uncertainty

$p_{\text{vote}} \neq p$

- We can calibrate and obtain coverage against

A Serious Example

Observation

Annotations

b[1]: {*Pyogenic granuloma*} {*Hemangioma*} {*Melanoma*}
b[2] {*Angiokeratoma of skin*} {*Atypical Nevus*}
b[3]: {*Hemangioma*} {*Melanocytic Nevus, Melanoma, O/E - ecchymoses present*}
b[4]: {*Hemangioma, Melanoma, Skin Tag*}
b[5]: {*Melanoma*}
b[6]: {*Hemangioma*} {*Melanoma*} {*Melanocytic Nevus*}

Majority vote

Hemangioma
= benign

A Serious Example

Observation | Annotations | Majority vote

b[1]**:** *{Pyogenic granuloma} {Hemangioma} {Melanoma}*
b[2] *{Angiokeratoma of skin} {Atypical Nevus}*
b[3]**:** *{Hemangioma} {Melanocytic Nevus, Melanoma, O/E - ecchymoses present}*
b[4]**:** *{Hemangioma, Melanoma, Skin Tag}*
b[5]**:** *{Melanoma}*
b[6]**:** *{Hemangioma} {Melanoma} {Melanocytic Nevus}*

Hemangioma = benign

- Shouldn't we at least check for *Melanoma* = cancerous?

Ignoring Ambiguity has Consequences

Calibrating against labels from p_{vote} misses *plausible* conditions:

Standard CP

{<u>Hemangioma</u>, Melanocytic Nevus, Atypical Nevus}

Do we consider CP successful when it includes the voted label?

Embracing Ambiguity in Conformal Prediction

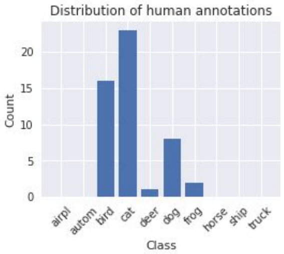
Use the annotations directly – for example, in terms of frequencies:

Unknown true label

?

Observation

Annotations



Majority vote

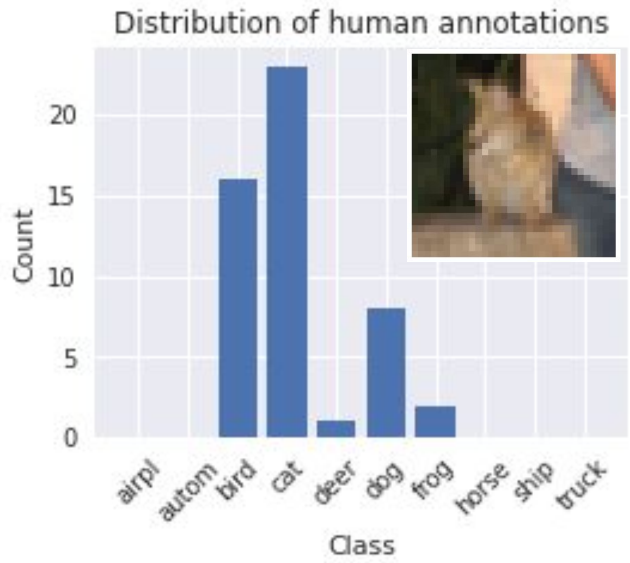
"cat"

$p_{\text{agg}} \approx p$

$\neq p_{\text{vote}}$

- Aggregating the annotations is our best option to approximate the true distribution p
(we can only be as good in this tasks as our expert annotators are)

For a Single Example



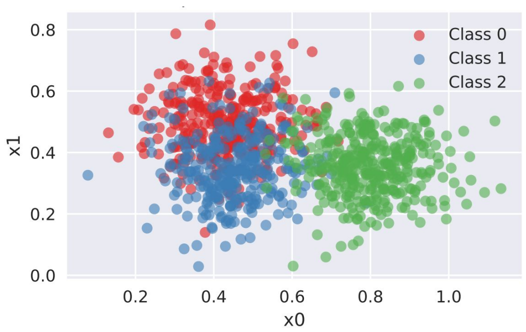
Annotator frequencies $\lambda = p_{\text{agg}}(y = k|x)$

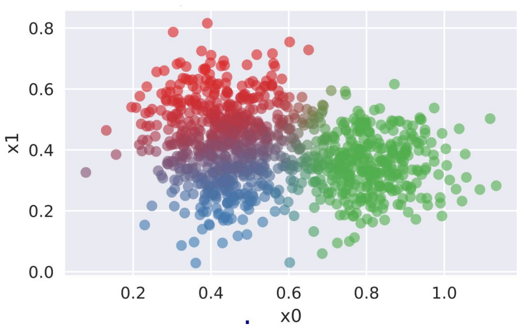
λ = (0, 0, 0.32, 0.46, 0.02, 016, 0.04, 0, 0, 0)

$C(x)$ = {cat, dog} – do we have coverage?

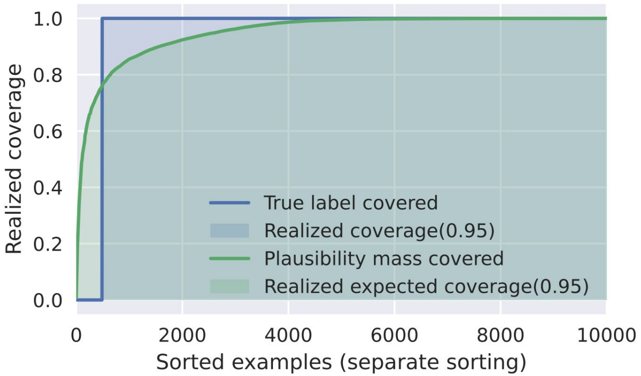
Voted/top-1 coverage	1
“Plausibility mass covered”	0.62 = 0.46 + 0.16

Across Examples





Evaluation
with true label



Evaluation
with
plausibilities

Evaluating Coverage with Plausibilities

Call estimates of $\lambda_{ik} = p_{\text{agg}}(y = k|x_i) \approx p(y|x_i)$ plausibilities:

- Guaranteeing anything for $p(y \in C(x))$ is impossible!
- We *can* guarantee $p_{\text{vote}}(y \in C(x))$
- Best we can hope to do:

$$p_{\text{agg}}(y \in C(x))$$

Guarantee coverage “against annotations”

Evaluating Coverage with Plausibilities

Call estimates of $\lambda_{ik} = p_{\text{agg}}(y = k|x_i) \approx p(y|x_i)$ plausibilities:

- Guaranteeing anything for $p(y \in C(x))$ is impossible!
- We *can* guarantee $p_{\text{vote}}(y \in C(x))$
- Best we can hope to do:

$$p_{\text{agg}}(y \in C(x)) = \mathbb{E}_{p_{\text{agg}}}[\delta[y \in C(x)]]$$

Binary event, express as expectation

Evaluating Coverage with Plausibilities

Call estimates of $\lambda_{ik} = p_{\text{agg}}(y = k|x_i) \approx p(y|x_i)$ plausibilities:

- Guaranteeing anything for $p(y \in C(x))$ is impossible!
- We *can* guarantee $p_{\text{vote}}(y \in C(x))$
- Best we can hope to do:

$$
\begin{aligned}
p_{\text{agg}}(y \in C(x)) &= \mathbb{E}_{p_{\text{agg}}}[\delta[y \in C(x)]] \\
&= \mathbb{E}_{x,y \sim p(x) p_{\text{agg}}(y|x)}[\delta[y \in C(x)]]
\end{aligned}
$$

Decompose joint probability

Evaluating Coverage with Plausibilities

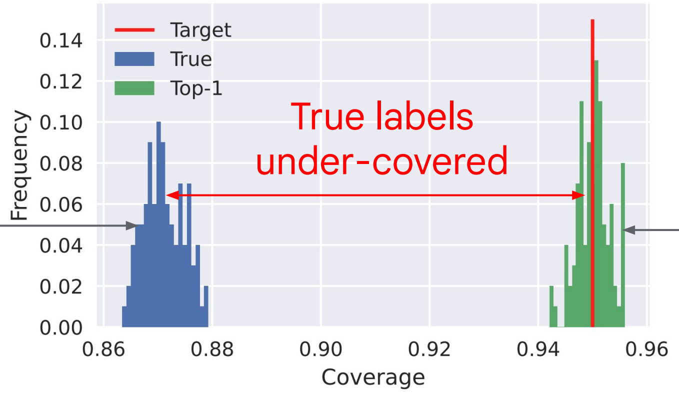
Call estimates of $\lambda_{ik} = p_{\text{agg}}(y = k|x_i) \approx p(y|x_i)$ plausibilities:

- Guaranteeing anything for $p(y \in C(x))$ is impossible!
- We *can* guarantee $p_{\text{vote}}(y \in C(x))$
- Best we can hope to do:

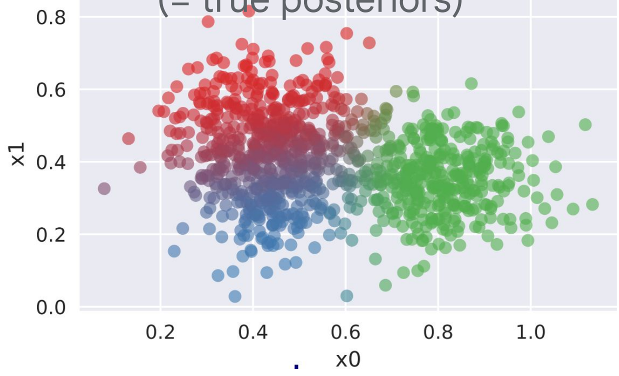
$$\begin{aligned} p_{\text{agg}}(y \in C(x)) &= \mathbb{E}_{p_{\text{agg}}}[\delta[y \in C(x)]] \\ &= \mathbb{E}_{x,y\sim p(x)p_{\text{agg}}(y|x)}[\delta[y \in C(x)]] \\ &= \mathbb{E}_{x\sim p(x)}[\underbrace{\mathbb{E}_{y\sim p_{\text{agg}}(y|x)}[\delta[y \in C(x)]]}_{\sum_k \lambda_k \delta[k \in C(x)]}] \end{aligned}$$

Calibrating with Voted/Top-1 Labels

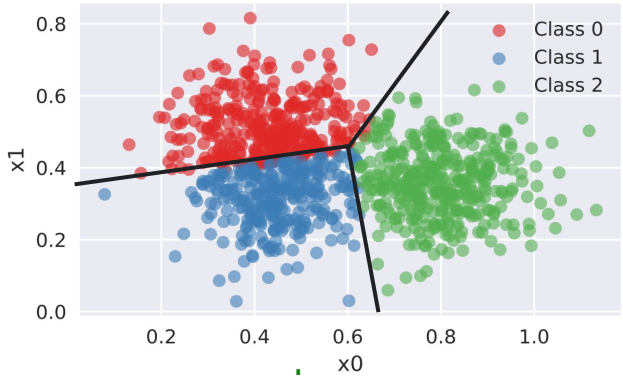
Calibrated w/ voted labels



(True) plausibilities
(= true posteriors)



Voted/top-1 labels



Monte Carlo Conformal Prediction

Monte Carlo conformal prediction:

- Use plausibilities for calibration: $\lambda_{ik} = p_{\text{agg}}(y = k|x_i) \approx p(y|x_i)$
- Repeat each calibration example M times
- Calibrate using the *augmented* calibration set

$$\{E(x_{ij}, y_{ij})\}_{i\in[N],j\in[M]} \qquad \text{with} \qquad y_{ij} \sim p_{\text{agg}}(y_{ij} = k|x_i) = \lambda_{ik}$$

- Adjust quantile computation to

$$\frac{\lfloor \alpha M(N+1) \rfloor - M + 1}{MN}$$

Coverage Guarantee

Monte Carlo conformal prediction breaks exchangeability for $M > 1$

- Can re-formulate as averaging M p-values
- This establishes a $1 - 2\alpha$ coverage guarantee
- Can improve to $(1-\alpha)(1-\delta)$ for $\delta > 0$ with additional calibration split
- Coverage w.r.t. $p_{\text{agg}} \approx p$ (the best we can do given the annotations)

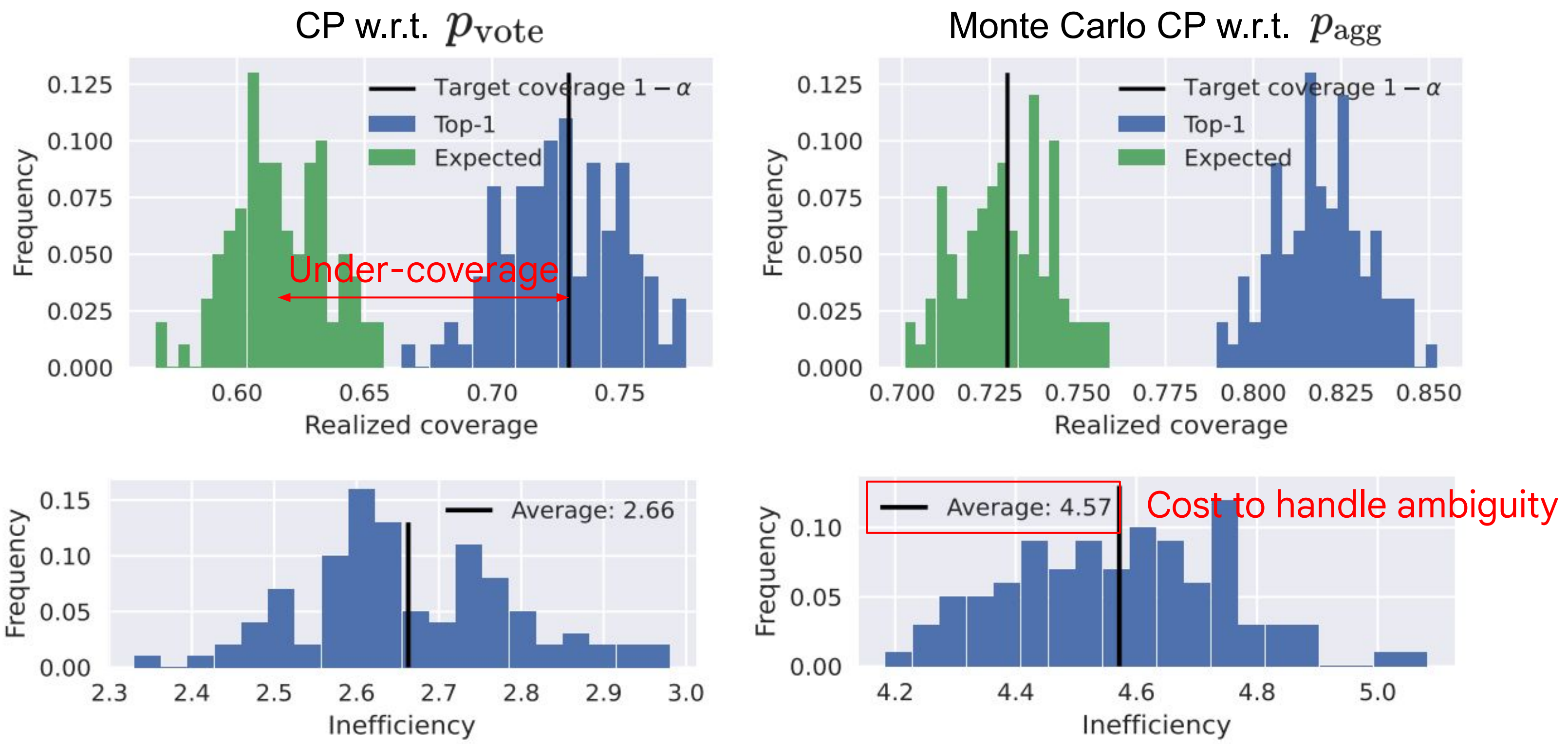
V. Vovk, Ruodu Wang. Candes. Combining P-Values Via Averaging. Biometrika, 2018.

Properties and Remarks

Some nice properties:

- Empirically, we always observe coverage $1 - \alpha$
- Without ambiguity, we recover standard conformal prediction (any M)
- Ambiguous examples, we improve coverage by sacrificing efficiency
- Unambiguous examples, it behaves like standard conformal prediction
- Also establishes coverage guarantee for multi-label classification and calibration with data augmentation

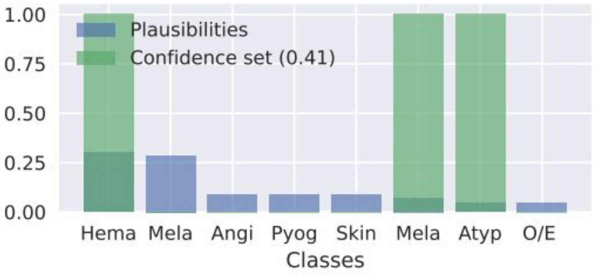
Results in Dermatology

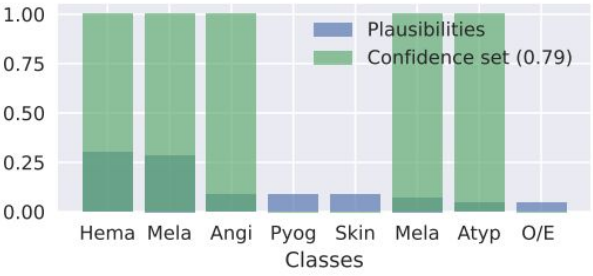


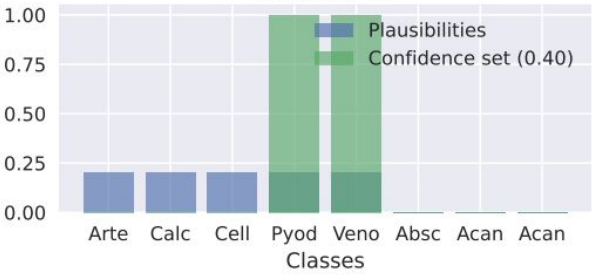
Qualitative Results in Dermatology

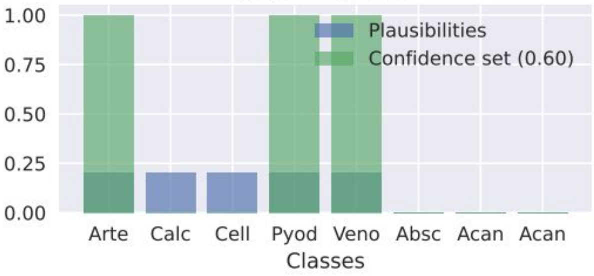
CP w.r.t. p_{vote}

Monte Carlo CP w.r.t. p_{agg}









Conclusion for Monte Carlo CP

= conformal prediction based on sampled labels from annotators/plausibilities.

- The labels we have access to are usually voted labels, from p_{vote}
- In ambiguous settings, voted labels can deviate from true labels:

$$p_{\text{vote}} \neq p$$

- Monte Carlo conformal prediction samples labels from $p_{\text{agg}} \approx p$
- Natural extension of standard conformal prediction to ambiguous tasks

Paper: arxiv.org/abs/2307.09302

Future Work for Monte Carlo CP

- Extension to conformal risk control with ambiguity
- Conformal prediction in ambiguous regression tasks (where plausibilities are not categorical distributions but could be modeled using various distributions, empirical or model-based)
- Conditional coverage on ambiguous examples
- …

Questions?

Conformal training:

- End-to-end training of deep models *for* conformal prediction
- Improve "efficiency" or application-specific losses

Paper: arxiv.org/abs/2110.09192

Monte Carlo conformal prediction:

- Calibrate and guarantee coverage on examples with ambiguous ground truth

Paper: arxiv.org/abs/2307.09302

Reach out: davidstutz.de | dstutz@google.com