Proprietary + Confidential

Google DeepMind

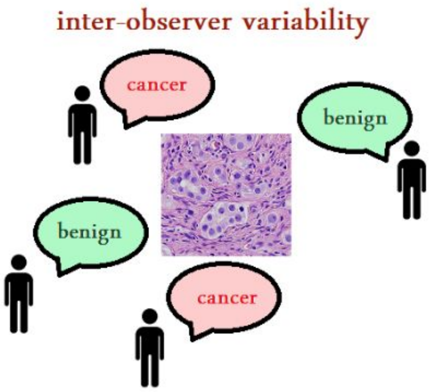
Conformal prediction under ambiguous ground truth

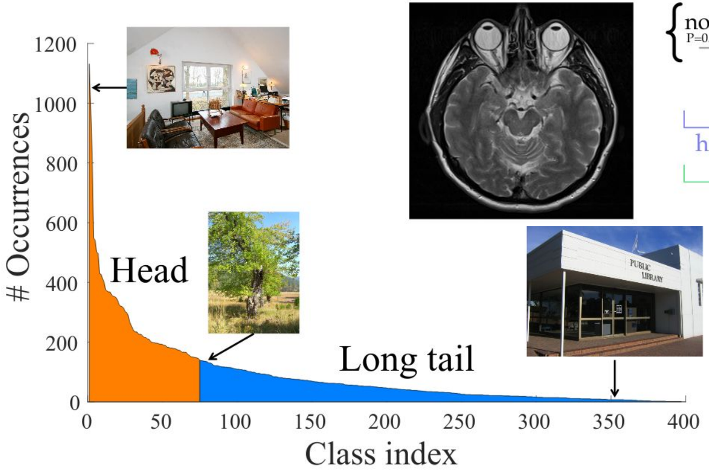
David Stutz

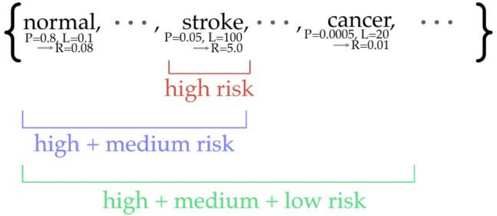
Feb 19th 2024

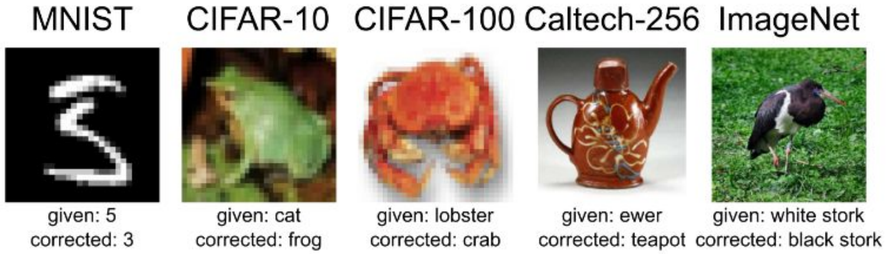
Motivation: Uncertainty Estimation in Classification

- 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.

Proprietary + Confidential

Talk Outline

Conformal prediction:

- Notation and background

Monte Carlo conformal prediction:

- Where does our ground truth for calibration come from?
- What if this ground truth is uncertain because annotators disagree?
- How can we handle this during calibration?

Paper: arxiv.org/abs/2307.09302

Proprietary + Confidential

Conformal Prediction

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

$$p(y \in C(x)) \geq 1 - \alpha \qquad \text{(\textit{coverage} guarantee)}$$

- confidence level α user-specified

Proprietary + Confidential

Conformal Prediction

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

$$p(y \in C(x)) \geq 1 - \alpha \qquad \text{(}\textit{coverage}\text{ guarantee)}$$

- confidence level α user-specified
- *inefficiency* = average confidence set size $|C(x)|$
- requires *exchangeability*, independent of model and distribution
- coverage marginal over examples!

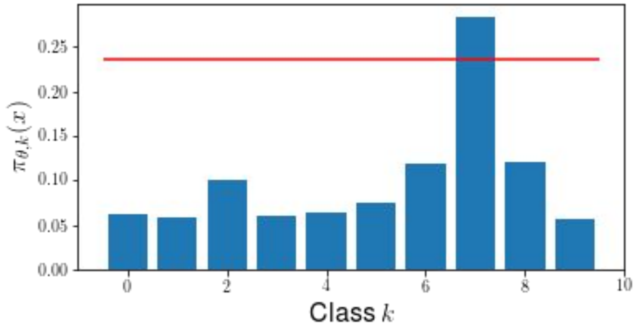
Split Conformal Prediction

Split conformal prediction with two steps: prediction and calibration:

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

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

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.

Proprietary + Confidential

Split Conformal Prediction

Split conformal prediction with two steps: prediction and calibration:

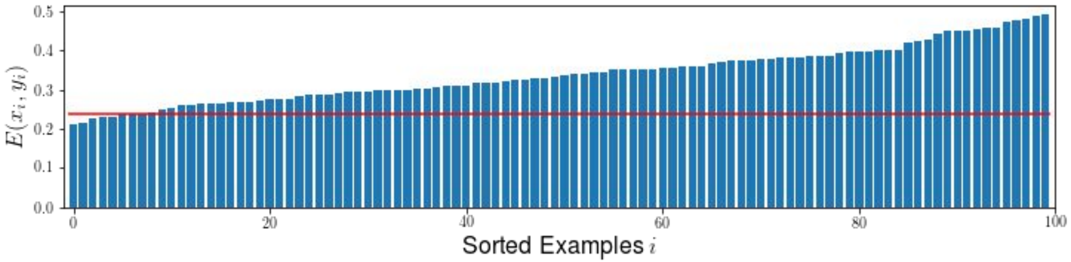
1. Prediction (test time): define how confidence sets are constructed

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

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

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

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



Conformal p-values

Alternative view (will be important later):

1. We test the null hypothesis that k is the true label of test example x:

$$H_k : y = k$$

2. Compute a p-value for this hypothesis using:

$$\rho_k = \frac{\sum_{i=1}^{N} \delta[E(x_i, y_i) \leq E(x, k)] + 1}{N + 1}$$

3. Construct confidence set

$$C(x) = \{k \in [K] : \rho_k \geq \alpha\}$$

Proprietary + Confidential
Public

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




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

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

Talk Outline

Conformal prediction:

- Notation and background

Monte Carlo conformal prediction:

- Where does our ground truth for calibration come from?
- What if this ground truth is uncertain because annotators disagree?
- How can we handle this during calibration?

Paper: arxiv.org/abs/2307.09302

Obtaining Calibration Labels

Need conformity scores of the true labels $E(x_i, y_i)$ for $x_i, y_i \sim p(x_i, y_i)$:

Unknown
true label

?

y_i

Observation

x_i

y_i = bird?

Proprietary + Confidential

Obtaining Calibration Labels

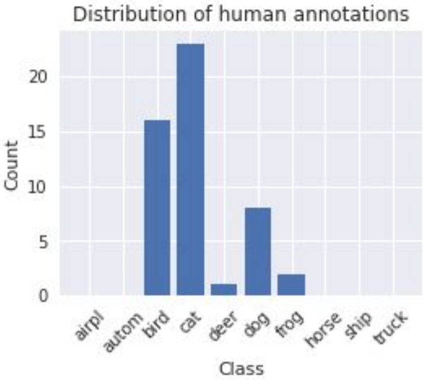
Need conformity scores of the true labels $E(x_i, y_i)$ for $x_i, y_i \sim p(x_i, y_i)$:

Unknown
true label

?

Observation

Annotations



Majority vote

cat ~~bird?~~

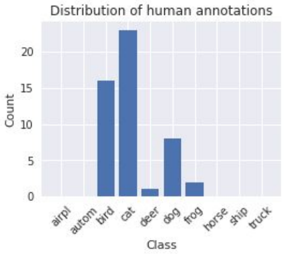
Proprietary + Confidential

Calibration Against Majority Voted Labels

Need conformity scores of the true labels $E(x_i, y_i)$ for $x_i, y_i \sim p(x_i, y_i)$:

Observation

Annotations

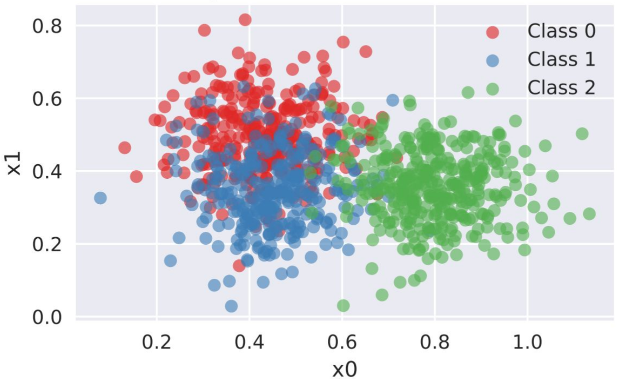


Majority vote $\longrightarrow$ cat

- We have access to majority voted labels $y_{\text{vote}} \sim p_{\text{vote}}(y|x)$
- For this example, clearly $p_{\text{vote}} \neq p$
- But we need "$p_{\text{vote}} = p$" to guarantee coverage w.r.t. p

A Simple Example

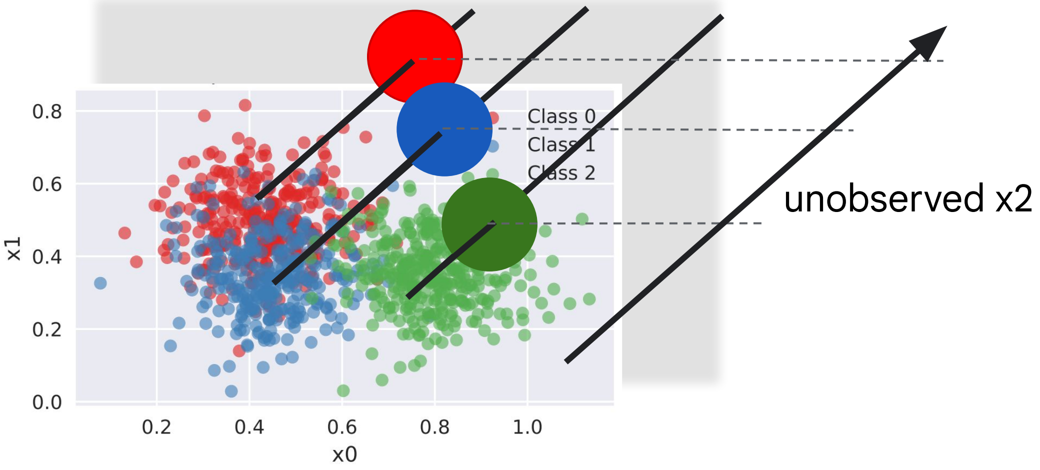
$$x, y \sim p(x, y)$$



Proprietary + Confidential

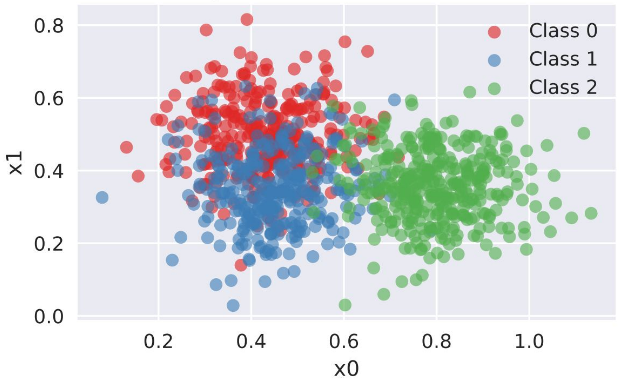
A Simple Example

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



A Simple Example

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

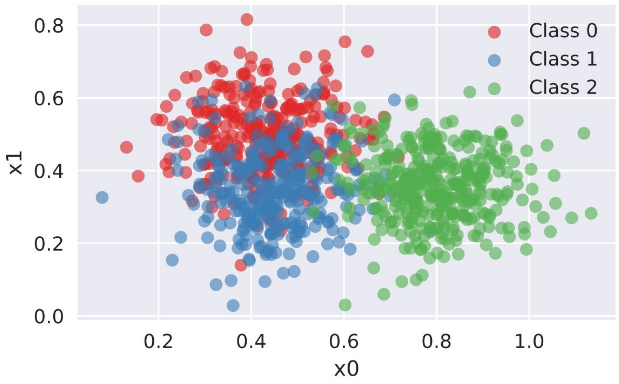


- In practice, we never observe these true labels
  (we cannot calibrate against them or obtain coverage against them)

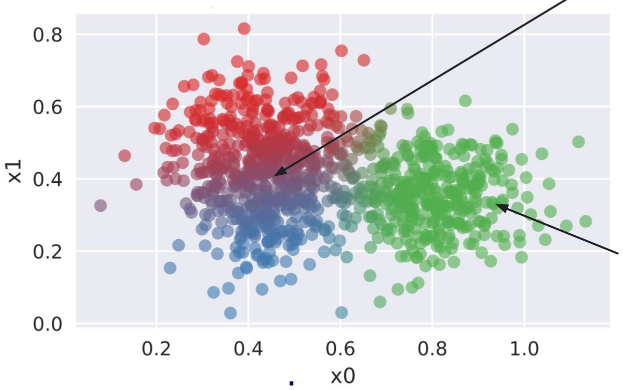
Proprietary + Confidential

A Simple Example

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



$$p(y|x)$$



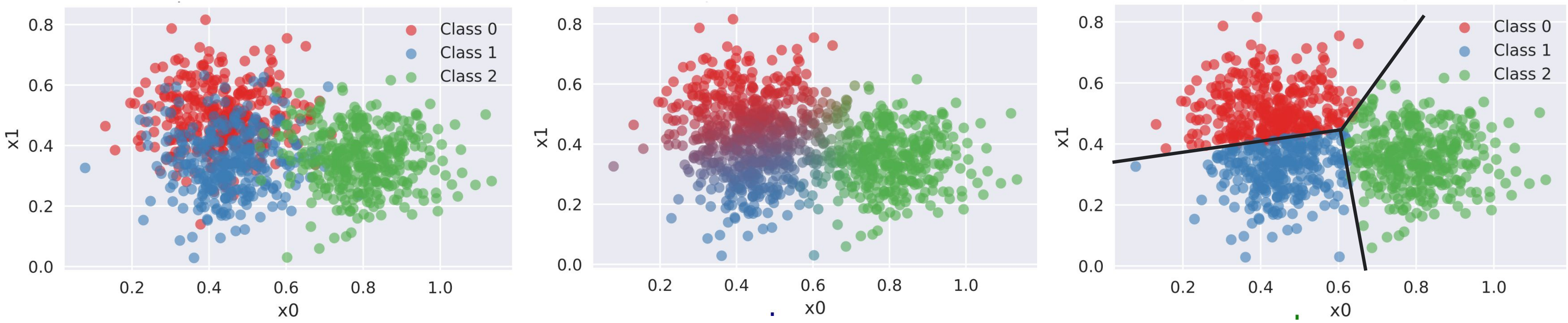
ambiguous example

“crisp” example

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

A Simple Example

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



- The “majority voted” label $y_{\text{vote}} \sim p_{\text{vote}}(y|x)$ ignores uncertainty
- We can calibrate and obtain coverage against $p_{\text{vote}} \neq p$

A Serious Example

Observation

Annotations

b¹: {*Pyogenic granuloma* (Low)} {*Hemangioma* (Med)} {*Melanoma* (High)}
b² {*Angiokeratoma of skin* (Low)} {*Atypical Nevus* (Med)}
b³: {*Hemangioma* (Med)} {*Melanocytic Nevus* (Low), *Melanoma* (High), *O/E - ecchymoses present* (Low)}
b⁴: {*Hemangioma* (Med), *Melanoma* (High), *Skin Tag* (Low)}
b⁵: {*Melanoma* (High)}
b⁶: {*Hemangioma* (Med)} {*Melanoma* (High)} {*Melanocytic Nevus* (Low)}

Conditions, Low/Med/High risk conditions

A Serious Example

Observation

Annotations

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

Conditions, Low/Med/High risk conditions

Majority vote

Hemangioma

A Serious Example

Observation

Annotations

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

Conditions, Low/Med/High risk conditions

Majority vote

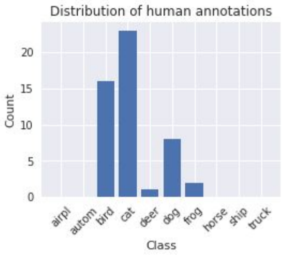
Hemangioma

Ignores a cancerous condition

Embracing Ambiguity in Conformal Prediction

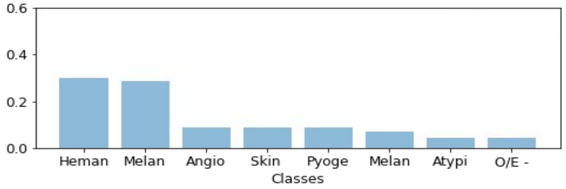
Use annotations directly – for example, in terms of aggregated frequencies:

Observation | Agg. Annotations



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

Observation | Agg. Annotations

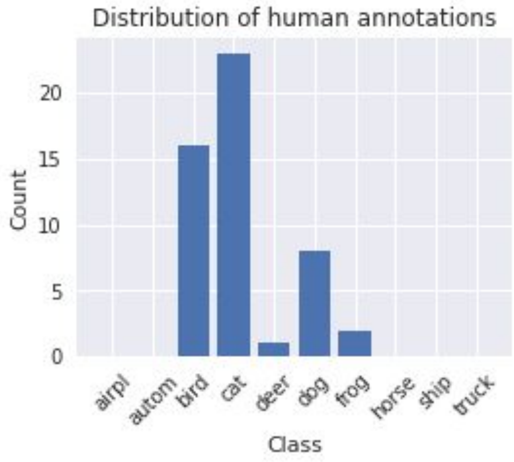


$$p_{\text{agg}} \approx p$$

- Aggregating annotations is our best option to approximate the true p (we can only be as good in this tasks as our expert annotators are)
- How can we calibrate for and evaluate coverage w.r.t. p_{agg} ?

Aggregated Coverage for a Single Example

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



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

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

Majority-voted coverage	1
Aggregated coverage	0.62 = 0.46 + 0.16

"Covered plausibility mass"

Aggregated Coverage with Plausibilities

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

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

Guarantee coverage “against annotations”

Aggregated Coverage with Plausibilities

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

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

Binary event, express as expectation

Proprietary + Confidential

Aggregated Coverage with Plausibilities

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

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

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

Decompose joint probability

Aggregated Coverage with Plausibilities

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

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

Distribute coverage across plausibilities ⟶ $\sum_k \lambda_k \delta[k \in C(x)]$

Aggregated Coverage with Plausibilities

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

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

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

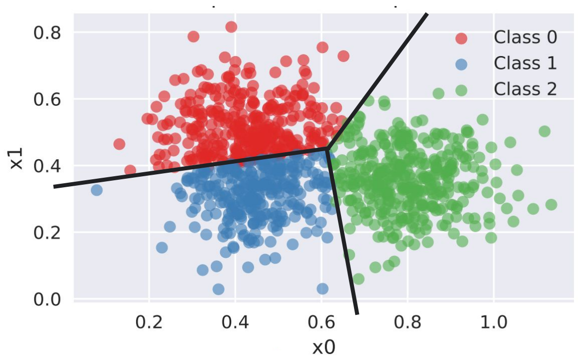
$$= \mathbb{E}_{x \sim p(x)}[\underbrace{\mathbb{E}_{y \sim p_{\text{agg}}(y|x)}[\delta[y \in C(x)]]}_{\sum_k \lambda_k \delta[k \in C(x)]}]$$

Distribute coverage across plausibilities $\longrightarrow$ $\sum_k \lambda_k \delta[k \in C(x)]$

- ➔ Coverage is marginal over examples and labels!
- ➔ If $p_{\text{agg}} = p$, this is coverage wrt. the true labels!

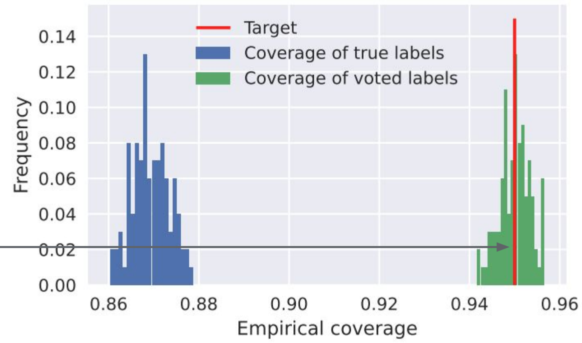
Calibrating with Voted Labels

Voted labels



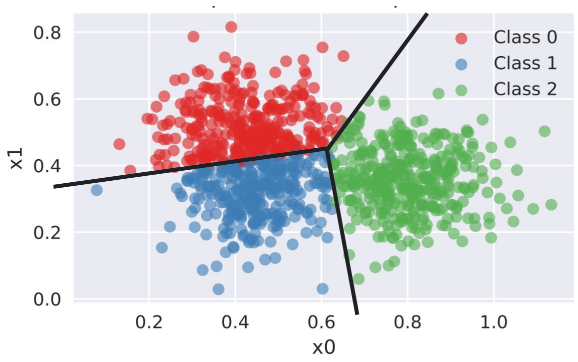
Evaluate “standard coverage” with voted labels

Calibrate with voted labels

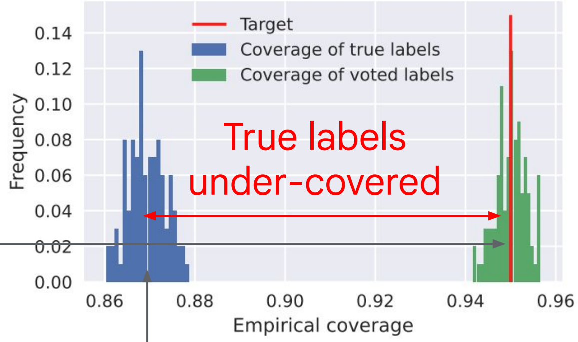


Calibrating with Voted Labels

Voted labels

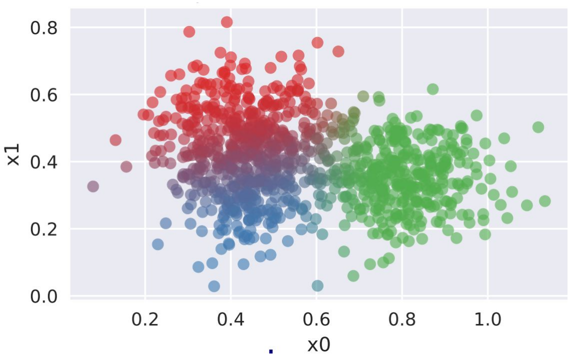


Calibrate with voted labels



Evaluate "standard coverage" with voted labels

Plausibilities = true posteriors



Evaluate *aggregated coverage* (= true coverage as $p_{\text{agg}} = p$)

Calibrate Against *Sampled* Labels

Basic idea:

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

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

Proprietary + Confidential

Calibrate Against *Sampled* Labels

Basic idea:

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

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

Problem:

- Invalidates coverage by breaking exchangeability:

$p(\underline{z_{11}, z_{12}, z_{13},} \ldots, z_{21}, \ldots, z_{NM}, z)$ for $z_{ij} = (x_i, y_{ij})$ and test example z

I know the first M examples are repeated

Monte Carlo Conformal Prediction

Solution:

- 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

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

- **Adjust quantile computation** to

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

Proprietary + Confidential

Obtaining Coverage $1 - 2\alpha$

Consider the p-values computed for standard conformal prediction:

$$\rho_k = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} \delta[E(x_i, y_{ij}) \leq E(x, k)] + 1}{M \cdot N + 1}$$

Fixing Coverage

Consider the p-values computed for standard conformal prediction:

$$\rho_k = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} \delta[E(x_i, y_{ij}) \leq E(x, k)] \cancel{+1}}{M \cdot N \cancel{+1}}$$

Fixing Coverage

Consider the p-values computed for standard conformal prediction:

$$\rho_k = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} \delta[E(x_i, y_{ij}) \leq E(x, k)] + 1}{M \cdot N + 1}$$

$$\bar{\rho}_k = \frac{1}{M} \sum_{j=1}^{M} \rho_k^j \longrightarrow \rho_k^j = \frac{\sum_{i=1}^{N} \delta[E(x_i, y_{ij}) \leq E(x, k)]}{N}$$

Fixing Coverage

Consider the p-values computed for standard conformal prediction:

$$\rho_k = \frac{\sum_{i=1}^N \sum_{j=1}^M \delta[E(x_i, y_{ij}) \leq E(x, k)] + 1}{M \cdot N + 1}$$

$$\bar{\rho}_k = \frac{1}{M} \sum_{j=1}^M \rho_k^j \longrightarrow \rho_k^j \overset{!}{=} \frac{\sum_{i=1}^N \delta[E(x_i, y_{ij}) \leq E(x, k)] + 1}{N + 1}$$

Proprietary + Confidential

Obtaining Coverage $1 - 2\alpha$

Consider the p-values computed for standard conformal prediction:

$$\rho_k = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} \delta[E(x_i, y_{ij}) \leq E(x, k)] + 1}{M(N+1)}$$

$$M(N+1)$$

$$\bar{\rho}_k = \frac{1}{M} \sum_{j=1}^{M} \rho_k^j \qquad \rho_k^j = \frac{\sum_{i=1}^{N} \delta[E(x_i, y_{ij}) \leq E(x, k)] + 1}{N+1}$$

➔ Vovk and Wang establish coverage $1 - 2\alpha$ when averaging p-values

Coverage Beyond $1 - 2\alpha$

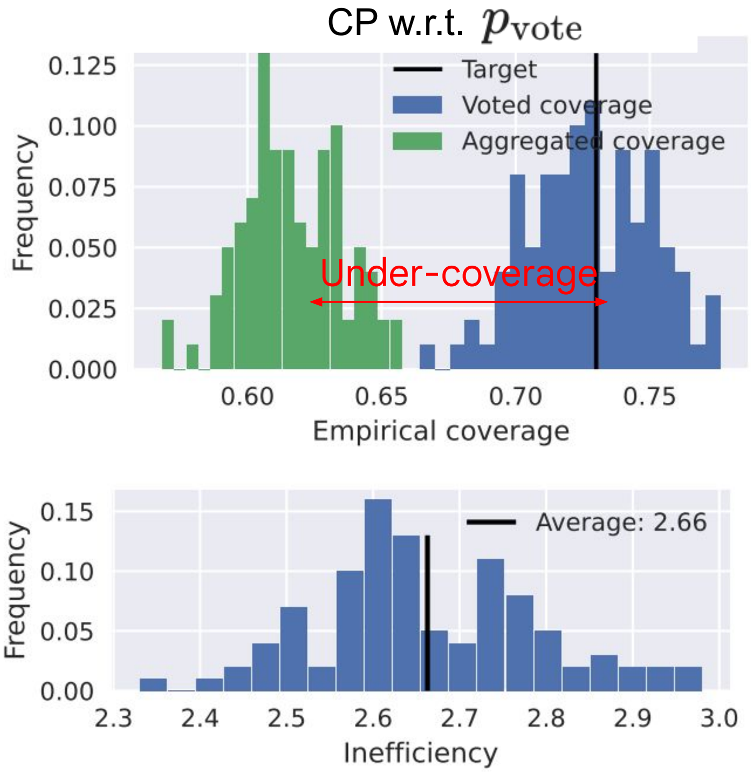
Monte Carlo conformal prediction:

- Can be re-formulated 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

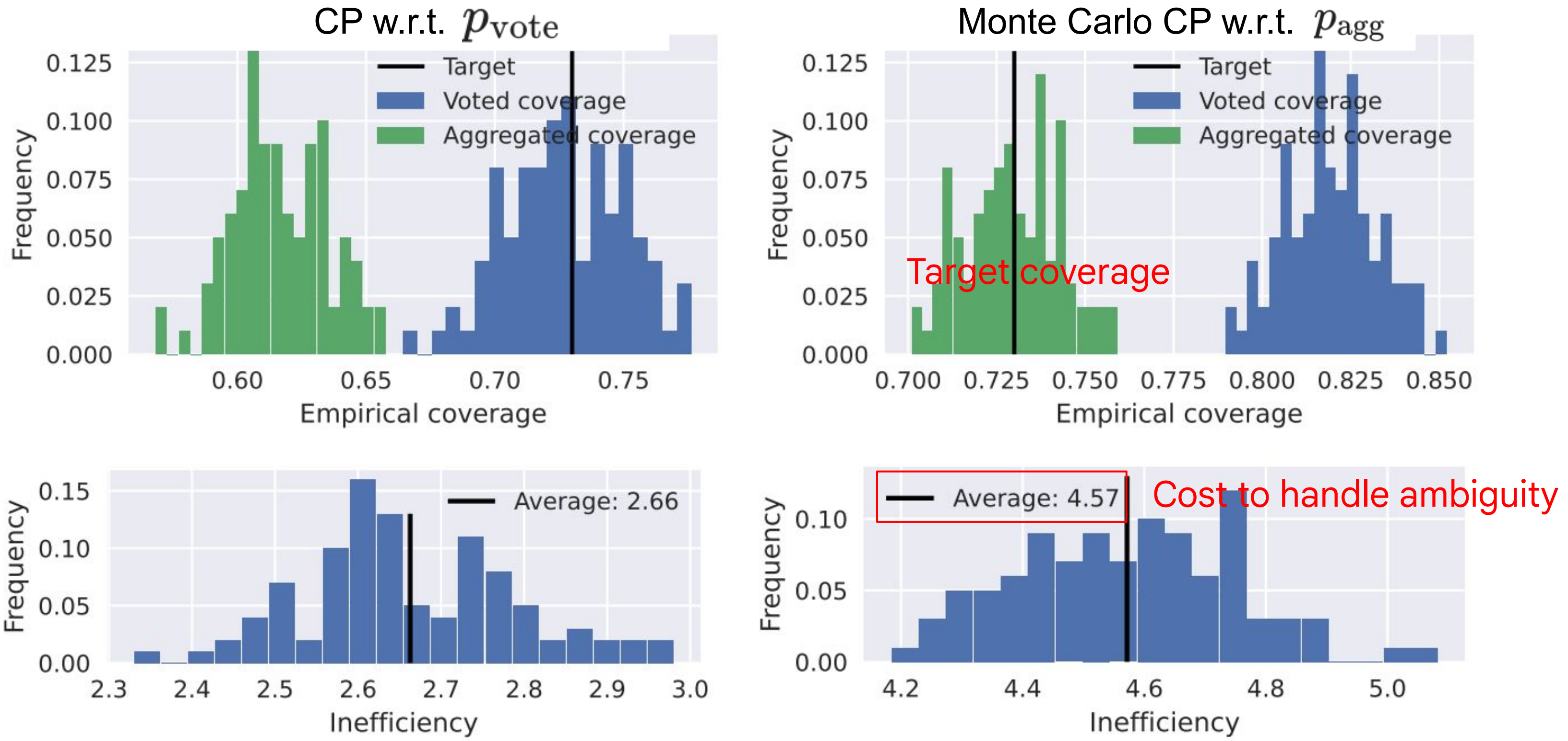
Remarks:

- 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

Results in Dermatology



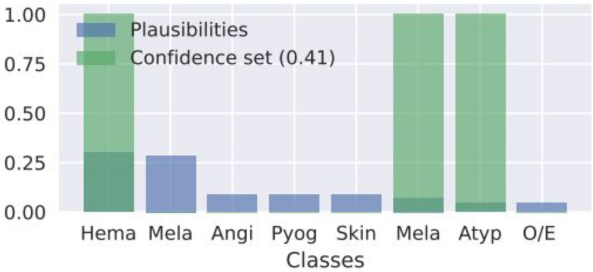
Results in Dermatology

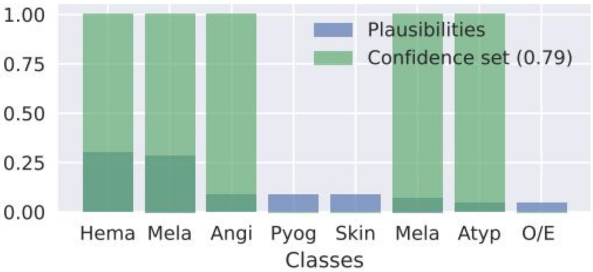


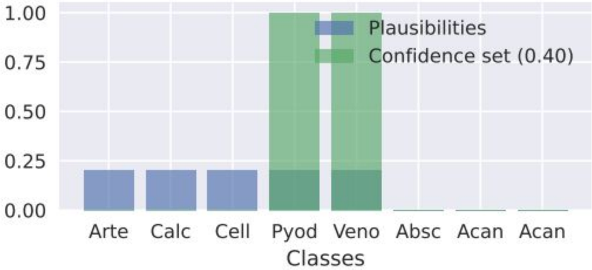
Qualitative Results in Dermatology

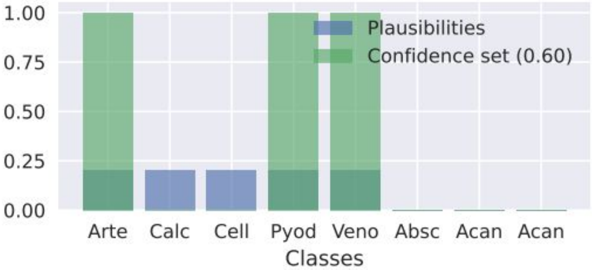
CP w.r.t. p_{vote}

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









Proprietary + Confidential

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 $p_{\text{agg}} \approx p$
- The best we can do: “calibrate wrt. to annotators”
- Establishes coverage guarantees for multi-label classification and calibration with data augmentation

Paper: arxiv.org/abs/2307.09302 | Contact: davidstutz.de / dstutz@google.com