DeepMind

Conformal prediction tutorial

dstutz@google.com

Outline

Overall outline:

- Introduction to uncertainty estimation
- Conformal prediction:
  - Theory and assumptions
  - Threshold conformal predictor
  - Understanding coverage
- Advanced topics
- Conclusion

Links:

- Conformal prediction tutorial (Angelopoulos and Bates)
- Conformal training
- Monte Carlo conformal prediction

DeepMind

Motivation

Why uncertainty estimation and calibration

Why quantify uncertainty?

- In medicine, science, and engineering, measurements are error-prone
  - Core ingredient for decision making
- In machine learning, “let me know when my model is wrong”
  - Confidence ≅ accuracy

Goal of this presentation:

- Introduce conformal prediction as principled uncertainty estimation technique
- Convince you that you should always calibrate if possible (not calibrating is a wasted opportunity)

Approaches to uncertainty estimation

Goal: quantify uncertainty using *sets* or *intervals* that "likely contain the true prediction"

Frequentist: confidence set/interval

Probability = frequency of repeated events

For machine learning:

- Examples are modeled as random
- Parameters are fixed

Confidence set = there is a X% probability that, when constructing confidence sets/intervals from data "like this", the true value will be included

Bayesian: credibility set/interval

Probability = degree of certainty about values

For machine learning:

- Data is fixed
- Parameters are modeled as random

Credibility set = given the data, there is a X% probability that the true value is included (assuming our model assumption is correct)

Approaches to uncertainty estimation

Goal: quantify uncertainty using *sets* or *intervals* that "likely contain the true prediction"

Frequentist: confidence set/interval

Probability = frequency of repeated events

For machine learning:

- Examples are modeled as random
- Parameters are fixed
- More like aleatoric uncertainty

Confidence set = there is a X% probability that, when constructing confidence sets/intervals from data "like this", the true value will be included

Bayesian: credibility set/interval

Probability = degree of certainty about values

For machine learning:

- Data is fixed
- Parameters are modeled as random
- More like epistemic uncertainty

Credibility set = given the data, there is a X% probability that the true value is included (assuming our model assumption is correct)

Conformal prediction

The "theory": coverage guarantee

Given a (fixed) model $\pi_y(x) \approx p(y|x)$, a set of *exchangeable* calibration examples $(x_i, y_i), \ldots, (x_n, y_n)$ and a test example x_{n+1}, construct a confidence set $C(x_{n+1}) \subseteq [K]$ of labels that contains the true labels y_{n+1} with high probability:

$$p_{x_1,\ldots,x_{n+1}}(y_{n+1} \in C(x_{n+1})) \geq 1 - \alpha \qquad \text{(coverage guarantee)}$$

- Coverage guarantee is marginal across examples *and* calibration sets
- α is a user-specified confidence level independent of data distribution and model
- Coverage guarantee can be tight (i.e., with tight upper bound)
- The set size $|C(x_{n+1})|$ is called inefficiency – we want to obtain coverage as efficiently as possible

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

A word on the underlying assumption

Basic assumption for conformal prediction is *exchangeability* of calibration and test example(s):

$$p(x_1, \ldots, x_n, x_{n+1}) = p(x_{i_1}, \ldots, x_{i_n}, x_{i_{n+1}})$$

with i_j being any permutation of $(1, \ldots, n+1)$.

Remarks:

- Assumption of i.i.d. data implies exchangeability but not vice-versa
- Different pixels of the same images are not exchangeable, time series examples are not exchangeable, etc.
- Distribution shift clearly results in non-exchangeability

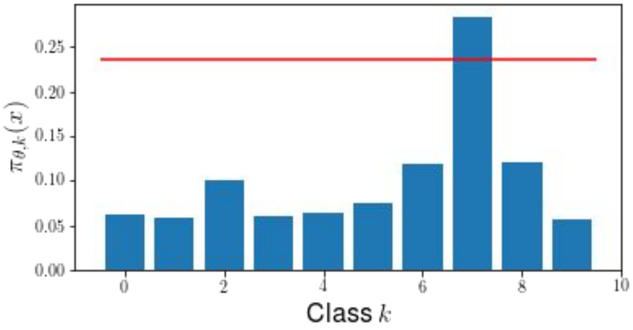
This sounds limiting – but:

- Still weaker than i.i.d.
- Coverage guarantee independent of model *and* data distribution
  (i.e., distribution-free guarantee and no risk of model mis-specification as in Bayesian approaches)

A simple conformal predictor: *thresholding*

The "thresholding" view on conformal prediction:

1. Define confidence sets $C(x_{n+1}) = \{k : \pi_k(x_{n+1}) \geq \tau\}$



Colab session

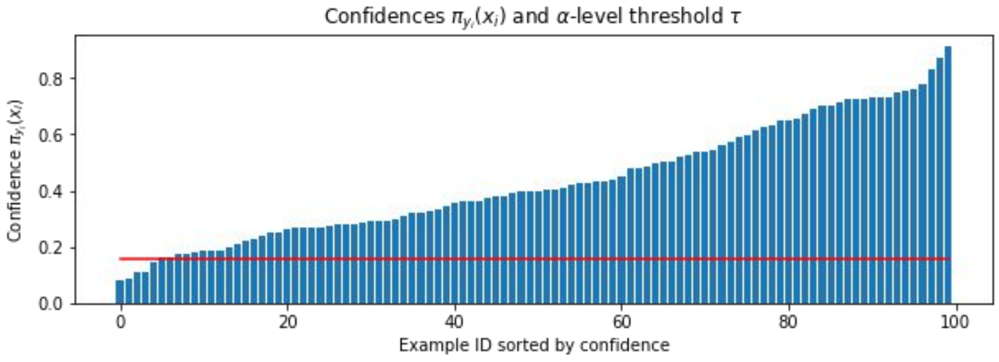
(constructing confidence sets 1-4)

A simple conformal predictor: *thresholding*

The "thresholding" view on conformal prediction:

1. Define confidence sets $C(x_{n+1}) = \{k : \pi_k(x_{n+1}) \geq \tau\}$
2. Calibrate threshold τ on calibration examples:

$$\tau = \alpha\left(1 + \frac{1}{n}\right)\text{-quantile of } \{\pi_{y_i}(x_i)\}_{i=1,\ldots,n}$$



Colab session

(calibrating confidence sets 5-6)

A simple conformal predictor: *thresholding*

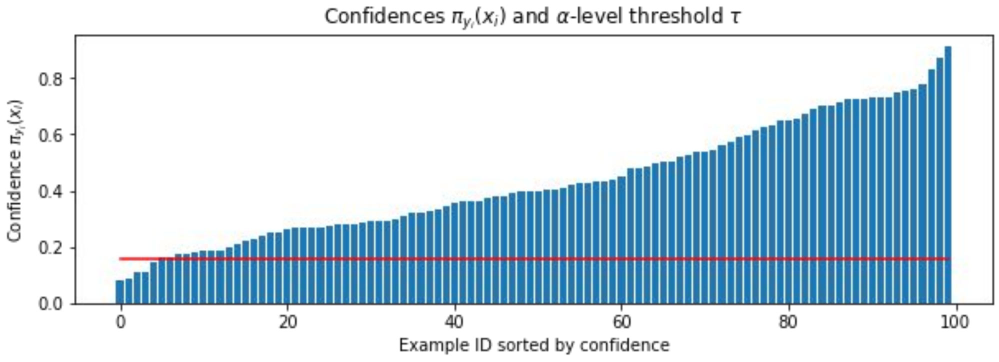
The "thresholding" view on conformal prediction:

1. Define confidence sets $C(x_{n+1}) = \{k : \pi_k(x_{n+1}) \geq \tau\}$
2. Calibrate threshold τ on calibration examples:

$$\tau = \alpha\left(\frac{1}{n+1}\right)\text{-quantile of } \{\pi_{y_i}(x_i)\}_{i=1,\ldots,n}$$

Conformity scores:

- $E(x, k) = \pi_k(x)$ can be replaced with any arbitrary "score" that is higher the more likely k is to be included in $C(x)$ – other conformal predictors define other scores



Alternative formulation: *p-values*

Let us look at one particular example x_{n+1} and class k ; the quantile can be re-formulated as a p-value:

$$\rho_k = \frac{|\{i = 1, \ldots, n : E(x_i, y_i) \leq E(x_{n+1}, k)\}|}{n+1}$$

This is a valid p-value, i.e., $p(\rho_k \leq \alpha) = \alpha$, which means

$$C(x_{n+1}) = \{k : \rho_k \geq \alpha\}$$

are valid confidence sets.

Colab session

(p-values 7 w/o details)

Alternative formulation: *p-values*

Let us look at one particular example x_{n+1} and class k ;the quantile can be re-formulated as a p-value:

$$\rho_k = \frac{|\{i = 1, \ldots, n : E(x_i, y_i) \leq E(x_{n+1}, k)\}|}{n+1}$$

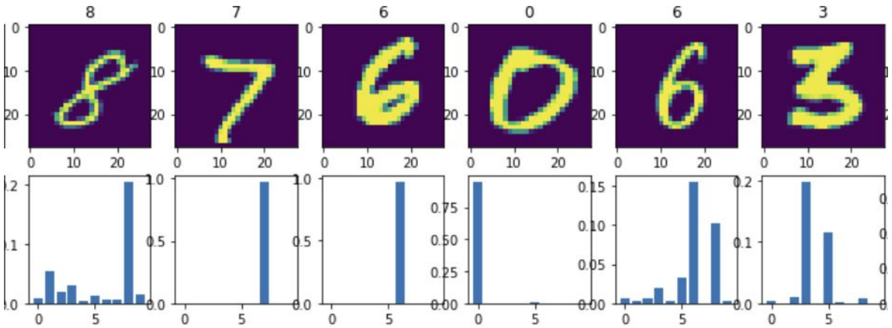
This is a valid p-value, i.e., $p(\rho_k \leq \alpha) = \alpha$, which means

$$C(x_{n+1}) = \{k : \rho_k \geq \alpha\}$$

are valid confidence sets.

Remarks:

- p-values ρ_k are *independent* of the confidence level α
- But calibrating a threshold is usually computationally easier to handle



Colab session

(understanding coverage 8-10)

Understanding the coverage guarantee

Take-aways about coverage guarantee:

- Coverage guarantee is *marginal* across examples
- Guarantee is *in expectation* over calibration sets
- Conditional coverage is not guaranteed by default

→ We can always have “bad luck”

Conclusion:

- “Frequentist” calibration method to predict confidence sets with coverage guarantee
- Independent of problem and model
- Only assumption is exchangeability
- Important to understand the coverage guarantee
  - Marginal – unconditionally and “in expectation”

DeepMind

Advanced topics

Overview of topics

Colab topics:

- Calibration-set conditional coverage
- Class-conditional coverage (a.k.a. *fairness*)

Advanced topics (get in touch!):

- Our work: *learning* conformal prediction
- Our work: handling ambiguous ground truth
- Conformal (multivariate) regression
- Multi-label conformal prediction
- Sample efficient conformal prediction
- Conformal risk
- Distribution shift and robustness
- Private conformal prediction

Calibration set conditional coverage

Calibration-set conditional coverage::

- For split conformal prediction, let e_n the mis-coverage probability with n calibration examples:

$$e_n = p(y_{n+1} \notin C(x_{n+1}))$$

- Then, it can be shown that

$$p(e_n \leq \alpha + \sqrt{\frac{\log 1/\delta}{2n}}) \geq 1 - \delta$$

Colab session

(calibration-set conditional coverage 11)

Calibration set conditional coverage

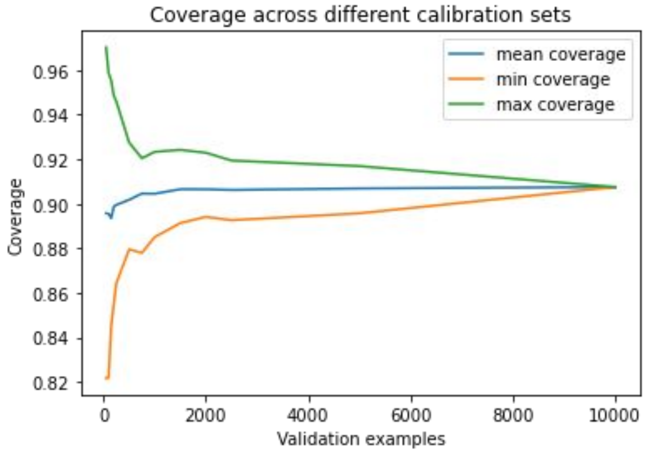
Calibration-set conditional coverage:

- For split conformal prediction, let e_n the mis-coverage probability with n calibration examples:

$$e_n = p(y_{n+1} \notin C(x_{n+1}))$$

- Then, it can be shown that

$$p(e_n \leq \alpha + \sqrt{\frac{\log 1/\delta}{2n}}) \geq 1 - \delta$$



Take-away: with enough samples, we can be pretty sure that we obtain coverage for a fixed calibration set.

- Generally not the case for full, cross-validation or bagging conformal prediction!

Class-conditional coverage

Class-conditional coverage:

- Remember that the coverage guarantee is marginal across examples
- Class-conditional coverage is possible using

Marginal: $$\rho_k = \frac{|\{i = 1, \ldots, n : E(x_i, y_i) \leq E(x_{n+1}, k)\}|}{n+1}$$

Class-conditional: $$\rho_k = \frac{|\{i = 1, \ldots, n \cap y_i = k : E(x_i, y_i) \leq E(x_{n+1}, k)\}|}{|\{i = 1, \ldots, n : y_i = k\}| + 1}$$

Colab session

(class-conditional coverage 12)

Class-conditional coverage

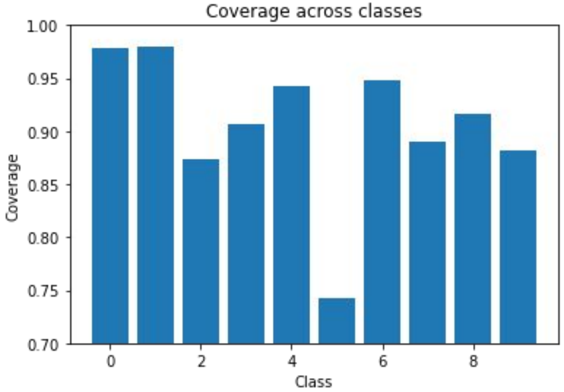
Class-conditional coverage:

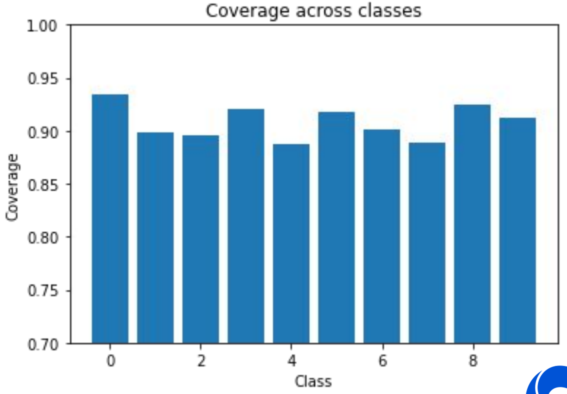
- Remember that the coverage guarantee is marginal across examples
- Class-conditional coverage is possible using

$$\rho_k = \frac{|\{i = 1, \ldots, n \cap y_i = k : E(x_i, y_i) \leq E(x_{n+1}, k)\}|}{|\{i = 1, \ldots, n : y_i = k\}| + 1}$$

- Sacrifices label efficiency for class-conditional coverage
- Attribute-conditional coverage (a.k.a. fairness) possible if attributes known at test time.
- But: coverage conditioned on arbitrary (previously unknown) groups generally impossible.

Take-away: conditional coverage possible assuming knowledge about groups.





Conformal regression

Given $y_i, \pi(x_i) \in \mathbb{R}$ we construct confidence *intervals* in the same way:

$$C(x_{n+1}) := \{r \in \mathbb{R} : E(x_{n+1}, r) \geq \tau\}$$

with the conformity score being defined as

$$E(x, r) = \exp(-|\pi(x) - r|)$$

Key problem: how do we evaluate infinitely many $r \in \mathbb{R}$ in practice?

Conformal regression

Given $y_i, \pi(x_i) \in \mathbb{R}$ we construct confidence *intervals* in the same way:

$$C(x_{n+1}) := \{r \in \mathbb{R} : E(x_{n+1}, r) \geq \tau\}$$

with the conformity score being defined as

$$E(x, r) = \exp(-|\pi(x) - r|)$$

Key problem: how do we evaluate infinitely many $r \in \mathbb{R}$ in practice?

- Can use a one-dimensional grid $\mapsto$ impossible for multivariate regression
- Learn a mean regressor and define

$$C(x_{n+1}) = [\pi(x_{n+1}) - \tau, \pi(x_{n+1}) + \tau]$$

(Results in non-adaptive confidence intervals, which can be fixed using quantile regression)

Take-away: conformal regression is possible, even in high dimensions but additional care needs to be taken!

Multi-label conformal prediction

Given $y_i \subseteq [K]$ and π be a multi-label classifier (e.g., with sigmoids per class).

Can we perform conformal prediction?

- Work on power sets, confidence sets are a sets-of-sets $\mapsto$ requires greedy approaches for large K
- What about just repeating each example for each $k \in y_i$?

Beware of exchangeability: is

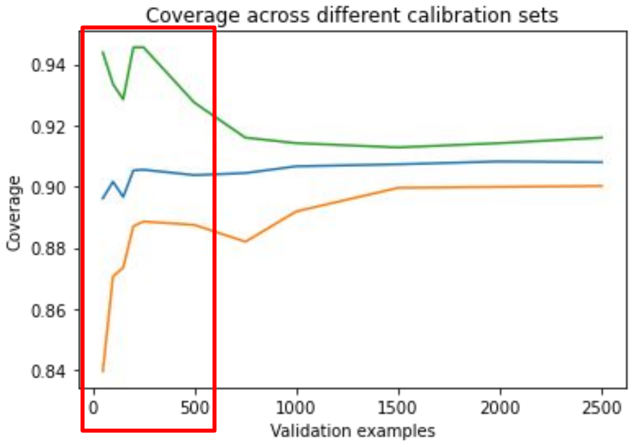
$$p((x_{1,1}, y_{1,1}), (x_{1,2}, y_{1,2}), \ldots)$$

exchangeable?

Take-away: multi-label conformal prediction is possible but involves undesired trade-offs!

Sample-efficient conformal prediction

Split conformal prediction is nice, but with few examples I want to use all of them for training! Also:



Can we do conformal prediction while sharing training and calibration sets?

- "Full" conformal prediction, jackknife+ and cross-validation variants

“Full” conformal prediction

Let $\pi^{(i,k)}$ be the model trained from scratch on

$$(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x_{n+1}, k), (x_{i+1}, y_{i+1}), \ldots, (x_n, y_n)$$

Define

$$\rho_k = \frac{|\{i = 1, \ldots, n : \pi^{(i,k)}_{y_i}(x_i) \leq \pi^{(n+1,k)}_{k}(x_{n+1})\}|}{n+1}$$

- This allows us to use all examples $x_1, \ldots, x_n$ for training *and* calibration
- We need to train $(n+1) \cdot K$ models for each prediction!

Sample-efficient conformal prediction

Let $\pi^{(i,k)}$ be the model trained from scratch on

$$(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x_{n+1}, k), (x_{i+1}, y_{i+1}), \ldots, (x_n, y_n)$$

Define

$$\rho_k = \frac{|\{i = 1, \ldots, n : \pi^{(i,k)}_{y_i}(x_i) \leq \pi^{(n+1,k)}_k(x_{n+1})\}|}{n+1}$$

- This allows us to use all examples $x_1, \ldots, x_n$ for training *and* calibration
- We need to train $(n+1) \cdot K$ models for each prediction!

Alternatives:

- Could we try a cross-validation or bagging approach to avoid re-training these models at test time?
- Yes, but this only provides coverage $1 - 2\alpha$!

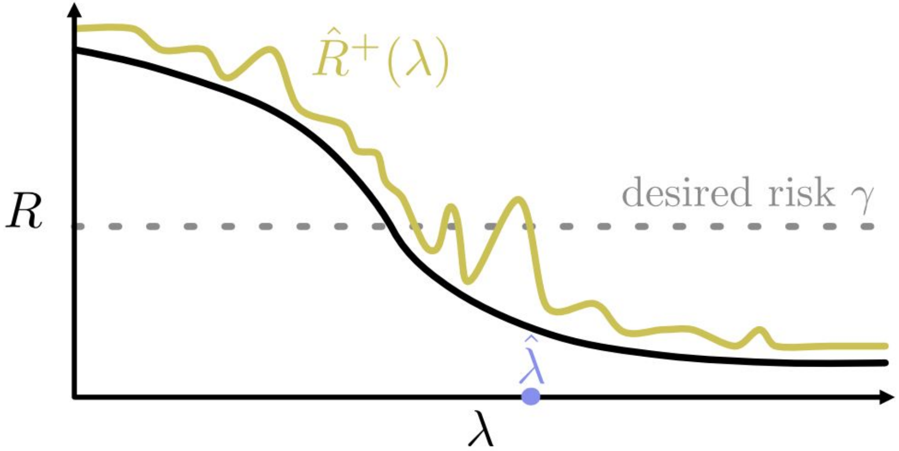
Take-away: more sample-efficient conformal prediction is possible with trade-offs.

- But simply combining p-values is generally not “free”.

Conformal risk

Can we obtain statistical guarantees on other risks (i.e., not coverage)?

- We can define confidence sets C_λ that get smaller for larger λ
- If the risk is monotone and we can upper bound the empirical risk $\hat{R}$ by $\hat{R}^+$, we can calibrate λ
  - Allows guarantees for structured predictions and many other tasks



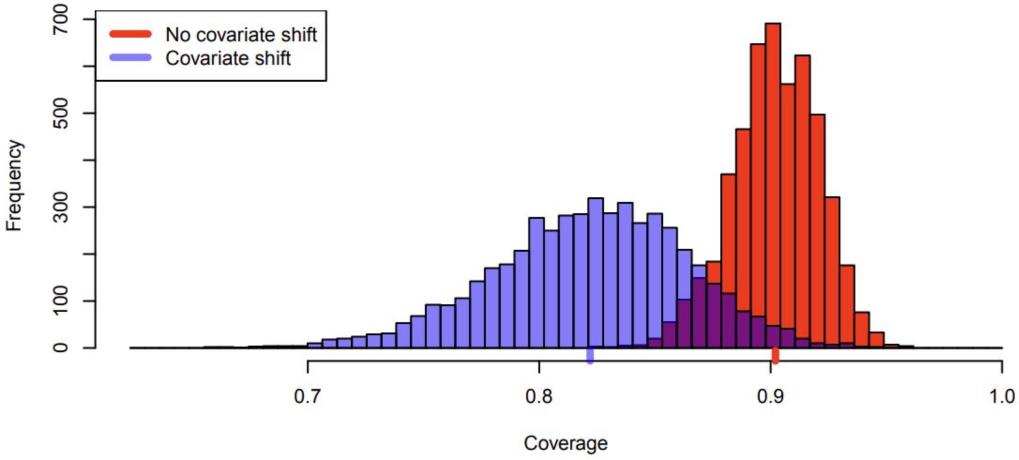
➔ Recently extended to non-monotonic risk functions!

Distribution shift and robustness

Any type of distribution shifts violates the exchangeability assumption! But not all hope is lost:

- (Covariate shift = input distribution shifts, but condition label distributions do not shift)
- If the distribution shift is known, we can work with likelihood ratios and weighted quantiles

$$w_i = \frac{\tilde{p}(x_i)}{p(x_i)}$$

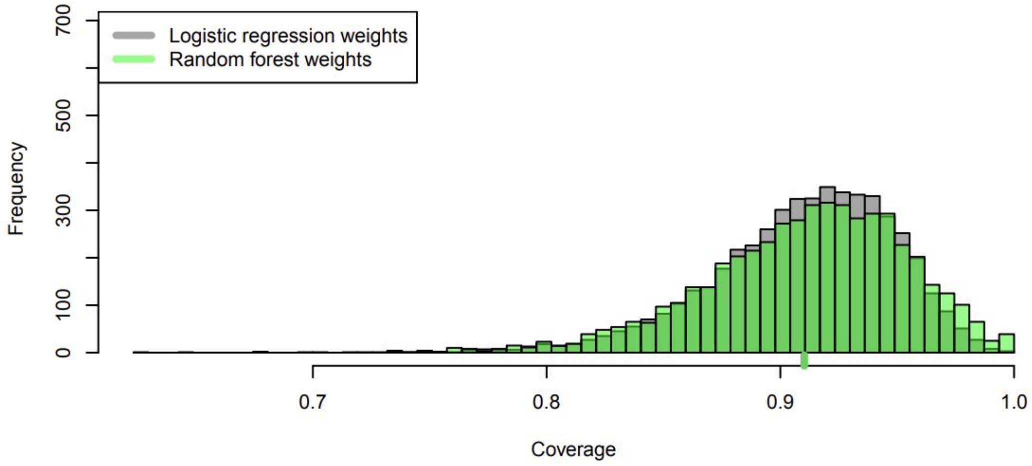


Distribution shift and robustness

Any type of distribution shifts violates the exchangeability assumption! But not all hope is lost:

- (Covariate shift = input distribution shifts, but condition label distributions do not shift)
- If the distribution shift is known, we can work with likelihood ratios and weighted quantiles

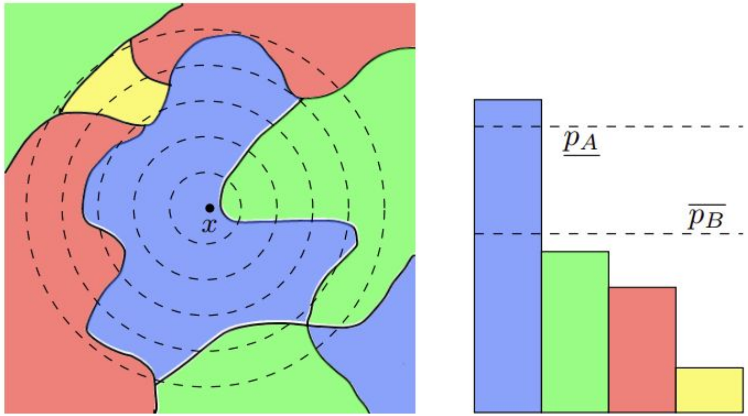
$$w_i = \frac{\tilde{p}(x_i)}{p(x_i)}$$

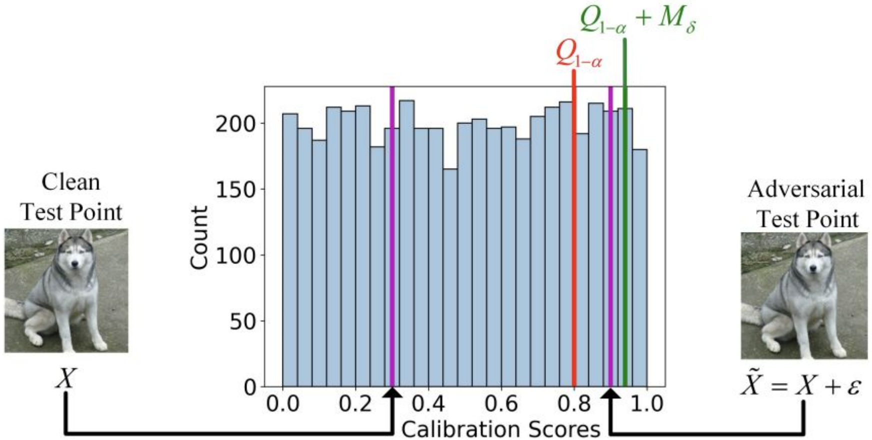


Distribution shift and robustness

Any type of distribution shifts violates the exchangeability assumption! But not all hope is lost:

- If the distribution shift is known and “quantifiable”, we can work with likelihood ratios
- Adversarial examples can be handled using certified defenses (in very limited settings)





Distribution shift and robustness

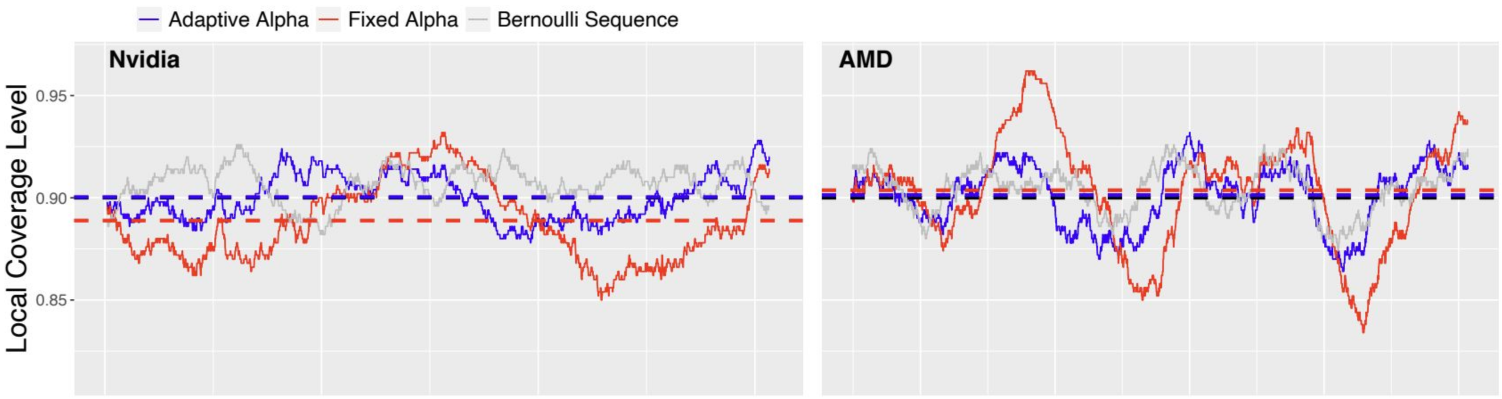
Any type of distribution shifts violates the exchangeability assumption! But not all hope is lost:

- If the distribution shift is known and “quantifiable”, we can work with likelihood ratios
- Adversarial examples can be handled using certified defenses (in very limited settings)
- For out-of-distribution examples, we can give a guarantee on false positives (i.e., on the in-distribution) (Even though many papers claim this was not possible before, most OOD papers do this implicitly by calibrating with respect to the true positive rate)

Distribution shift and robustness

Any type of distribution shifts violates the exchangeability assumption! But not all hope is lost:

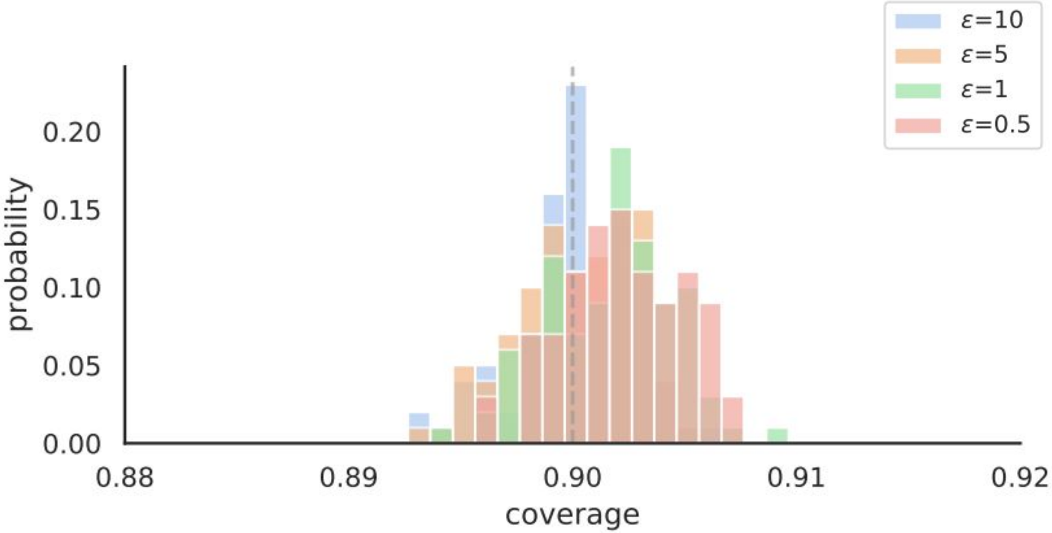
- If the distribution shift is known and “quantifiable”, we can work with likelihood ratios
- Adversarial examples can be handled using certified defenses (in very limited settings)
- For out-of-distribution examples, we can give a guarantee on false positives (i.e., on the in-distribution) (Even though many papers claim this was not possible before, most OOD papers do this implicitly by calibrating with respect to the true positive rate)
- Unknown distribution shifts are most difficult and require “adaptive”/online methods



Private conformal prediction

Problem: can we do conformal calibration privately?
(Note that we do not care about privacy on the *training* set)

- Use a differentially private quantile computation
- Essentially done by discretizing conformity scores into bins
- “Costs” over-estimation of coverage



Learning conformal prediction

Question: can we *learn* how to perform conformal prediction?

Two directions:

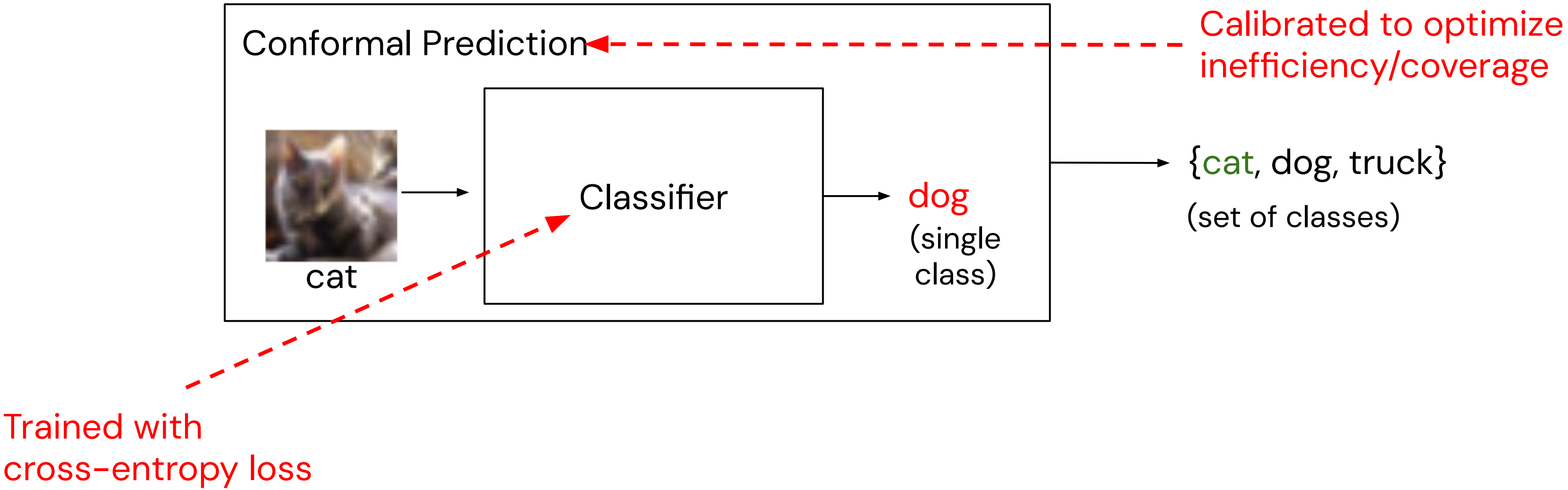
- Learning *models* for/with conformal predictors $\mapsto$ conformal training
    - Independent follow-up work uses conformal training to improve conditional coverage
- Formulating *calibration* as a learning/optimization problem

Our work: conformal training

Question: can we *learn* how to perform conformal prediction?

Two directions:

- Learning *models* for/with conformal predictors ↦ addresses mis-alignment between training/calibration
- Formulating *calibration* as a learning problem

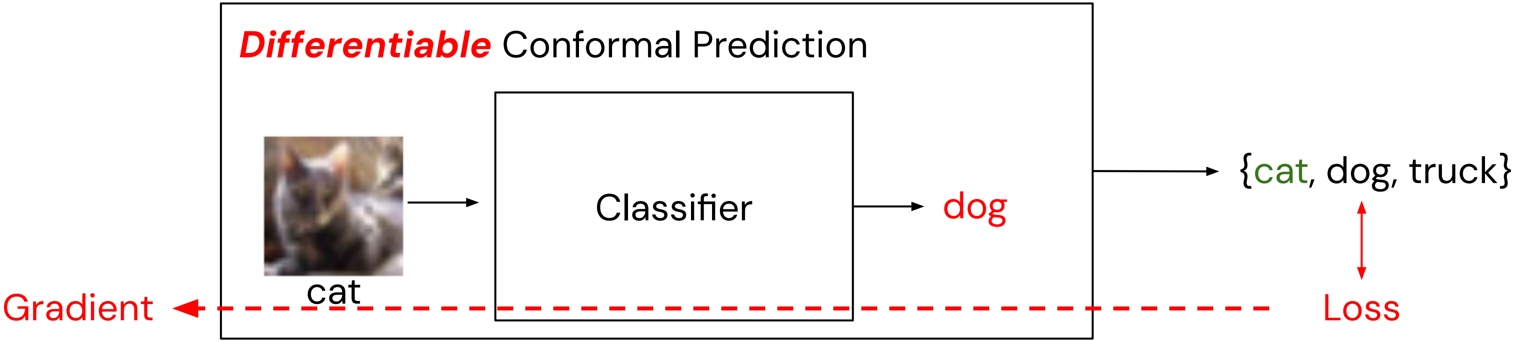


Our work: conformal training

Question: can we *learn* how to perform conformal prediction?

Two directions:

- Learning *models* for/with conformal predictors ↦ addresses mis-alignment between training/calibration
- Formulating *calibration* as a learning problem

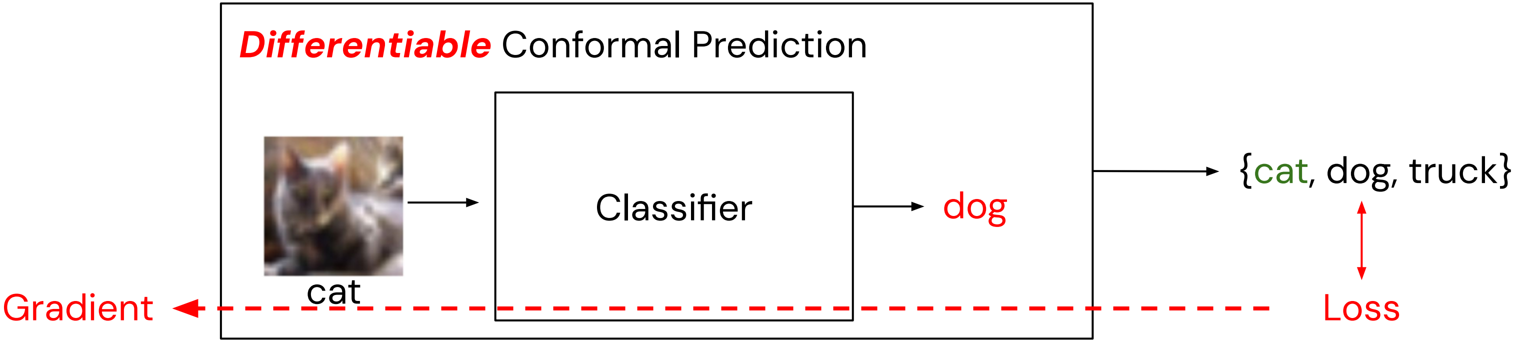


Our work: conformal training

Question: can we *learn* how to perform conformal prediction?

Two directions:

- Learning *models* for/with conformal predictors ↦ addresses mis-alignment between training/calibration
- Formulating *calibration* as a learning problem



➔ Allows to optimize arbitrary losses (minimize inefficiency, improve conditional coverage etc.)
➔ Independent of the coverage guarantee applied at test time

Conclusion

Conformal prediction can be useful for you:

- If you are already calibrating your model, but without obtaining *valid* uncertainty estimates
- If you need uncertainty estimates (confidence sets/intervals, p-values etc.)
- If statistical performance guarantees are required
- If you want to “fix”/calibrate for specific shortcomings (e.g., fairness, robustness)
- If you want to “bridge” performance gaps

Current research tries to:

- Obtain conditional coverage
- Consider more interesting settings (multivariate regression, multi-label classification etc.)
- Obtain guarantees on arbitrary risks
- Go beyond exchangeability (time-series data, distribution shift etc.)
- Integrate conformal prediction into training

Contact: davidstutz.de / dstutz@google.com

Our work: *learning* conformal prediction | handling ambiguous ground truth