Conformal Prediction is Robust to Label Noise

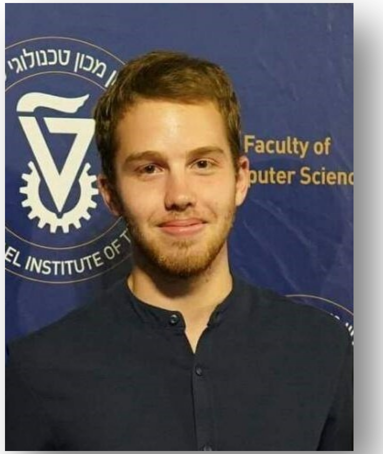
Yaniv Romano

Technion – Israel Institute of Technology

COPA Conference
14 September 2023

Joint work with

Bat-Sheva Einbinder



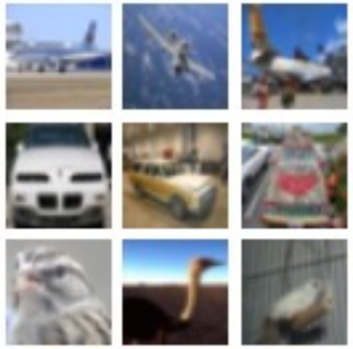
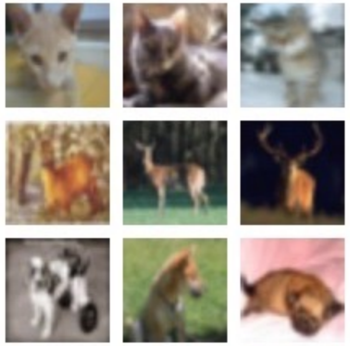
Shai Feldman

Asaf Gendler

Stephen Bates

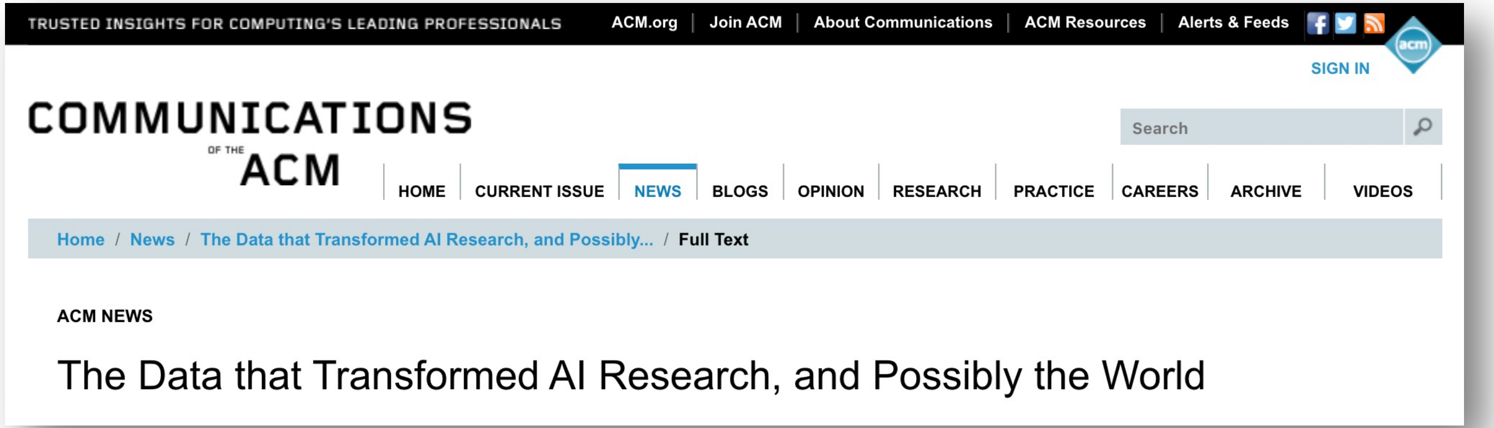
Anastasios Angelopoulos

The fuel of ML is clean, labeled data

MNIST	CIFAR 10	CIFAR 100	Imagenet
			
60K images 10 classes	60K images 10 classes	60K images 100 classes	14M images 22K classes

time

Collecting clean, annotated data is hard and expensive


TRUSTED INSIGHTS FOR COMPUTING'S LEADING PROFESSIONALS | ACM.org | Join ACM | About Communications | ACM Resources | Alerts & Feeds

SIGN IN

COMMUNICATIONS OF THE ACM

Search

HOME | CURRENT ISSUE | NEWS | BLOGS | OPINION | RESEARCH | PRACTICE | CAREERS | ARCHIVE | VIDEOS

Home / News / The Data that Transformed AI Research, and Possibly... / Full Text

ACM NEWS

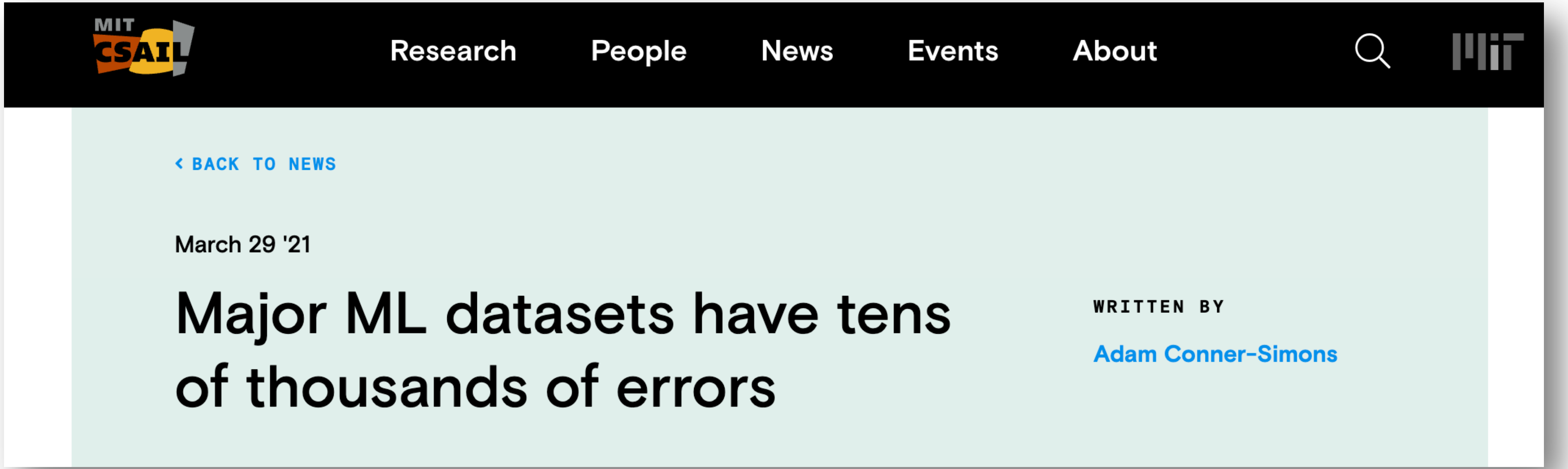
The Data that Transformed AI Research, and Possibly the World

- 14M images, 22K classes
- 49K annotators
- 2 ½ years project
- $$$

amazon mechanical turk beta

Crowdsourcing

No 100% accurate annotations



- Analysis of 10 datasets that have been cited over 100,000 times
- 3.4% of incorrect labels on average
- 6% wrong labels in ImageNet

Various sources of errors [Carniero et al. ‘21]

- Labeling in a rush

1. Dough (ImageNet label)
2. Pizza
3. Soup bowl
4. …

1. Bald eagle
2. Kite (ImageNet label)
3. Soup bowl
4. …

Pay less $$$ and get more

company

VS

Get more $$$
work fast/too much

annotator

Various sources of errors [Carniero et al. ‘21]

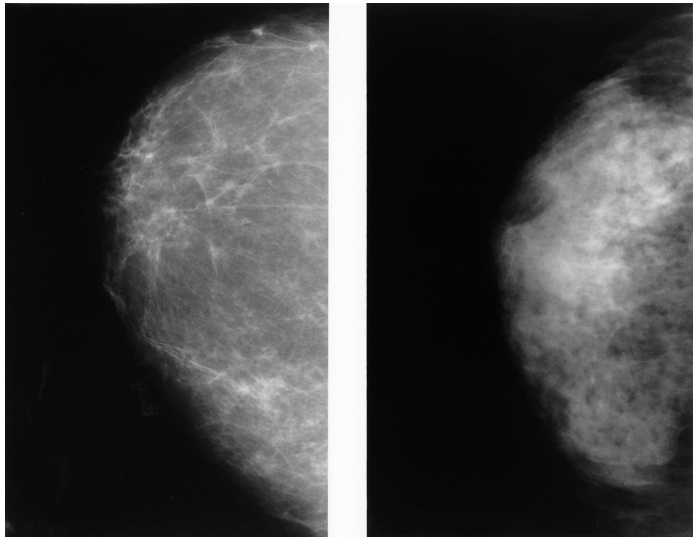
- Labeling in a rush
- Low-quality data, uncertainty

1. Airplane (CIFAR10 label)
2. Ship
3. Car
4. …

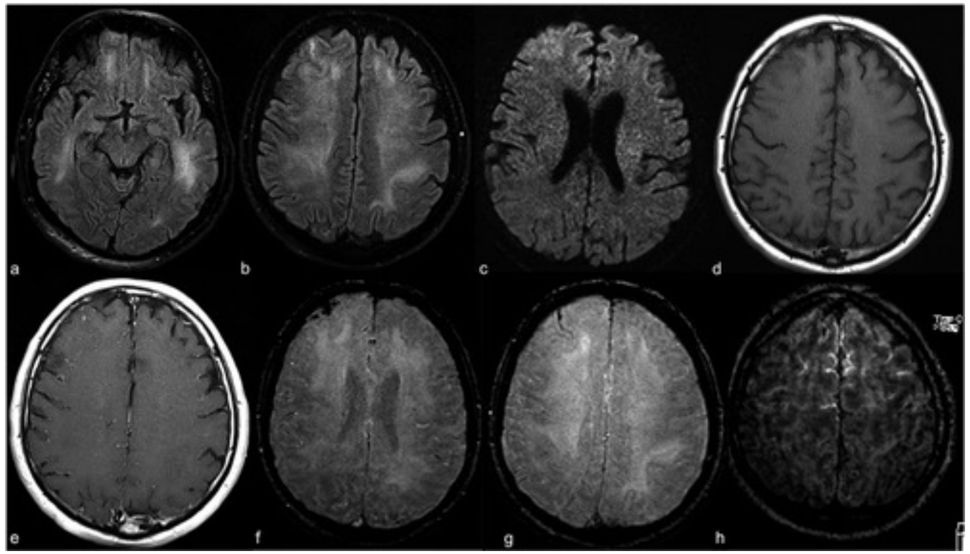
1. Truck
2. Cat (CIFAR10 label)
3. Dog
4. …

Various sources of errors [Carniero et al. '21]

- Labeling in a rush
- Low-quality data, uncertainty
- Challenging problems



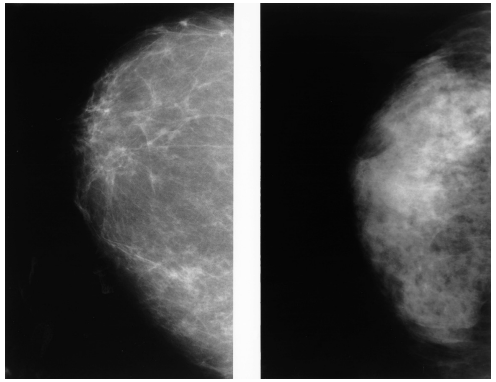
mammography



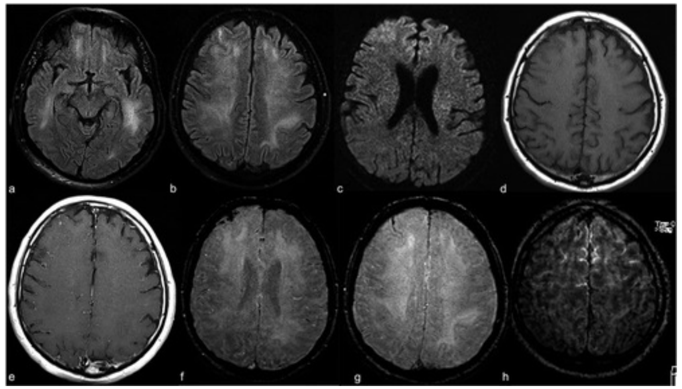
brain MRI

Various sources of errors

- Labeling in a rush
- Low-quality data, uncertainty
- Challenging problems
- Difficult to hire experts



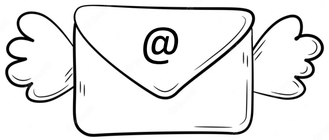
mammography

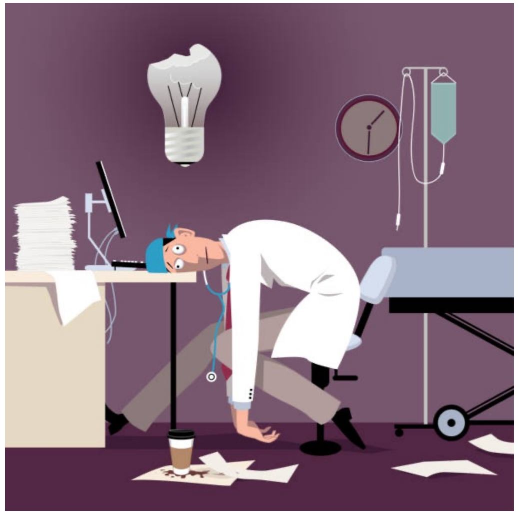


brain MRI

company

"I hope this email finds you well"

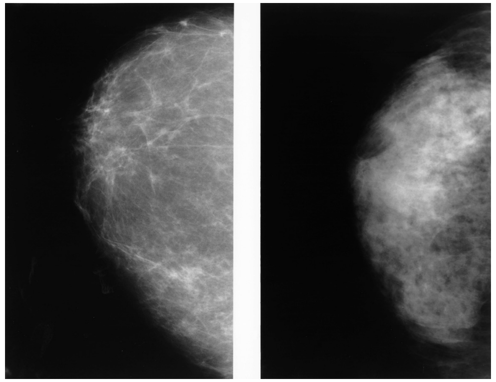




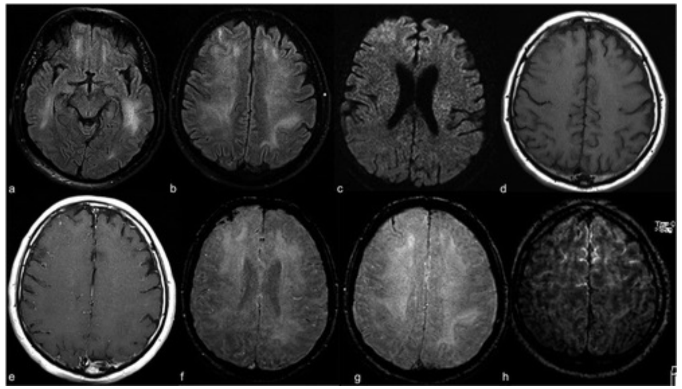
expert

Various sources of errors

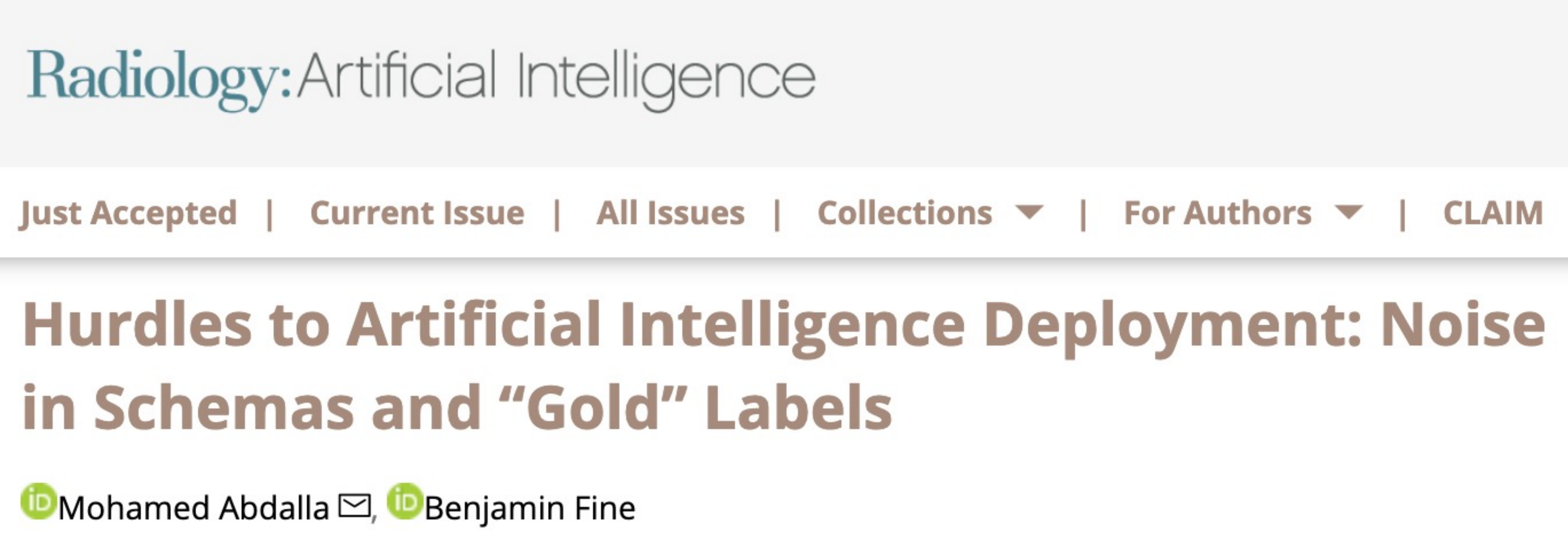
- Labeling in a rush
- Low-quality data, uncertainty
- Challenging problems
- Difficult to hire experts
- Subjective options, there is no consensus



mammography



brain MRI



Radiology: Artificial Intelligence

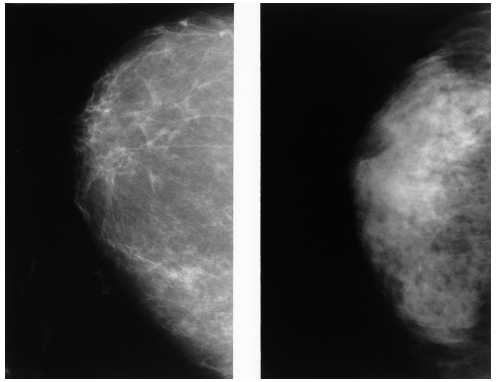
Just Accepted | Current Issue | All Issues | Collections | For Authors | CLAIM

Hurdles to Artificial Intelligence Deployment: Noise in Schemas and "Gold" Labels

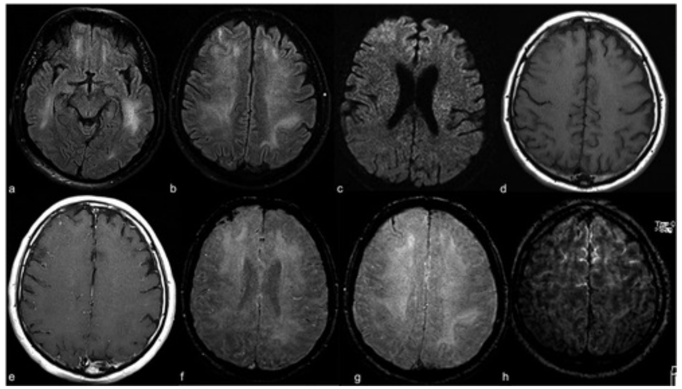
Mohamed Abdalla, Benjamin Fine

Various sources of errors

- Labeling in a rush
- Low-quality data, uncertainty
- Challenging problems
- Difficult to hire experts
- Subjective options, there is no consensus
- Sensor noise
- Data entry mistakes
- …



mammography



brain MRI

No 100% accurate labels

→ noisy labels

Uncertainty is inevitable!

Ultimate goal: reliable UQ under label noise

- **Input:** n noisy training points $(X_1, \tilde{Y}_1), \ldots, (X_n, \tilde{Y}_n)$ and a test point $(X_{\text{test}}, ?)$
  → exchangeable (e.g., i.i.d.) samples from unknown joint dist. $P_{X\tilde{Y}}^{\text{noisy}}$
- $X \in \mathcal{X}$: features
- $\tilde{Y} \in \mathcal{Y}$: noisy label/response
- $Y \in \mathcal{Y}$: ground-truth, clean label (*unobserved*)

Ultimate goal: reliable UQ under label noise

- **Input:** n noisy training points $(X_1, \tilde{Y}_1), \ldots, (X_n, \tilde{Y}_n)$ and a test point $(X_{\text{test}}, ?)$
  → exchangeable (e.g., i.i.d.) samples from unknown joint dist. $P_{X\tilde{Y}}^{\text{noisy}}$
- $X \in \mathcal{X}$: features
- $\tilde{Y} \in \mathcal{Y}$: noisy label/response
- $Y \in \mathcal{Y}$: ground-truth, clean label (*unobserved*)

Wish to use any ML algorithm to construct a marginal **distribution-free prediction set**

$$\mathbb{P}\left[Y_{\text{test}} \in C^{\text{noisy}}(X_{\text{test}})\right] \geq 1 - \alpha \quad \text{(e.g., 90\%)}$$

$\alpha \in (0,1)$ is a user-specified miscoverage rate

- Construct $C^{\text{noisy}}(X_{\text{test}})$ using the *observed* noisy data
- Guarentee that clean Y_{test} is covered in $C^{\text{noisy}}(X_{\text{test}})$

Ultimate goal: reliable UQ under label noise

- **Input:** n noisy training points $(X_1, \tilde{Y}_1), \dots, (X_n, \tilde{Y}_n)$ and a test point $(X_{\text{test}}, ?)$

  → exchangeable (e.g., i.i.d.) samples from unknown joint dist. $P_{X\tilde{Y}}^{\text{noisy}}$
- $X \in \mathcal{X}$: features
- $\tilde{Y} \in \mathcal{Y}$: noisy label/response
- $Y \in \mathcal{Y}$: ground-truth, clean label (*unobserved*)

Wish to use any ML algorithm to construct a marginal **distribution-free prediction set**

$$\mathbb{P}\left[Y_{\text{test}} \in C^{\text{noisy}}(X_{\text{test}})\right] \geq 1 - \alpha \quad \text{(e.g., 90\%)}$$

$\alpha \in (0,1)$ is a user-specified miscoverage rate

- Construct $C^{\text{noisy}}(X_{\text{test}})$ using the *observed* noisy data
- Guarentee that clean Y_{test} is covered in $C^{\text{noisy}}(X_{\text{test}})$

how and under what conditions is it possible?

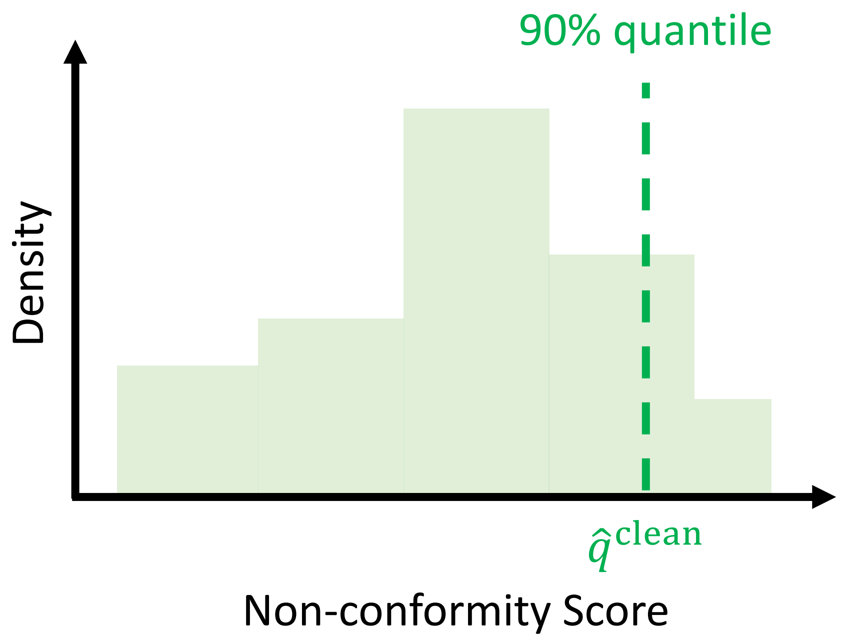
Conformal prediction: notations

Conformal prediction [Vovk et al. ’99; Papadopoulos et al. ‘12, Lei et al. ’18; …]

- **Input:** pre-trained predictive model $\hat{f}$, and holdout calibration set $\{(X_i, Y_i)\}_{i=1}^{n}$
- **Process**
  – Compute non-conformity scores $s_i = s(X_i, Y_i)$ for all $i = 1, \ldots, n$
  a measure of goodness-of-fit (the lower the better), e.g., $s_i = \left|\hat{f}(X_i) - Y_i\right|$

Conformal prediction [Vovk et al. '99; Papadopoulos et al. '12, Lei et al. '18; ...]

- **Input:** pre-trained predictive model $\hat{f}$, and holdout calibration set $\{(X_i, Y_i)\}_{i=1}^n$
- **Process**
  - Compute non-conformity scores $s_i = s(X_i, Y_i)$ for all $i = 1, \ldots, n$
    a measure of goodness-of-fit (the lower the better), e.g., $s_i = |\hat{f}(X_i) - Y_i|$
  - Compute* $\hat{q}^{\text{clean}}$ = the $(1-\alpha)$-empirical quantile of $\{s_i\}_{i=1}^n$



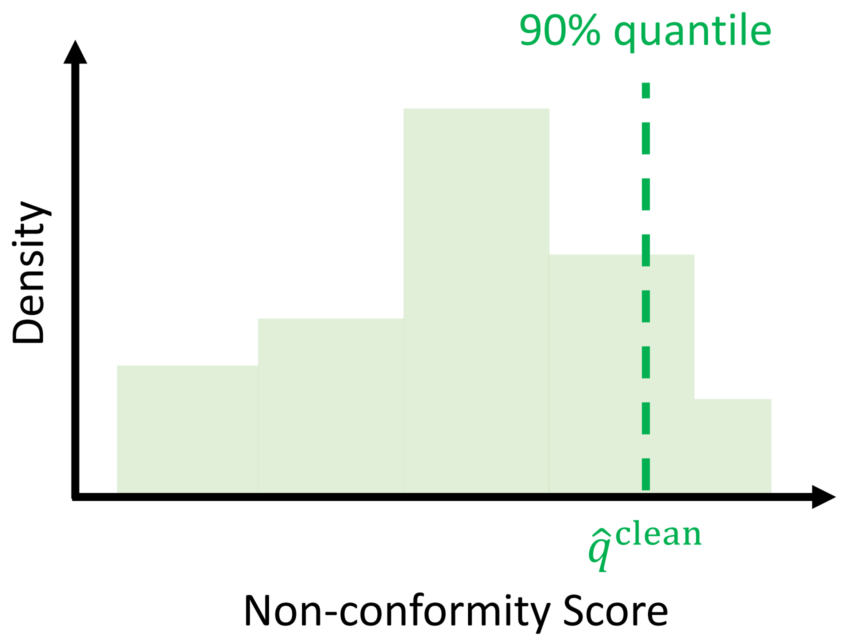
*missing a small correction term

Conformal prediction [Vovk et al. ’99; Papadopoulos et al. ‘12, Lei et al. ’18; …]

- **Input:** pre-trained predictive model $\hat{f}$, and holdout calibration set $\{(X_i, Y_i)\}_{i=1}^n$
- **Process**
  - Compute non-conformity scores $s_i = s(X_i, Y_i)$ for all $i = 1, \ldots, n$
    a measure of goodness-of-fit (the lower the better), e.g., $s_i = \left|\hat{f}(X_i) - Y_i\right|$
  - Compute* $\hat{q}^{\text{clean}}$ = the $(1-\alpha)$-empirical quantile of $\{s_i\}_{i=1}^n$
- **Output:** prediction set with 90% coverage

$$C\left(X_{\text{test}}, \hat{q}^{\text{clean}}\right) = \left\{y \in \mathcal{Y} : s(X_{\text{test}}, y) \leq \hat{q}^{\text{clean}}\right\}$$

Sweep over all $y \in \mathcal{Y}$ and return the guessed y‘s whose score falls below $\hat{q}^{\text{clean}}$



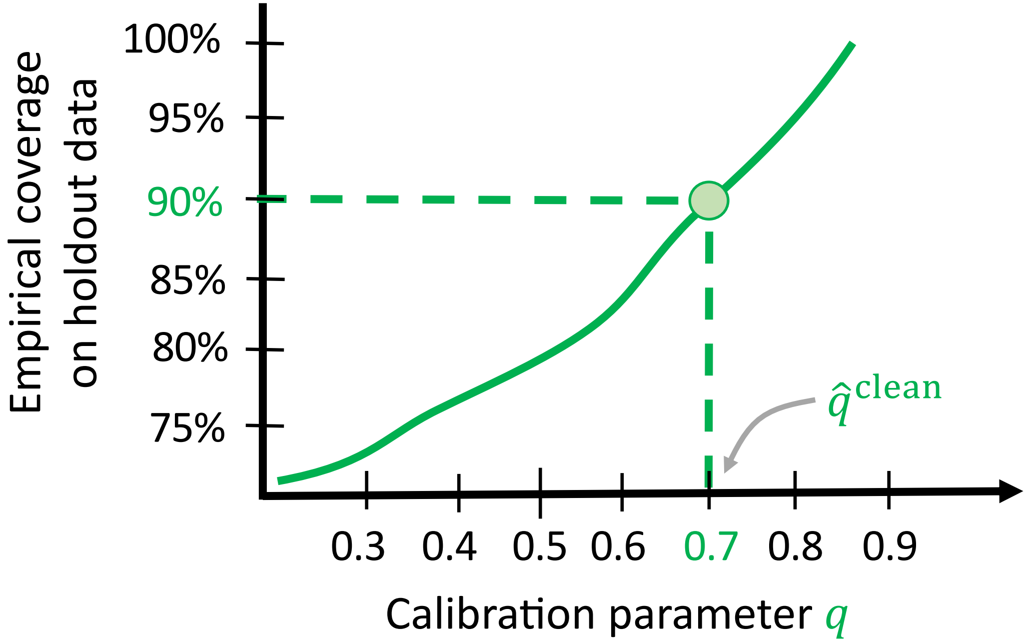
*missing a small correction term

Another way to view conformal prediction

- Given a set constructing function

$$C(x, q) = \{y \in \mathcal{Y} : s(x, y) \leq q\}$$

- Find the $\hat{q}^{\text{clean}}$ that achieves 90% coverage on the calibration set



$$\text{Emp. Coverage*}(q) = \frac{1}{n} \sum_{i=1}^{n} 1\{Y_i \in C(X_i, q)\}_{i=1}^{n}$$

*missing a small correction term

Conformal prediction is valid under exchangeability

<u>Theorem</u> (Vovk et al. '99; Papadopoulos et al. '12; Lei et al. '18; R., Patterson, Candes '19, …)

If $(X_1, Y_1), \dots, (X_n, Y_n)$ and $(X_{\text{test}}, Y_{\text{test}})$ are exchangeable (or i.i.d.). Then,

$$\mathbb{P}\left[Y_{\text{test}} \in C^{\text{clean}}\left(X_{\text{test}}, \hat{q}^{\text{clean}}\right)\right] \geq 1 - \alpha \quad \text{(e.g., 90\%)}$$

- Finite sample, dist. free guarantee!
- There is also an upper bound (guarantee is tight)
- Exchangeability is the only assumption

Conformal in action: the Washington Post election night model

Technology is based on *conformalized quantile regression* [R., Patterson, Candes ’19]


The Washington Post
Democracy Dies in Darkness

PostCode

From The Washington Post Engineering team.

Open positions

Don’t see the job you’re looking for?

How The Washington Post Estimates Outstanding Votes for the 2020 Presidential Election

By **Lenny Bronner**, **Jeremy Bowers** and **John Cherian**

Oct. 22, 2020 at 6:14 p.m GMT+3

Pennsylvania

20 ELECTORAL VOTES

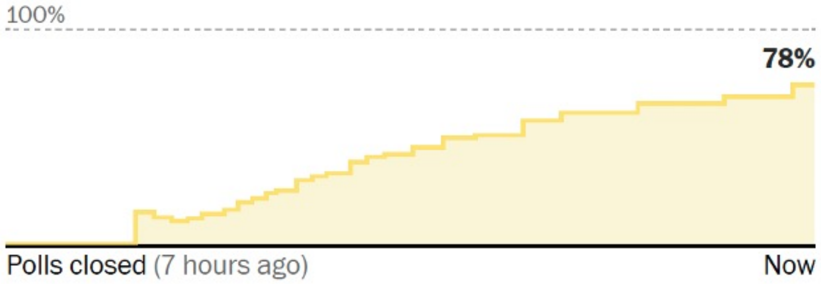
LIVE: Donald Trump (R) is leading. An estimated 78 percent of votes have been counted.

Biden
43.0%
2,283,656

Trump
55.7%
2,956,791

How much of the vote has been counted in Pennsylvania?

The Post estimates **78%** of votes cast have been counted here.



U.S. House District 10

Perry	58.9%
DePasquale	41.1%

An estimated 67% of votes have been counted

U.S. House District 17

Parnell	56.5%
Lamb	43.5%

An estimated 67% of votes have been counted

Pennsylvania has 18 U.S. House races. Jump to results

Note: Map colors on this page won't indicate a lead for a candidate until an estimated 35 percent of the vote has been reported there. Results updated at 2:50 a.m. ET

The Washington Post 4 November 2020, 11:50 PM

Pennsylvania

20 ELECTORAL VOTES

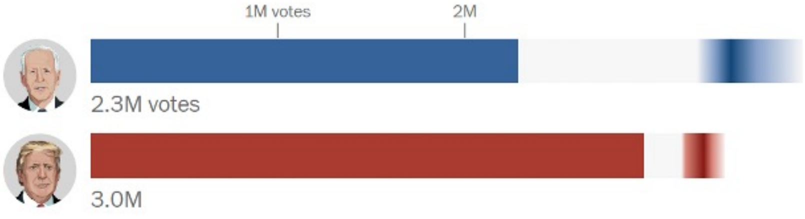
LIVE: Donald Trump (R) is leading. An estimated 78 percent of votes have been counted.

Where the vote could end up

These estimates are calculated based on past election returns as well as votes counted in the presidential race so far. View details

We estimate that 78 percent of the total votes cast have been counted. Biden is favored to win the state, but Trump still has a chance to win. These are the most likely outcomes.

Counted votes   Estimates of final vote tally
Lighter colors are less likely outcomes



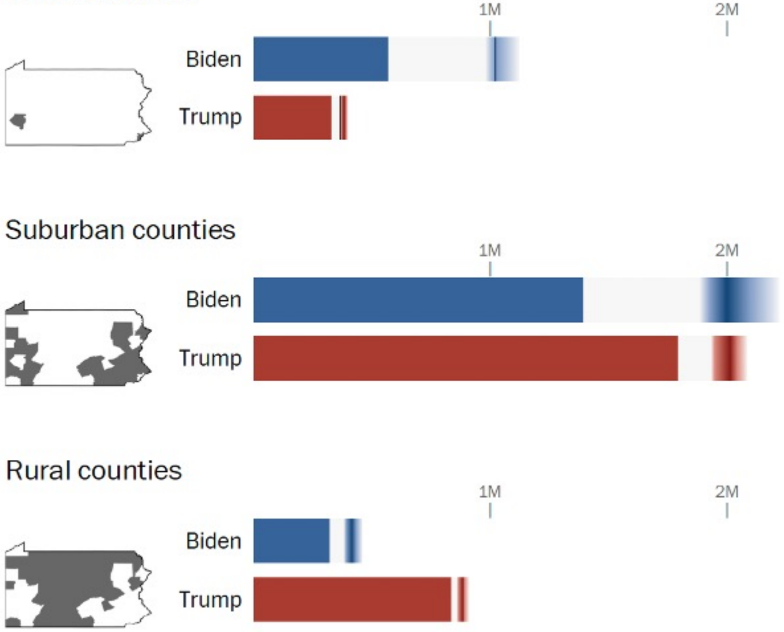
Core idea: use reported counties to forecast unreported counties

Breaking down the estimates

Urban counties

Suburban counties

Rural counties



The Washington Post   4 November 2020, 11:50 PM

Pennsylvania

20 ELECTORAL VOTES

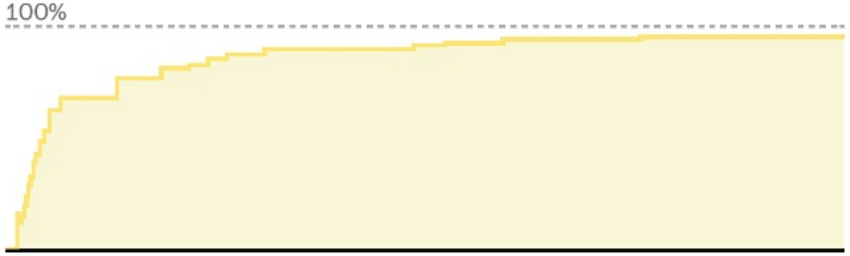
LIVE: Joe Biden (D) is leading by 30,908 votes. An estimated 95 percent of votes have been counted.

Biden
49.6%
3,339,318

Trump
49.1%
3,308,410

How much of the vote has been counted in Pennsylvania?

The Post estimates **95%** of votes cast have been counted here.



U.S. House District 10

Perry ✓ 53.4%
DePasquale 46.6%

An estimated 94% of votes have been counted

U.S. House District 17

Lamb 50.9%
Parnell 49.1%

An estimated 94% of votes have been counted

Pennsylvania has 18 U.S. House races. Jump to results

Note: Map colors on this page won't indicate a lead for a candidate until an estimated 35 percent of the vote has been reported there. Results updated at 11:24 a.m. ET

Amazing software packages

Unit tests passing | codecov 100% | docs passing | license BSD-3-Clause | python 3.7 | 3.8 | 3.9 | 3.10

pypi v0.7.0 | conda-forge v0.7.0 | release v0.7.0 | commits since v0.7.0 2 | 10.48550/arXiv.2207.12274

MAPIE - Model Agnostic Prediction Interval Estimator

Conformal in the cloud

aws

re:Invent Products Solutions Pricing Documentation Learn Partner Network AWS Marketplace Customer Enablement Events

Blog Home **Category** **Edition**

AWS Machine Learning Blog

Introducing Fortuna: A library for uncertainty quantification

by Gianluca Detommaso, Alberto Gasparin, Cedric Archambeau, Michele Donini, Matthias Seeger, and Andrew Gordon Wilson | on 16 DEC 2022 | in Amazon Machine Learning, Artificial Intelligence, Foundational (100) | Permalink | Comments | Share

https://github.com/awslabs/fortuna

Fortuna's documentation   Quickstart   Installation   Examples   Usage modes   API References   More

Conformal prediction methods

We support conformal prediction methods for classification and regression.

For classification:

- **A simple conformal prediction sets method** [Vovk et al., 2005]

  A simple conformal prediction method deriving a score function from the probability associated to the largest class.

- **An adaptive conformal prediction sets method** [Romano et al., 2020]

  A method for conformal prediction deriving a score function that makes use of the full vector of class probabilities.

- **Adaptive conformal inference** [Gibbs et al., 2021]

  A method for conformal prediction that aims at correcting the coverage of conformal prediction methods in a sequential prediction framework (e.g. time series forecasting) when the distribution of the data shifts over time.

For regression:

- **Conformalized quantile regression** [Romano et al., 2019]

  A conformal prediction method that takes in input a coverage interval and calibrates it.

- **Conformal interval from scalar uncertainty measure** [Angelopoulos et al., 2022]

  A conformal prediction method that takes in input a scalar measure of uncertainty (e.g. the standard deviation) and returns a conformal interval.

UQ methods we developed for image recovery tasks: Technion-Berkeley collaboration

Back to label noise…

Back to Label noise: what is the challenge?

Suppose we observe *only* the noisy labels

$\tilde{Y} = g(Y, U)$ e.g., randomly flip the true label w.p. ϵ

- g is a corruption function; U is random noise

Back to Label noise: what is the challenge?

Suppose we observe *only* the noisy labels

$$\tilde{Y} = g(Y, U) \quad \text{e.g., randomly flip the true label w.p. } \epsilon$$

- g is a corruption function; U is random noise

Imagine we run conformal prediction on noisy data **as if it is clean**

$$C(x, \hat{q}^{\text{noisy}}) = \{y \in \mathcal{Y} : s(X_{\text{test}}, y) \leq \hat{q}^{\text{noisy}}\}$$

$\hat{q}^{\text{noisy}}$ = $(1 - \alpha)$-empirical quantile of $\{s(X_i, \tilde{Y}_i)\}_{i=1}^{n}$

Back to Label noise: what is the challenge?

Suppose we observe *only* the noisy labels

$$\tilde{Y} = g(Y, U) \quad \text{e.g., randomly flip the true label w.p. } \epsilon$$

- g is a corruption function; U is random noise

Imagine we run conformal prediction on noisy data **as if it is clean**

$$C\left(x, \hat{q}^{\text{noisy}}\right) = \left\{y \in \mathcal{Y} : s(X_{\text{test}}, y) \leq \hat{q}^{\text{noisy}}\right\}$$

$\hat{q}^{\text{noisy}}$ = $(1-\alpha)$-empirical quantile of $\left\{s\left(X_i, \tilde{Y}_i\right)\right\}_{i=1}^{n}$

- It achieves valid cov. on noisy $\mathbb{P}\left(\tilde{Y}_{\text{test}} \in C\left(X_{\text{test}}, \hat{q}^{\text{noisy}}\right)\right) \geq 1-\alpha$
- Would it have valid cov. on clean? $\mathbb{P}\left(Y_{\text{test}} \in C\left(X_{\text{test}}, \hat{q}^{\text{noisy}}\right)\right) \geq 1-\alpha$

Problem:
distribution shift!
$P_{X,Y}^{\text{clean}} \neq P_{X,\tilde{Y}}^{\text{noisy}}$

Adversarial thinking about distribution shift

$\Downarrow$

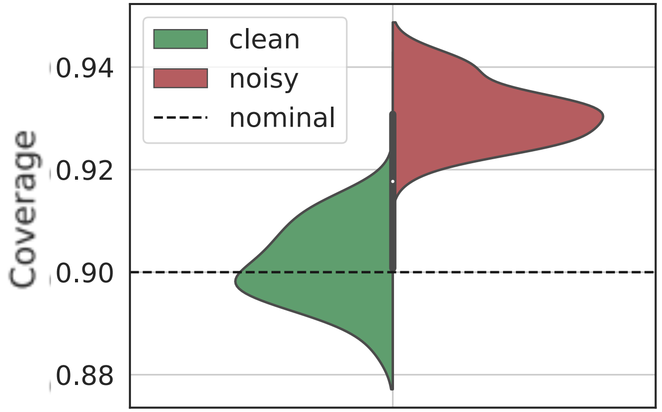
under-coverage

But is it really the case?

Let’s see some evidence on label noise robustness

Classification: CIFAR10H image data

- Task: classify the object in an image ($K = 10$ classes)
- Clean CIFAR10 : clean labels Y are the majority vote of ≈ 50 annotators
- Noisy CIFAR10H : noisy labels $\tilde{Y}$ are from a single annotator
- NNet classifier (resnet-18)



- True label: Cat
- Noisy: {Cat, Dog}
- Clean: {Cat}

- True label: Car
- Noisy: {Car, Ship, Cat}
- Clean: {Car}

- Exact coverage when calibrated on clean labels (not surprising)
- Conservative but valid coverage when calibrated on noisy labels

Regression: aesthetic visual analysis

- Data: pairs of images and their annotated aesthetic score, in a range of 1-10

[Murray et al. '12]

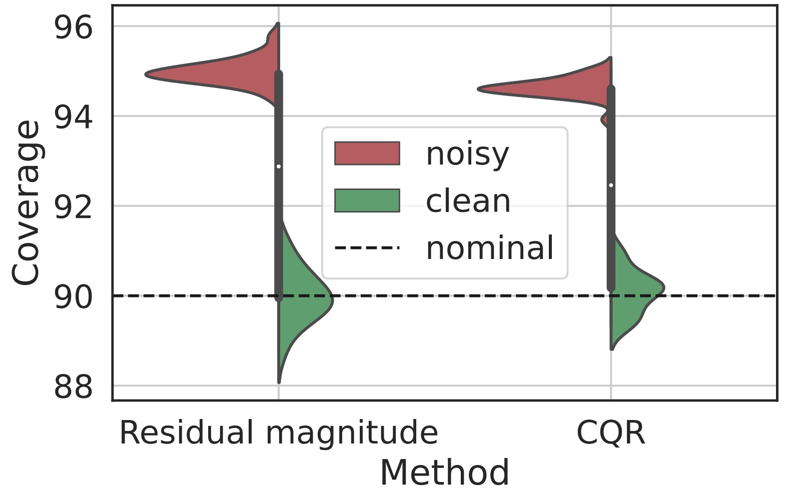
Ranked as "high-quality"
Aesthetic score = 9

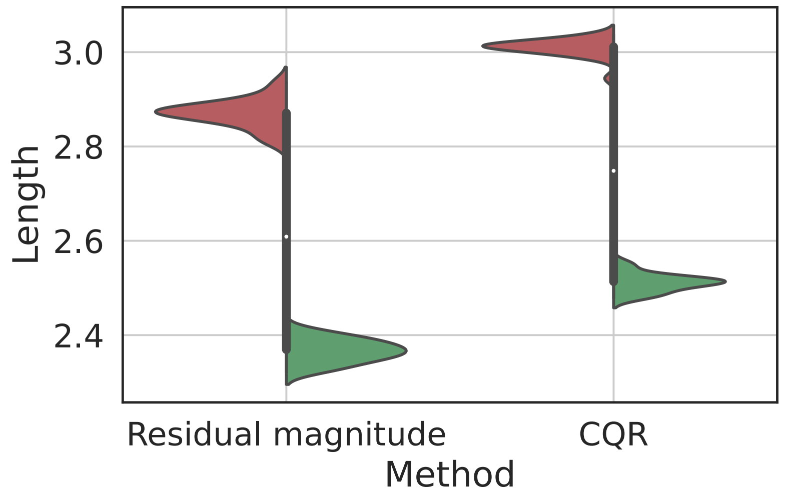
Ranked as "low-quality"
Aesthetic score = 2

Subjective options, uncertainty, ...

Regression: aesthetic visual analysis

- Data: pairs of images and their annotated aesthetic score, in a range of 1-10
- Task: predict the aesthetic score of a given image
  - Clean Y = average score of $\approx$ 200 annotators
  - Noisy $\tilde{Y}$ = average score of $\approx$ 10 annotators
- NNet regressor (fine-tuned VGG-16 model)
- Training ($\approx$ 35K images), calib. ($\approx$ 8K), testing ($\approx$ 8K)





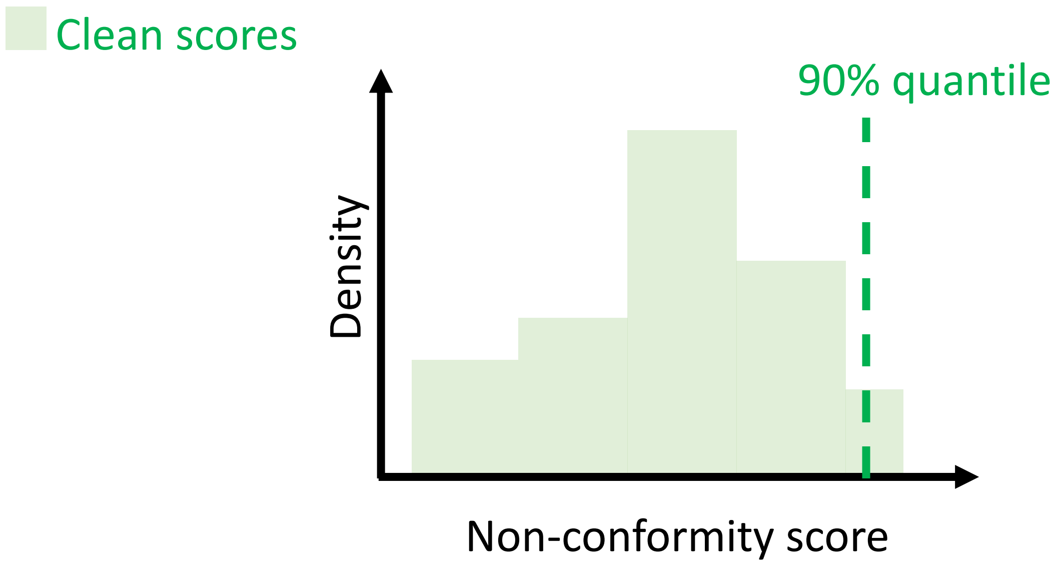
- Exact coverage when calibrated on clean
- Conservative coverage when calibrated on noisy
- Noisy intervals are **wider**

Empirical evidence: label noise $\Longrightarrow$ over-coverage

Let’s gain intuition: when and why this happens?

Contrastive vs. dispersive noise

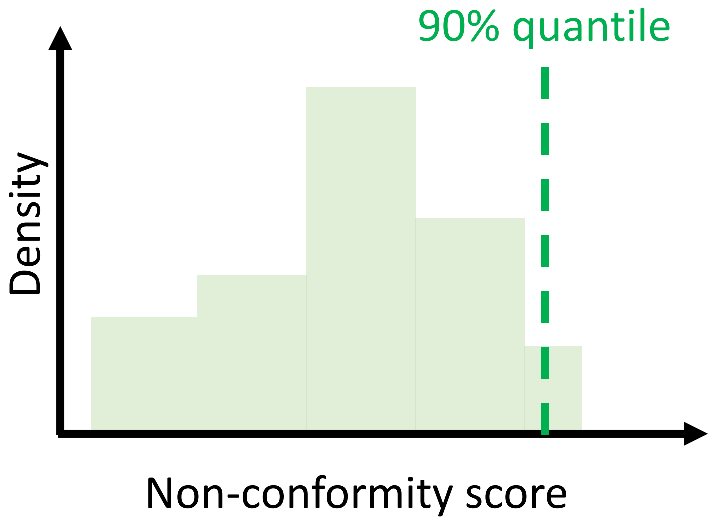
contractive noise



scores on clean > scores on noisy

e.g. $\mathrm{Var}(Y \mid X = x) > \mathrm{Var}(\tilde{Y} \mid X = x)$

dispersive noise

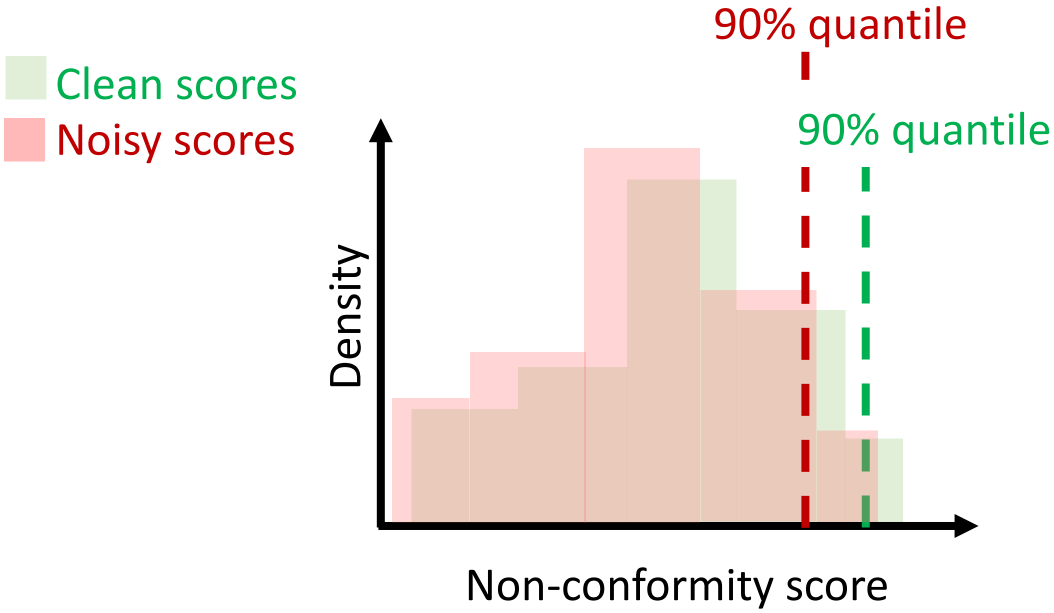


scores on clean < scores on noisy

e.g. $\mathrm{Var}(Y \mid X = x) < \mathrm{Var}(\tilde{Y} \mid X = x)$

Contractive vs. dispersive noise

contractive noise

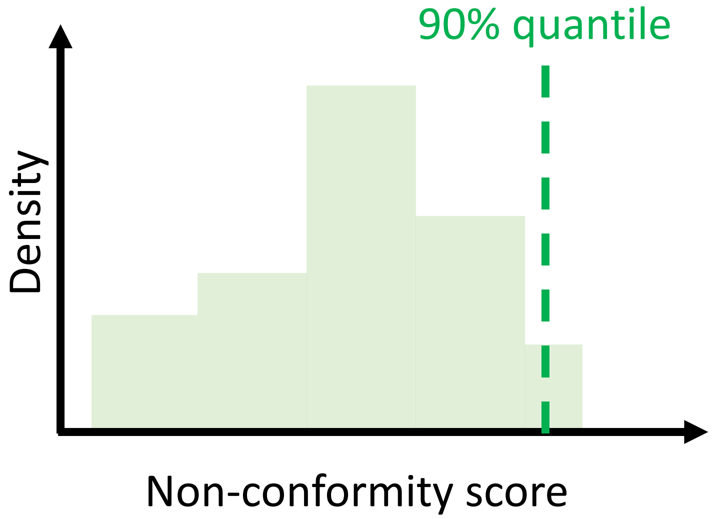


scores on clean > scores on noisy

e.g. $\mathrm{Var}(Y \mid X = x) > \mathrm{Var}(\tilde{Y} \mid X = x)$

<u>Effect</u>: **under**-coverage

dispersive noise

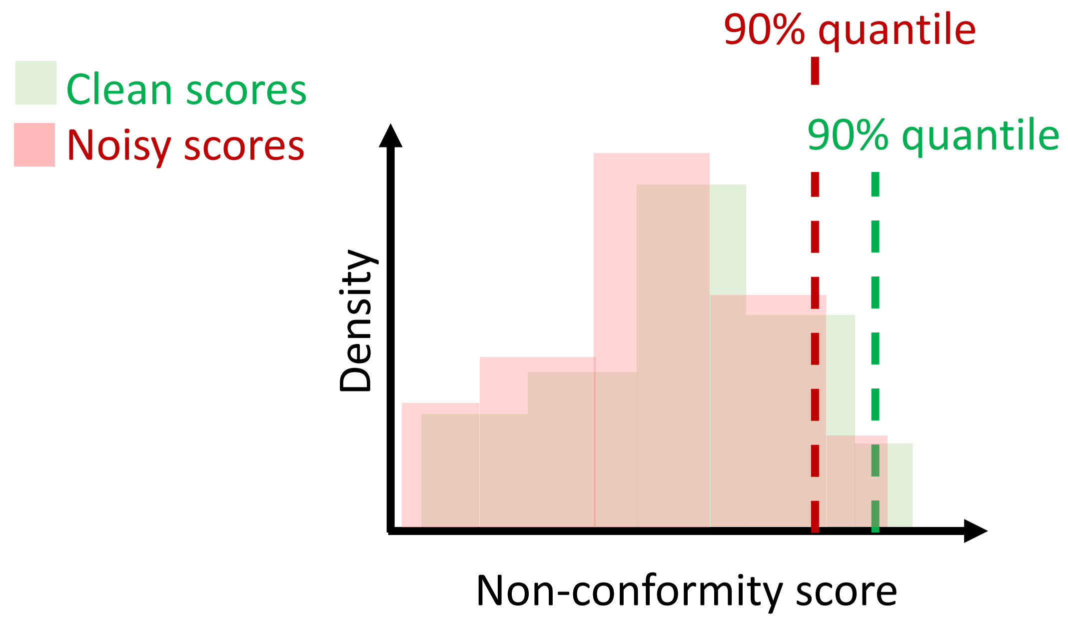


scores on clean < scores on noisy

e.g. $\mathrm{Var}(Y \mid X = x) < \mathrm{Var}(\tilde{Y} \mid X = x)$

Contractive vs. dispersive noise

contractive noise

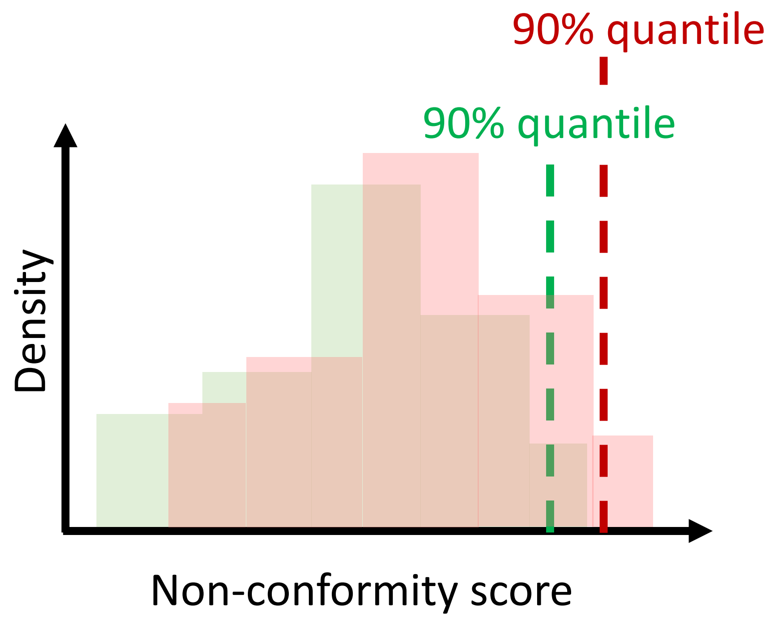


scores on clean > scores on noisy

e.g. $\mathrm{Var}(Y \mid X = x) > \mathrm{Var}(\tilde{Y} \mid X = x)$

Effect: **under**-coverage

dispersive noise



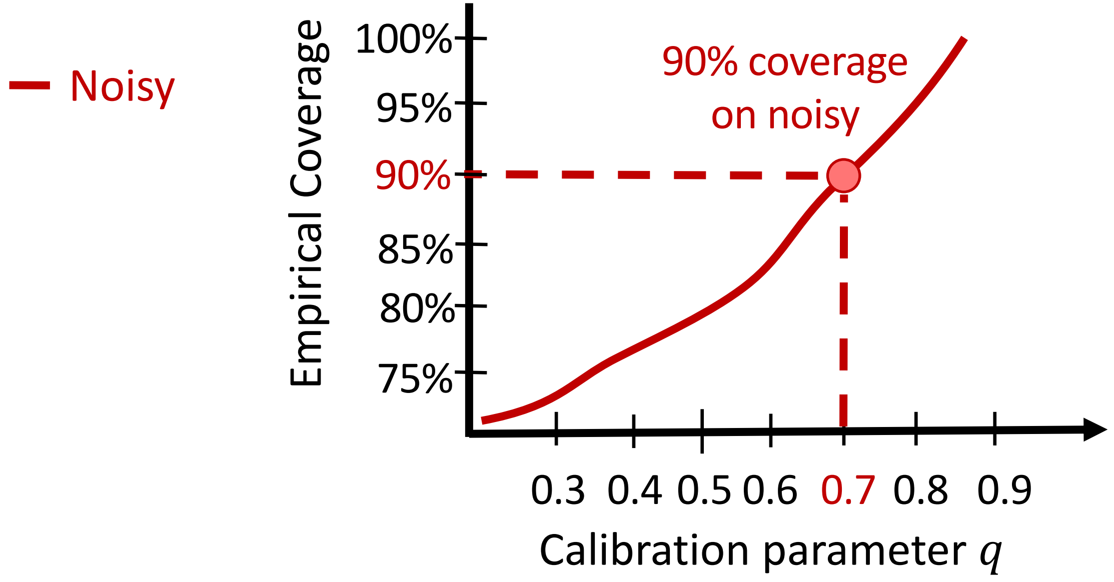
scores on clean < scores on noisy

e.g. $\mathrm{Var}(Y \mid X = x) < \mathrm{Var}(\tilde{Y} \mid X = x)$

Effect: **over**-coverage

Contractive vs. dispersive noise

contractive noise

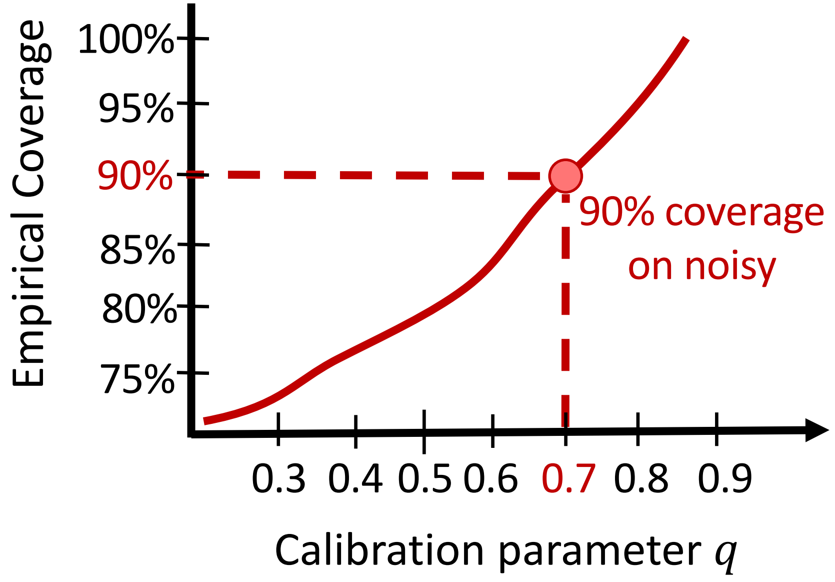


scores on clean > scores on noisy

e.g. $\mathrm{Var}(Y \mid X = x) > \mathrm{Var}(\tilde{Y} \mid X = x)$

Effect: **under**-coverage

dispersive noise



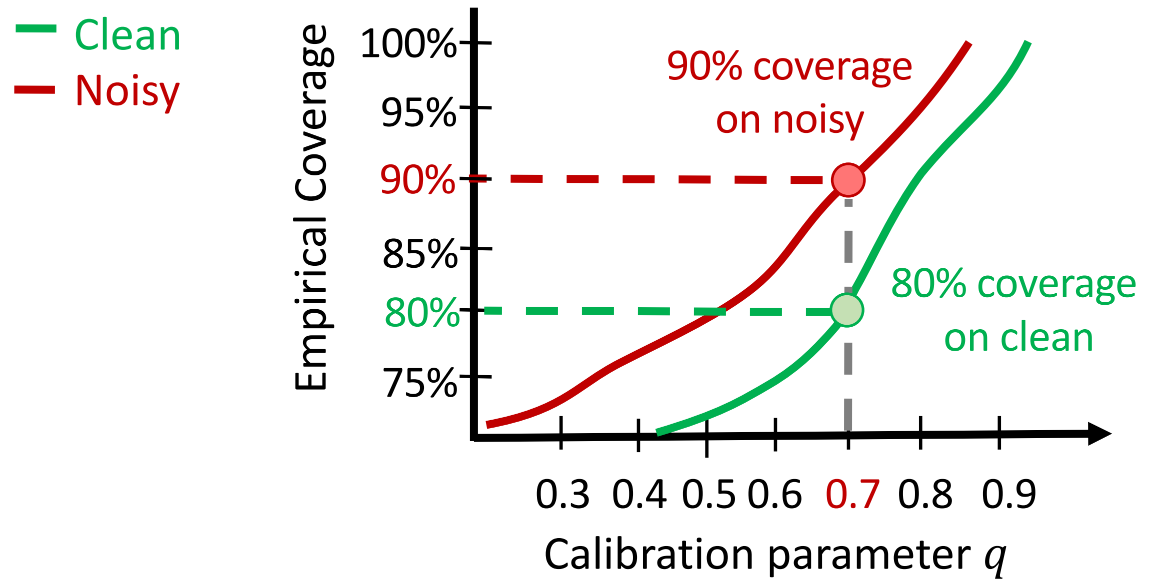
scores on clean < scores on noisy

e.g. $\mathrm{Var}(Y \mid X = x) < \mathrm{Var}(\tilde{Y} \mid X = x)$

Effect: **over**-coverage

Contractive vs. dispersive noise

contractive noise

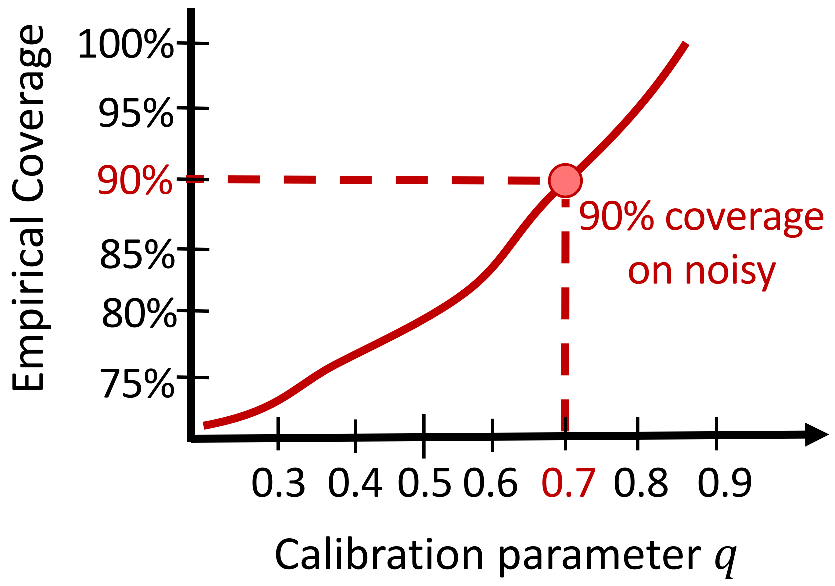


scores on clean > scores on noisy

e.g. $\mathrm{Var}(Y \mid X = x) > \mathrm{Var}(\tilde{Y} \mid X = x)$

Effect: **under**-coverage

dispersive noise



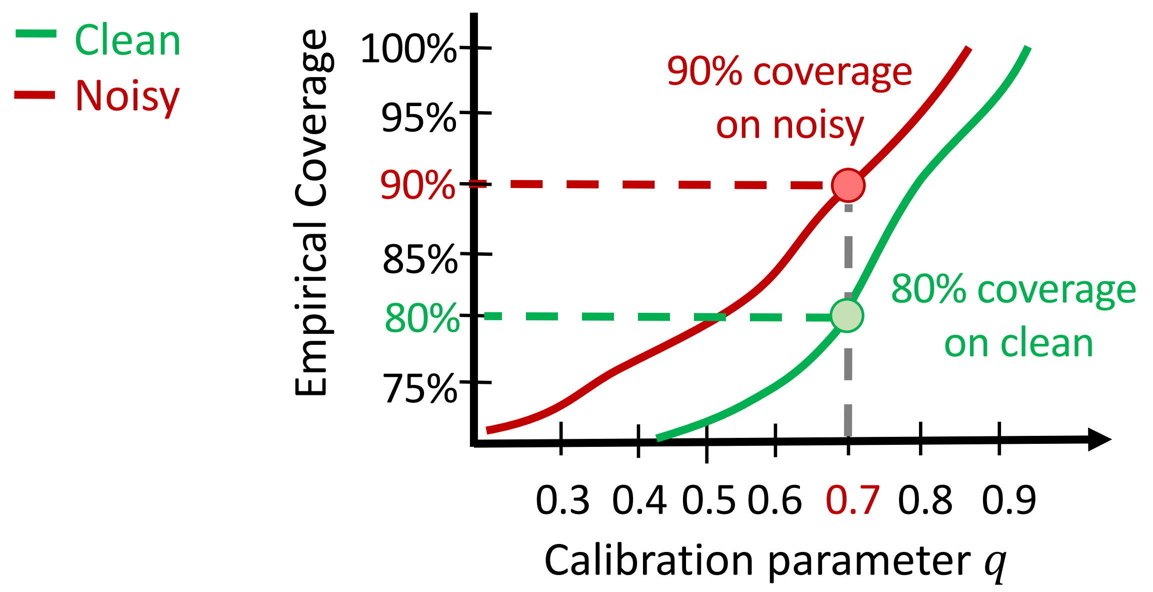
scores on clean < scores on noisy

e.g. $\mathrm{Var}(Y \mid X = x) < \mathrm{Var}(\tilde{Y} \mid X = x)$

Effect: **over**-coverage

Contractive vs. dispersive noise

contractive noise



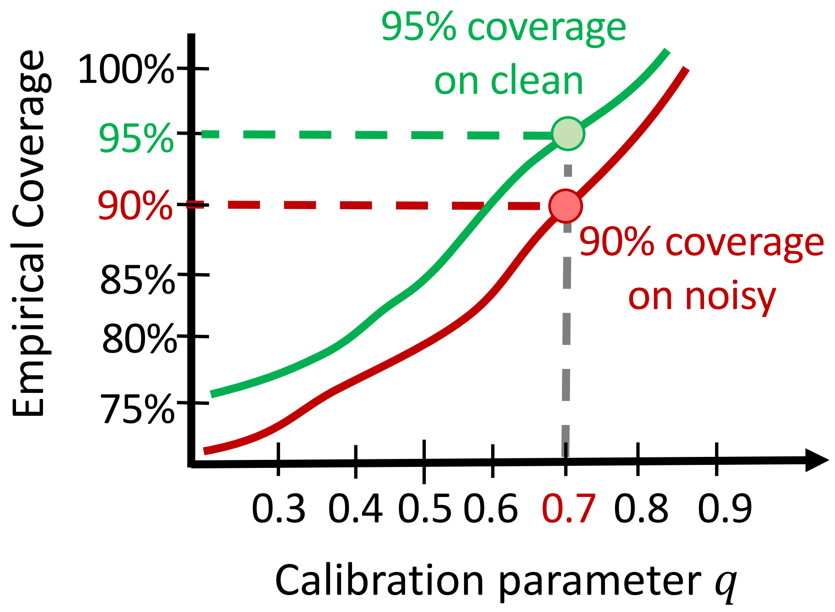
scores on clean > scores on noisy

e.g. $\mathrm{Var}(Y \mid X = x) > \mathrm{Var}(\tilde{Y} \mid X = x)$

<u>Effect</u>: **under**-coverage

VS

dispersive noise



scores on clean < scores on noisy

e.g. $\mathrm{Var}(Y \mid X = x) < \mathrm{Var}(\tilde{Y} \mid X = x)$

<u>Effect</u>: **over**-coverage

Formally: validity under dispersive noise

Theorem

If $\mathbb{P}\left(s\left(X_{\text{test}}, \tilde{Y}_{\text{test}}\right) \leq t\right) \leq \mathbb{P}(s(X_{\text{test}}, Y_{\text{test}}) \leq t)$ for all t, then

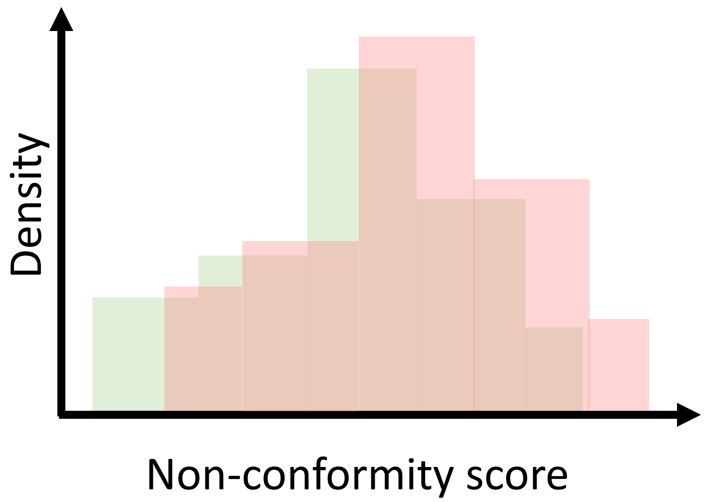
$$\mathbb{P}\left[Y_{\text{test}} \in C^{\text{noisy}}(X_{\text{test}})\right] \geq 1 - \alpha$$

- See paper for upper bound

Challenge: when does this assumption hold?

It's a function of (1) the clean data dist., (2) the noise, (3) the model performance, and (4) the score we use

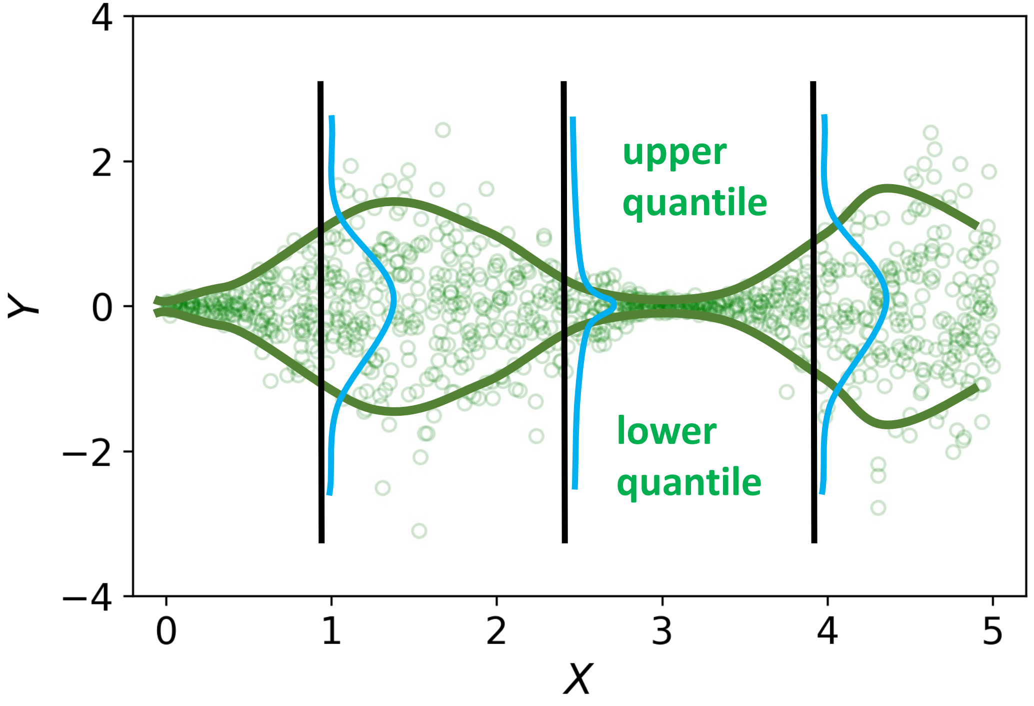
Visualizing the assumption



Regression

The ideal, oracle case

- Imagine we know the **true** conditional dist. of the clean data



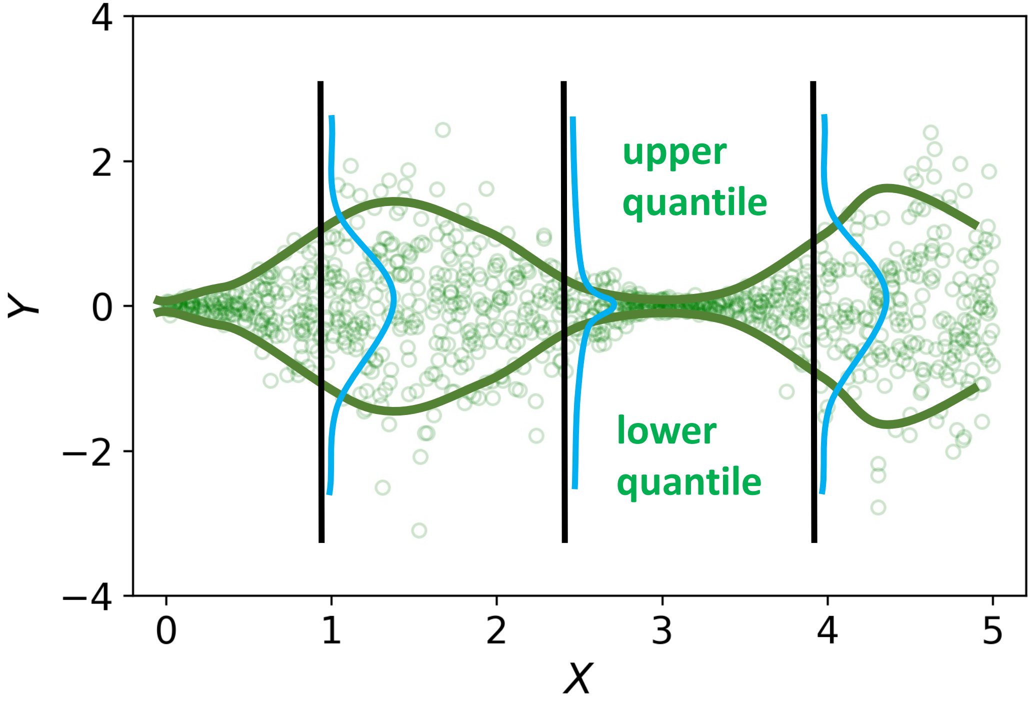
$\mathrm{lower}(x)$ = 0.05–th cond. quantile of $Y \mid X = x$

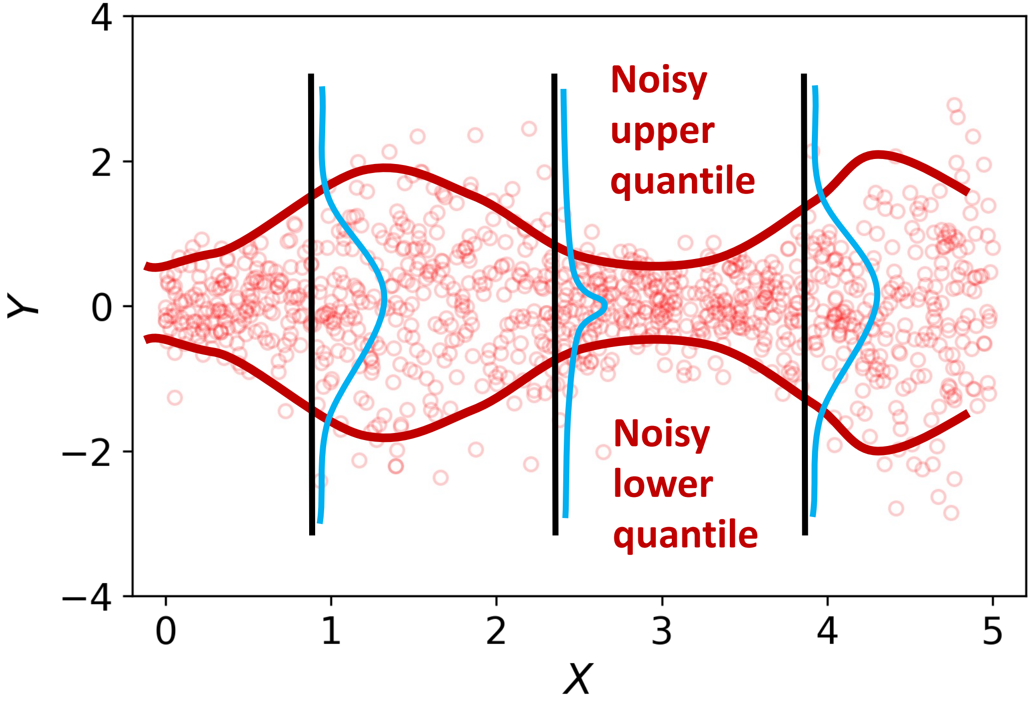
$\mathrm{upper}(x)$ = 0.95–th cond. quantile of $Y \mid X = x$

90% coverage by definition

The ideal, oracle case : noisy vs. clean

- Imagine we know the **true** conditional dist.
- What is the effect of noise? $\tilde{Y} = Y + Z$, the noise Z is symmetric around 0





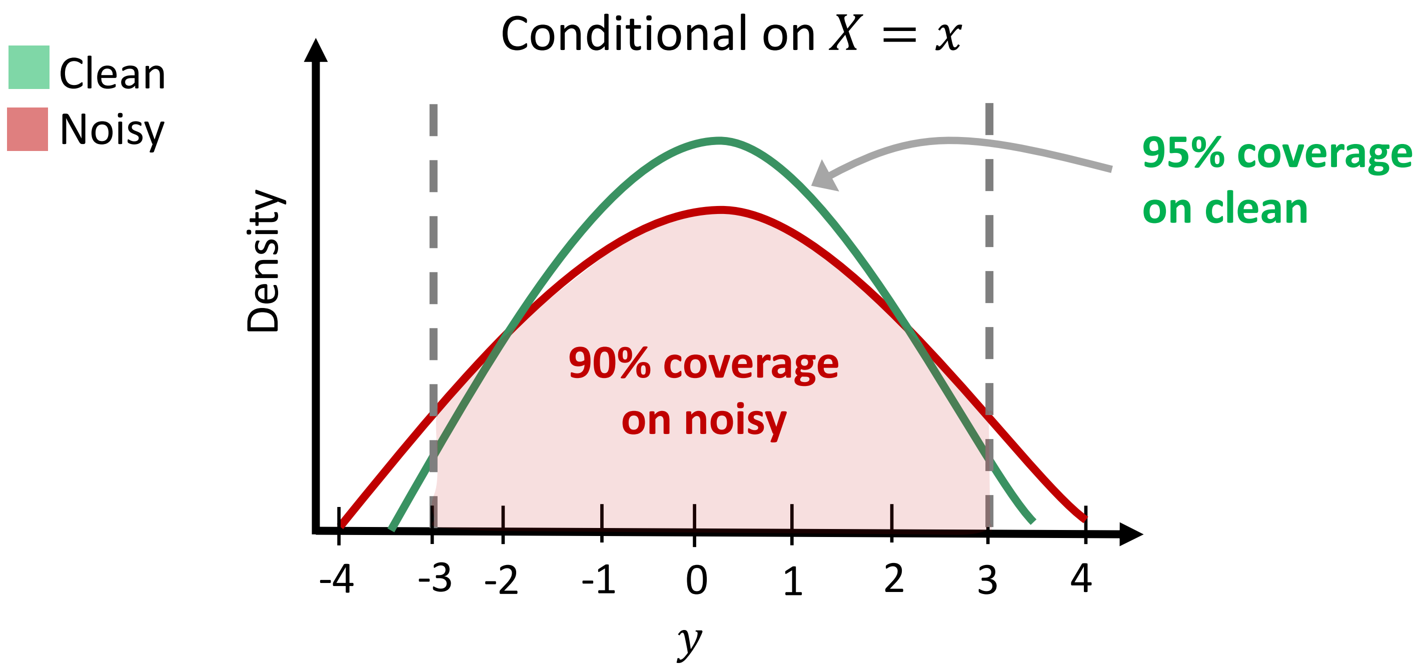
The noisy interval contains the clean interval

$\Downarrow$

higher coverage rate on clean

The ideal, oracle case : noisy vs. clean

- Imagine we know the **true** conditional dist.
- What is the effect of noise? $\tilde{Y} = Y + Z$, the noise Z is symmetric around 0



The noisy interval contains the clean interval

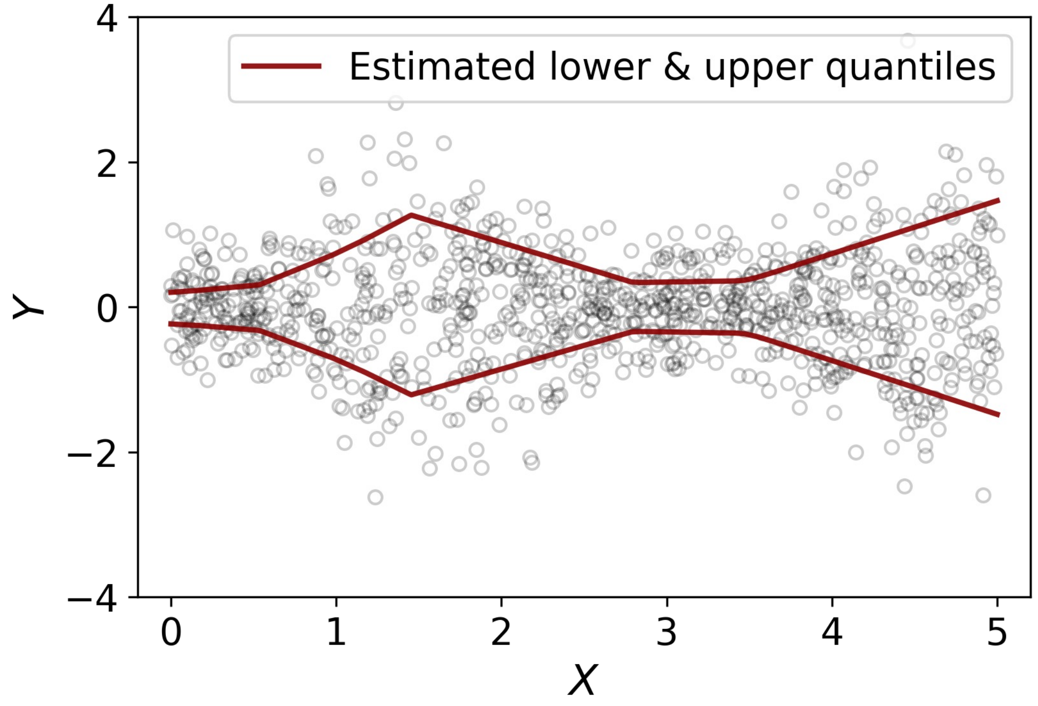
⇓

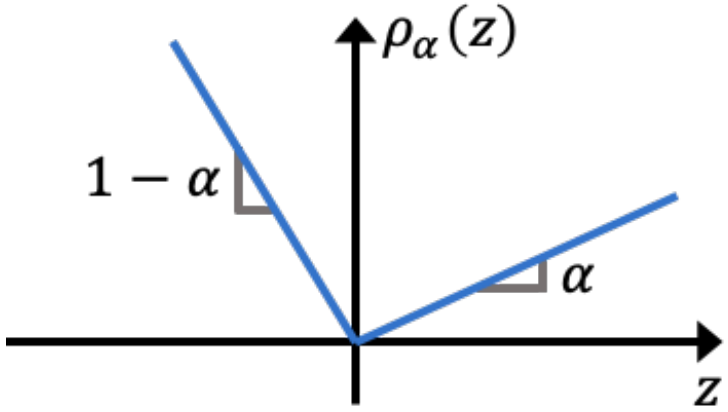
higher coverage rate on clean

Conformalized quantile regression (CQR) [R., Patterson, Candes ’19]

- Given a model that estimates the $\widehat{\text{lower}}(x)$ and $\widehat{\text{upper}}(x)$ cond. quantiles e.g., *quantile regression* model fitted to minimize the pinball loss [Koenker & Bassett ’78]

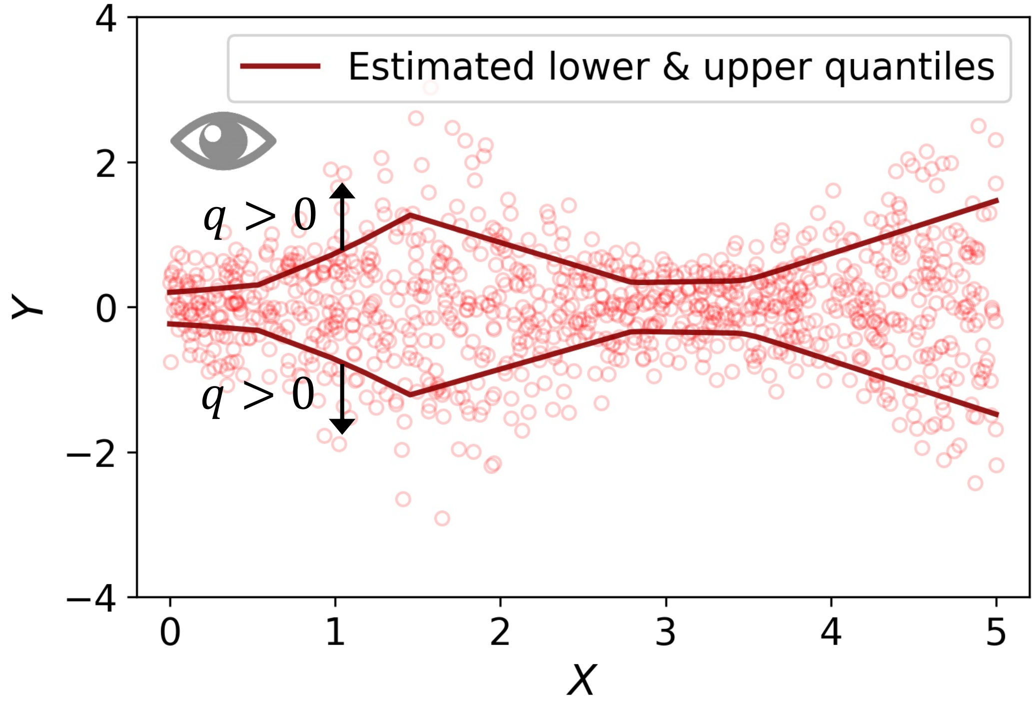
$$\widehat{\text{lower}}(x), \widehat{\text{upper}}(x) = \arg\min_{l,u} \textstyle\sum_i \rho_{\alpha_{\text{lo}}}\big(Y_i - l(X_i)\big) + \rho_{\alpha_{\text{up}}}\big(Y_i - u(X_i)\big)$$

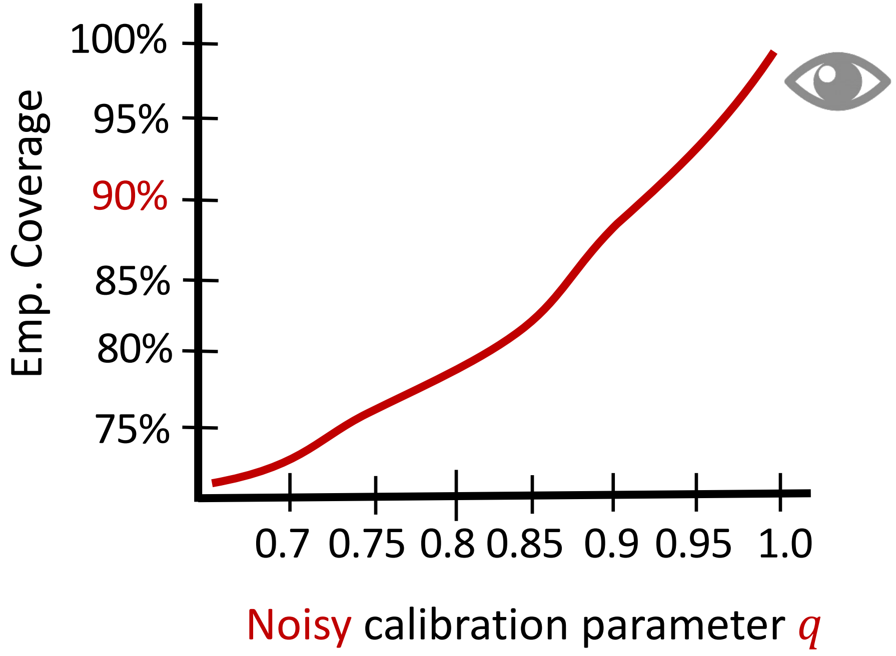




Conformalized quantile regression (CQR) [R., Patterson, Candes ’19]

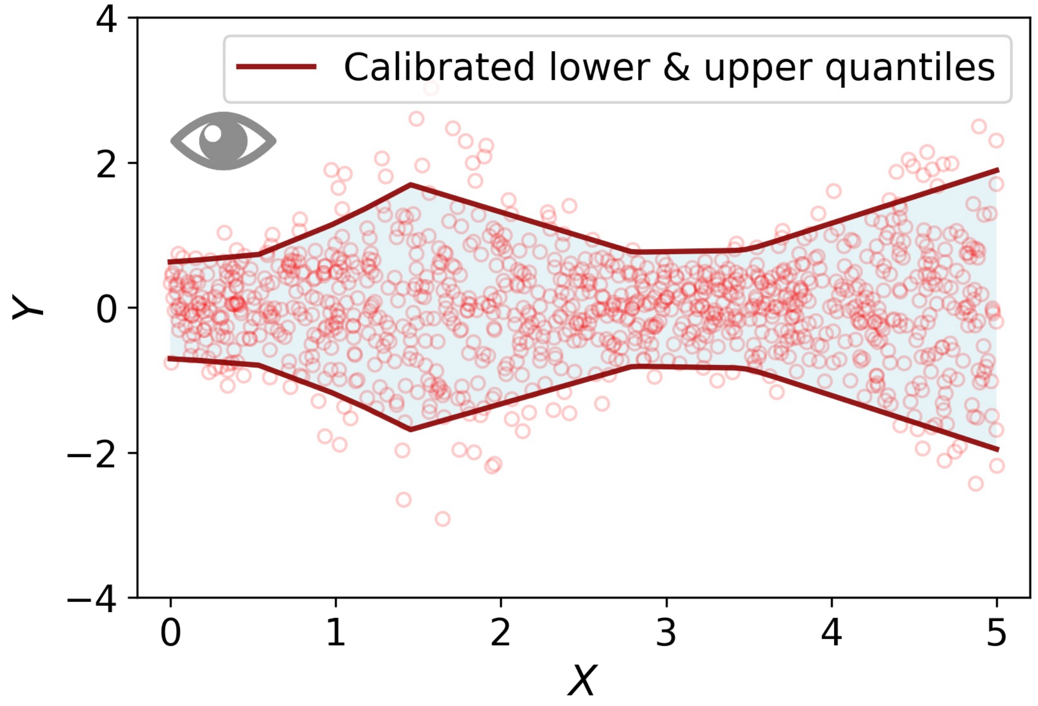
- Given a model that estimates the $\widehat{\text{lower}}(x)$ and $\widehat{\text{upper}}(x)$ cond. quantiles e.g., *quantile regression* model fitted to minimize the pinball loss
- CQR interval function: $C^{\text{noisy}}(x, q) = \left[\widehat{\text{lower}}(x) - q, \widehat{\text{upper}}(x) + q\right]$
- Calibrate the threshold $\hat{q}^{\text{noisy}}$ on the noisy calibration data

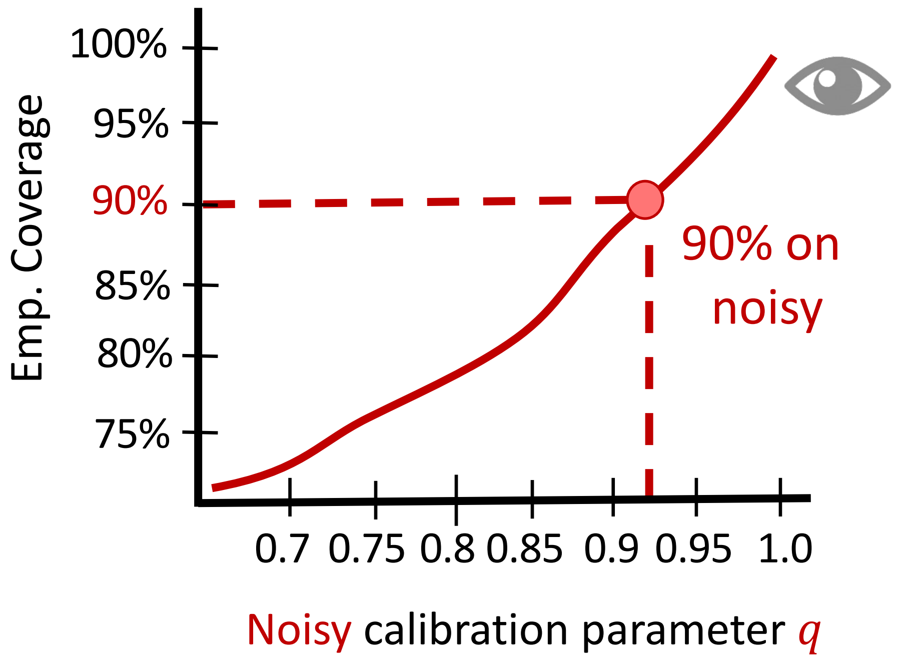




Conformalized quantile regression (CQR) [R., Patterson, Candes '19]

- Given a model that estimates the $\widehat{\text{lower}}(x)$ and $\widehat{\text{upper}}(x)$ cond. quantiles e.g., *quantile regression* model fitted to minimize the pinball loss
- CQR interval function: $C^{\text{noisy}}(x, q) = [\widehat{\text{lower}}(x) - q, \widehat{\text{upper}}(x) + q]$
- Calibrate the threshold $\hat{q}^{\text{noisy}}$ on the noisy calibration data





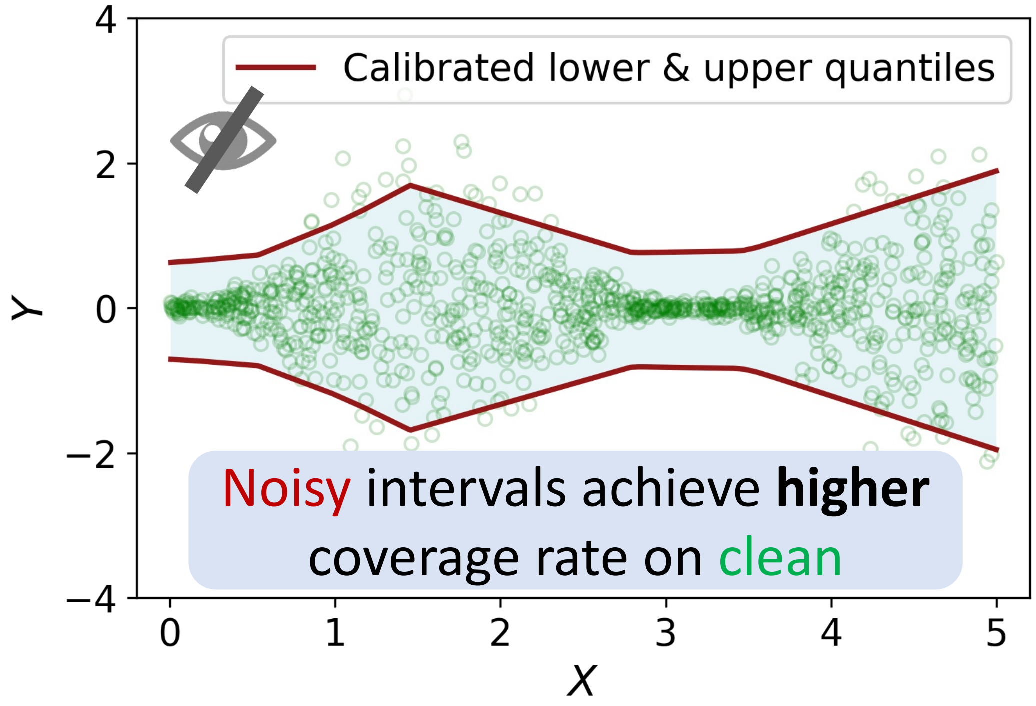
CQR is robust to dispersive noise

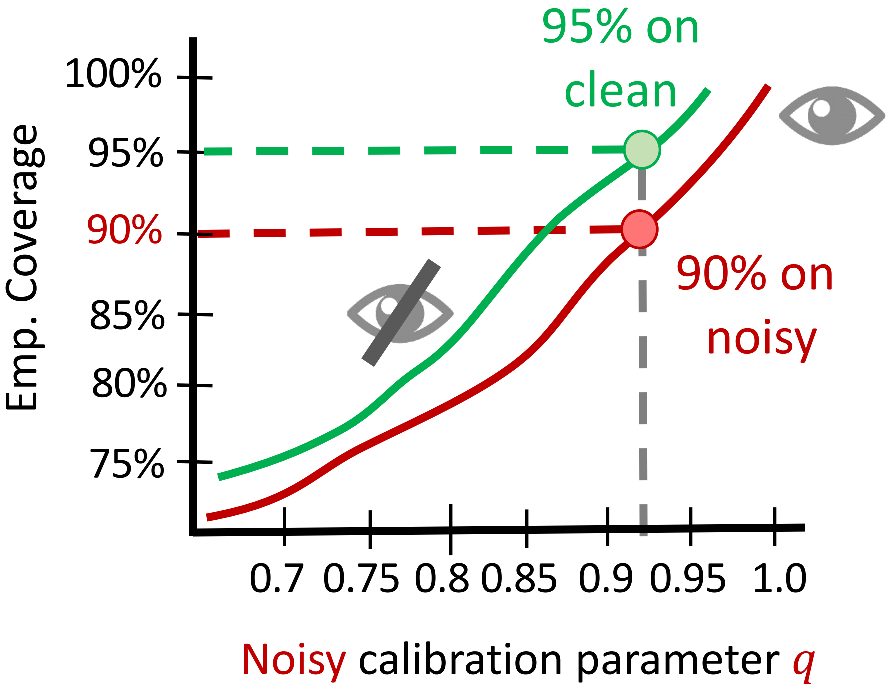
Assumptions

(1) $Y \mid X$ is symmetric & unimodal

(2) Z is symmetric around 0

- Calibrate the threshold $\hat{q}^{\text{noisy}}$ on the noisy calibration data

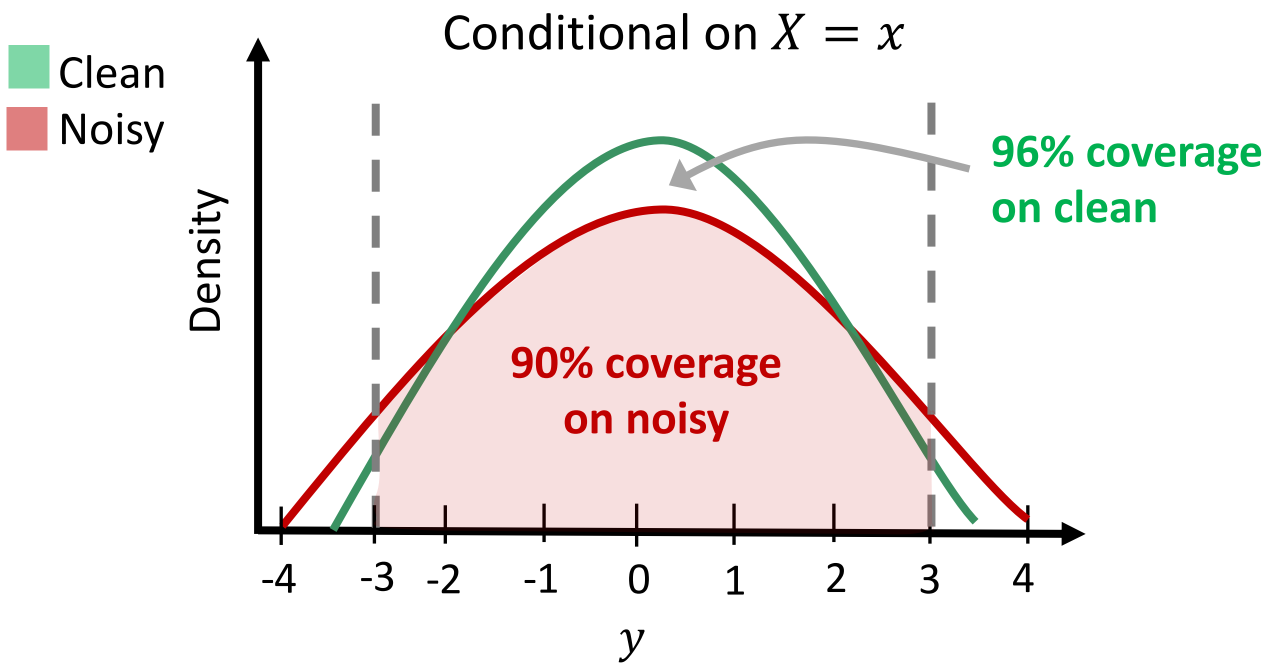




CQR is robust to dispersive noise

Suppose that $Y \mid X$ is symmetric and unimodal. Suppose further that noisy $\tilde{Y} = Y + Z$ where Z is symmetric around 0. If $\widehat{\mathrm{lower}}(x) \leq \mathrm{median}(x) \leq \widehat{\mathrm{upper}}(x)$, then

$$\mathbb{P}\left[Y_{\mathrm{test}} \in C^{\mathrm{noisy}}(X_{\mathrm{test}})\right] \geq 1 - \alpha$$



Remark

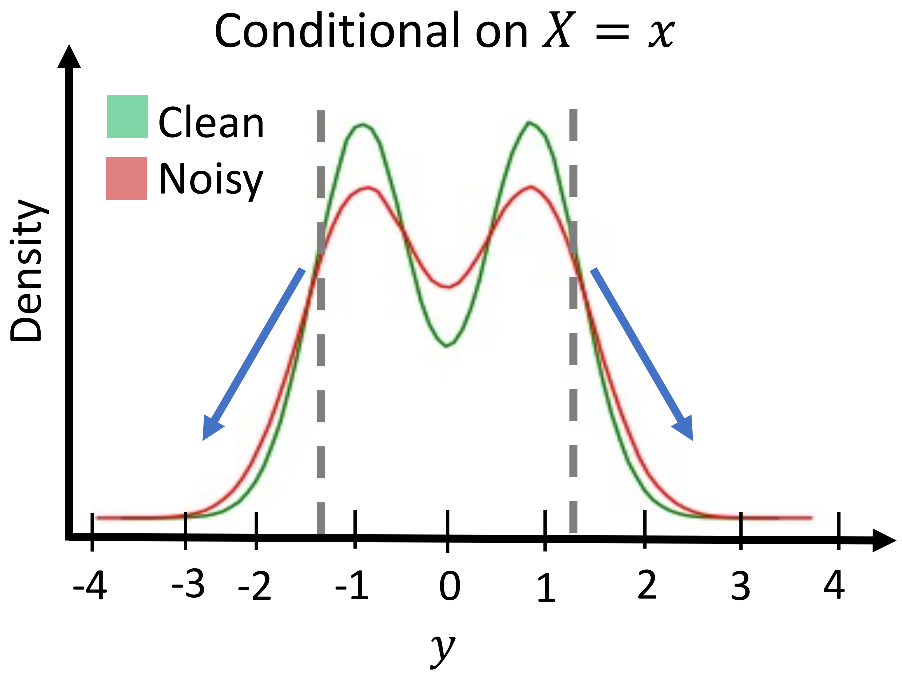
+ Weak assumption on the model

− Strong assumption on the data

Relaxing the distributional assumption

We say that the density of $Y \mid X = x$ is peaked inside the interval $\left[q^{\text{lower}}, q^{\text{upper}}\right]$ if for all $t \geq 0$:

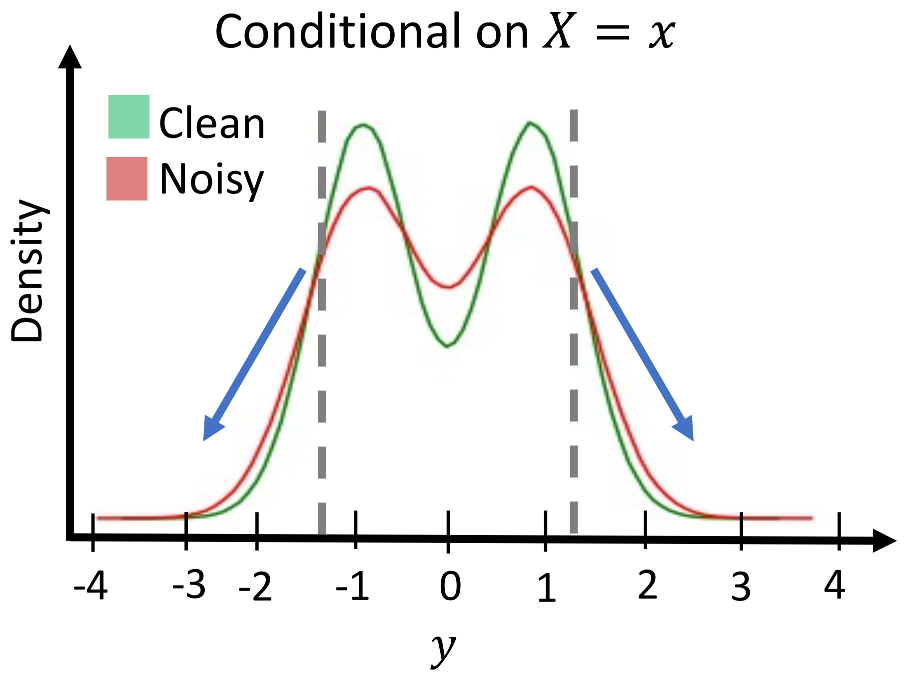
$$f_{Y|X=x}(q^{\text{upper}} + t) \leq f_{Y|X=x}(q^{\text{upper}} - t)$$
$$f_{Y|X=x}\left(q^{\text{lower}} + t\right) \geq f_{Y|X=x}\left(q^{\text{lower}} - t\right)$$



General robustness proposition

Suppose that $\tilde{Y} = Y + Z$ where Z is symmetric around 0. If the density of $Y \mid X = x$ is peaked inside $C^{\text{noisy}}(X_{\text{test}})$, then

$$\mathbb{P}\left[Y_{\text{test}} \in C^{\text{noisy}}(X_{\text{test}})\right] \geq 1 - \alpha$$



Remark

+ Weaker assumptions on the data

− Stronger assumptions on the model

Inclusion between results

Master Theorem $\supseteq$ General regression result $\supseteq$ Unimodal result

Multi-class classification

The noise setting

- Multi-class classification with K classes
- Random flip corruption

$$\tilde{Y} = g^{\text{flip}}(Y, U) = \begin{cases} Y & \text{w.p } 1-\varepsilon \\ Y' & \text{otherwise} \end{cases}$$

Y' is drawn uniformly from $\{1,2,\dots,K\}$

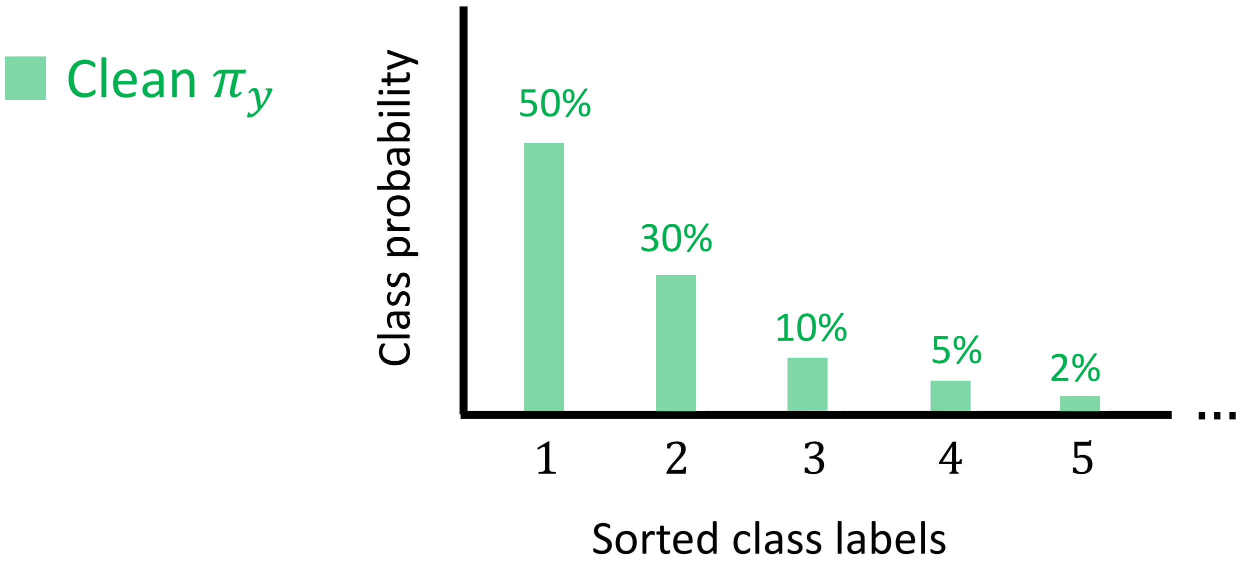
[Angluin & Laird ’88; Aslam & Decatur ’96; Ma et al. ‘18; Jenni & Favaro ’18; Yuan et al. ‘18]

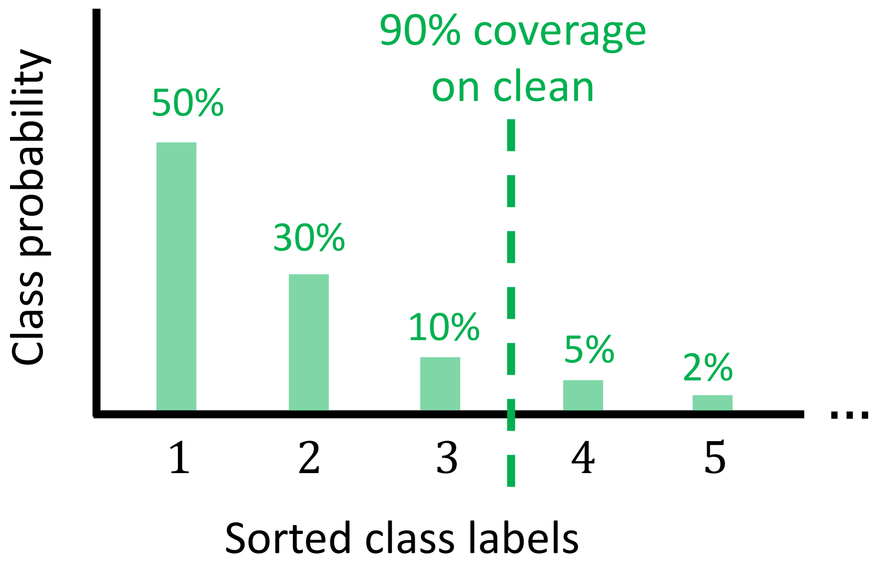
The ideal, oracle case

- Imagine we know the true conditional class probabilities of the clean data

$$\pi_y(x) = \mathbb{P}[Y = y \mid X = x]$$

- How to construct a prediction set for $Y \mid X$?





$$C^{\text{ideal}}(x, q = 0.9) = \{1,2,3\}$$

The ideal, oracle case: noisy vs. clean

- Imagine we know the true conditional class probabilities of the clean data

$$\pi_y(x) = \mathbb{P}[Y = y \mid X = x]$$

$\mathbb{P}[\tilde{Y} = y \mid X = x]$

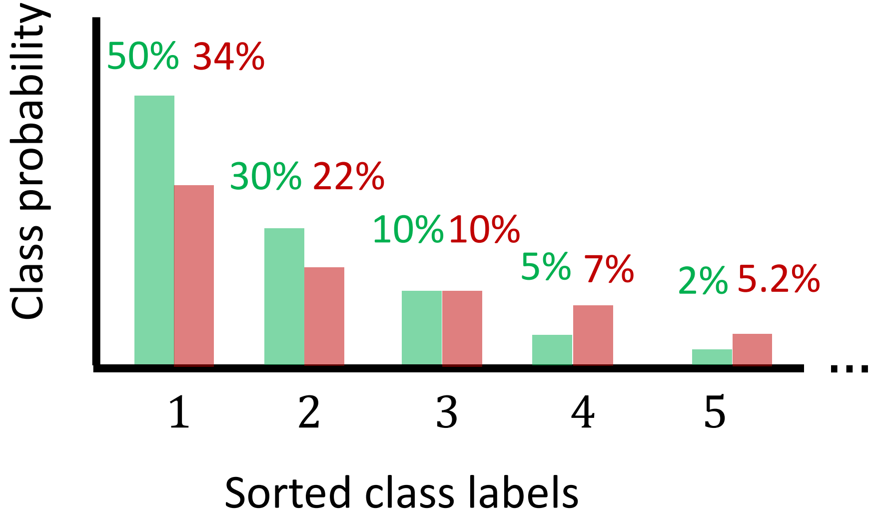
- What is the effect of noise = label is flipped w.p. ϵ ?

$$\tilde{\pi}_y(x) = (1-\epsilon)\pi_y(x) + \epsilon \frac{1}{K}$$

Clean π_y

Noisy $\tilde{\pi}_y$

Smaller noise ϵ



Higher noise ϵ

$K = 10, \varepsilon = 0.1, 0.4$

The ideal, oracle case: noisy vs. clean

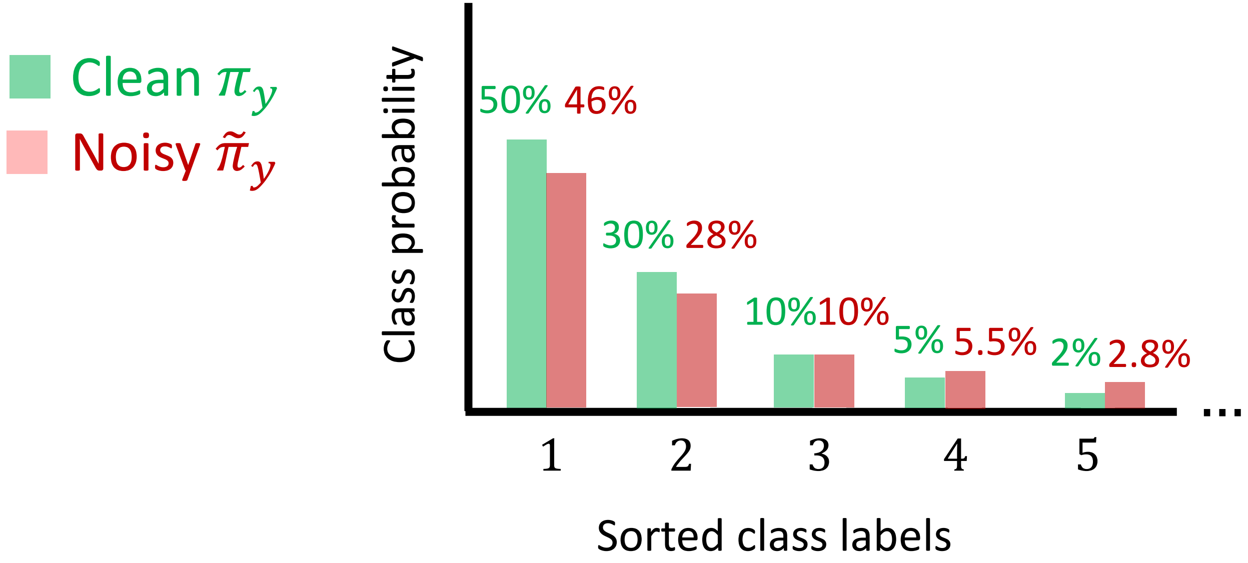
- Imagine we know the true conditional class probabilities of the clean data

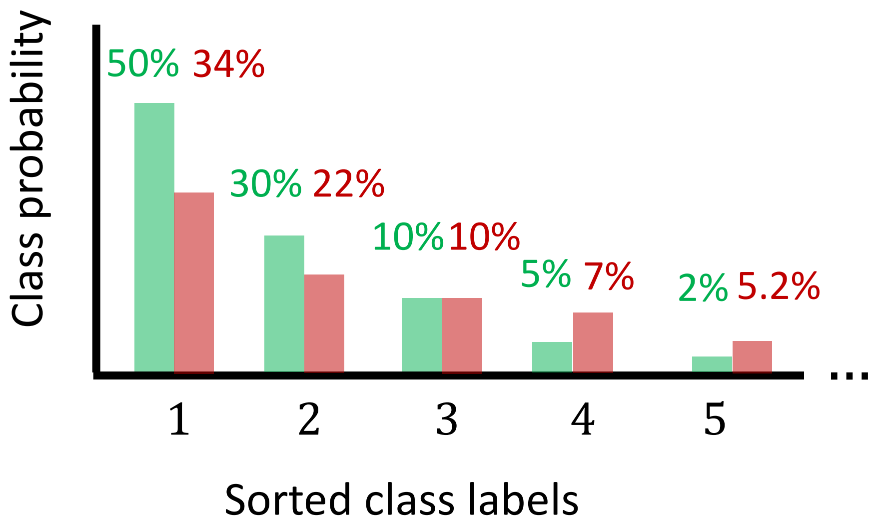
$$\pi_y(x) = \mathbb{P}[Y = y \mid X = x]$$

$\mathbb{P}[\tilde{Y} = y \mid X = x]$

- What is the effect of noise = label is flipped w.p. ϵ ?

$$\tilde{\pi}_y(x) = (1 - \epsilon)\pi_y(x) + \epsilon \frac{1}{K}$$

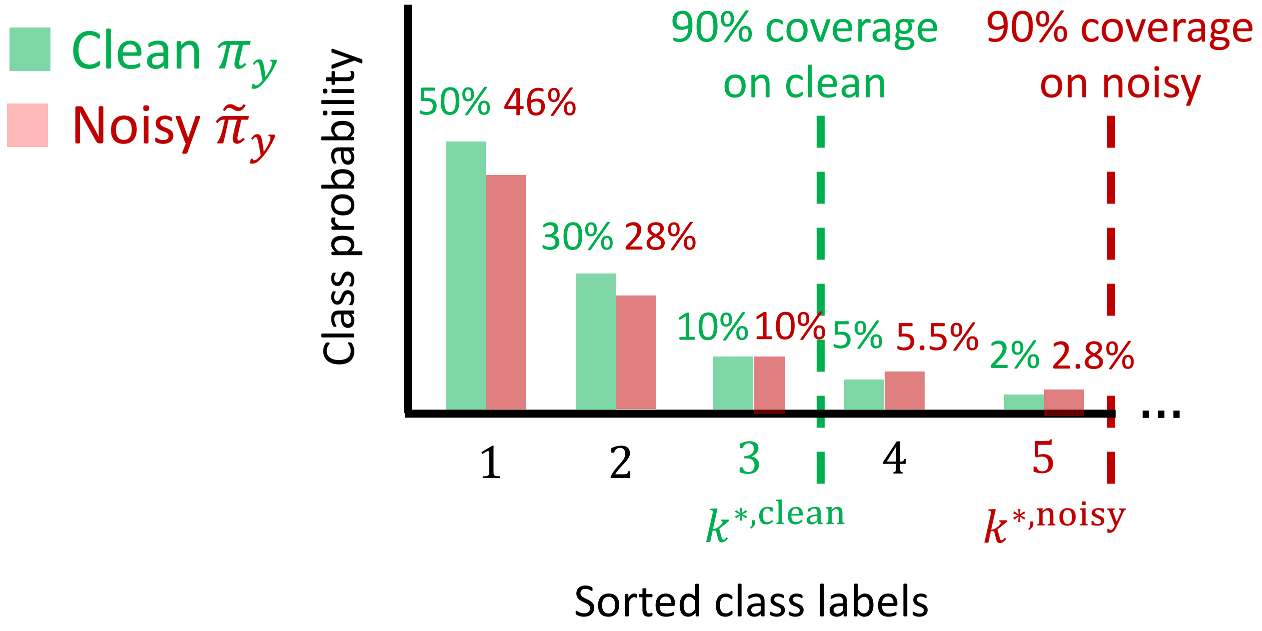




1. The noisy class probs. get closer to uniform as ϵ increases
2. The orderings of the clean/noisy class probs. are identical

Oracle achieves conservative coverage on clean

- Constructing sets with threshold $q^{\text{noisy}} = 0.9$; run the procedure as if data is clean



Ideal clean set $C^{\text{ideal}}_{\text{clean}}(x) = \{1,2,3\}$

Noisy set $C^{\text{noisy}}(x, q^{\text{noisy}} = 0.9) = \{1,2,3,4,5\}$

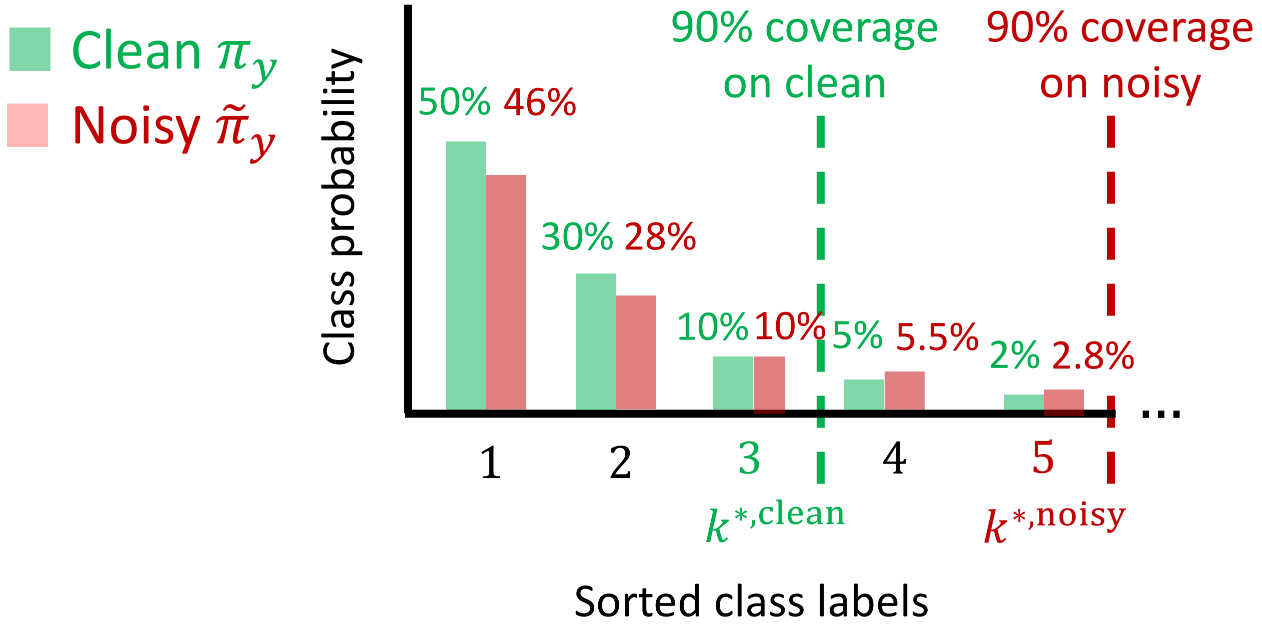
The noisy set contains all the labels of the clean

$\Downarrow$

higher coverage rate on clean

Oracle achieves conservative coverage on clean

- Constructing sets with threshold $q^{\text{noisy}} = 0.9$; run the procedure as if data is clean



$$k^{*,\text{noisy}} = \left\{ \min k : \sum_{j=1}^{k} \tilde{\pi}_{(j)}(x) \geq 0.9 \right\}$$

$$= \left\{ \min k : \sum_{j=1}^{k} \pi_{(j)}(x) + \epsilon \underbrace{\left(\frac{k}{K} - \sum_{j=1}^{k} \pi_{(j)}(x) \right)}_{\leq 0} \geq 0.9 \right\}$$

$$\geq k^{*,\text{clean}}$$

Ideal clean set $C^{\text{ideal}}_{\text{clean}}(x) = \{1,2,3\}$

Noisy set $C^{\text{noisy}}\left(x, q^{\text{noisy}} = 0.9\right) = \{1,2,3,4,5\}$

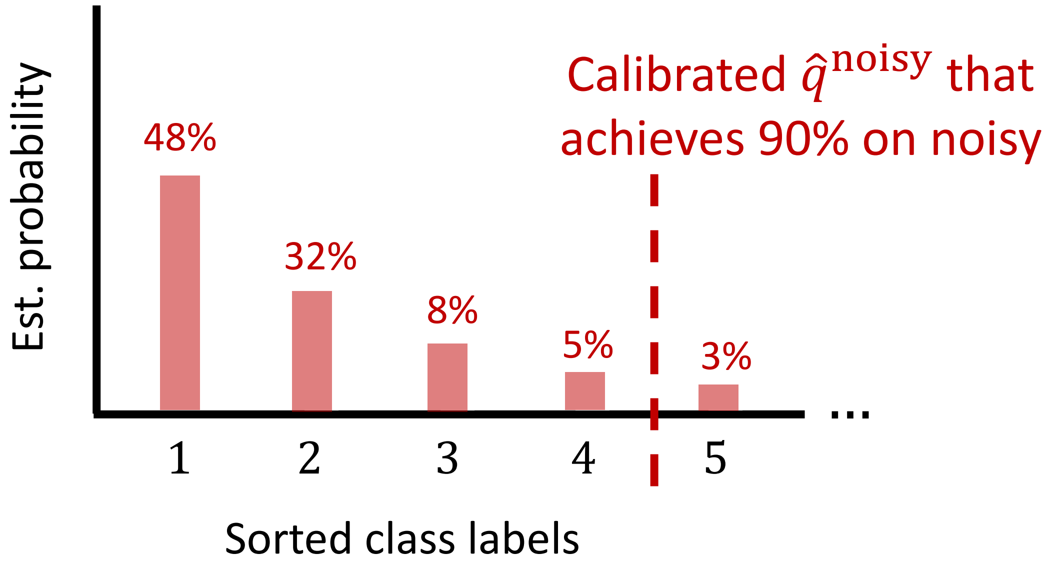
The noisy set contains all the labels of the clean

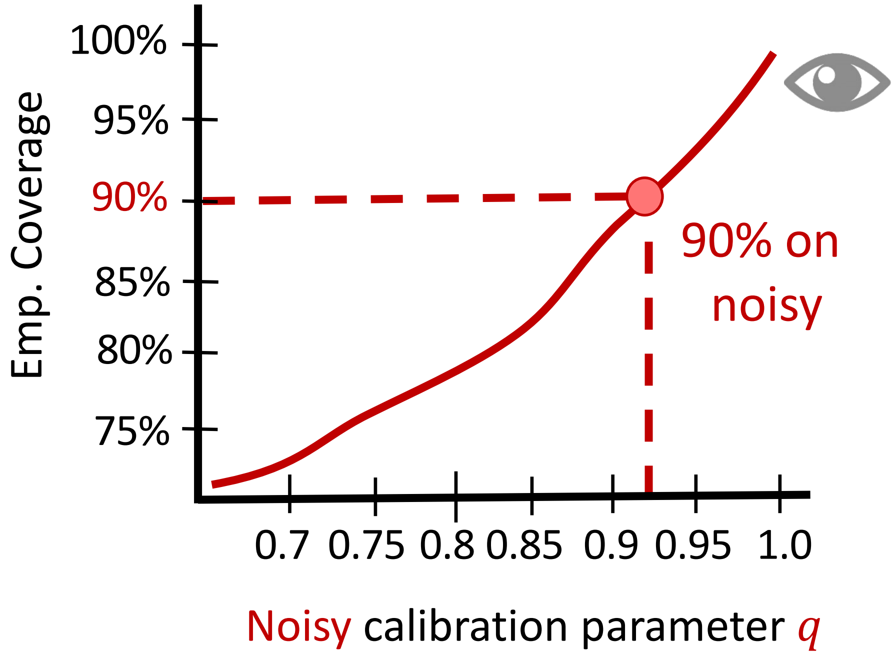
⇓

higher coverage rate on clean

Conformal APS [R., Sesia, Candes ('20)]

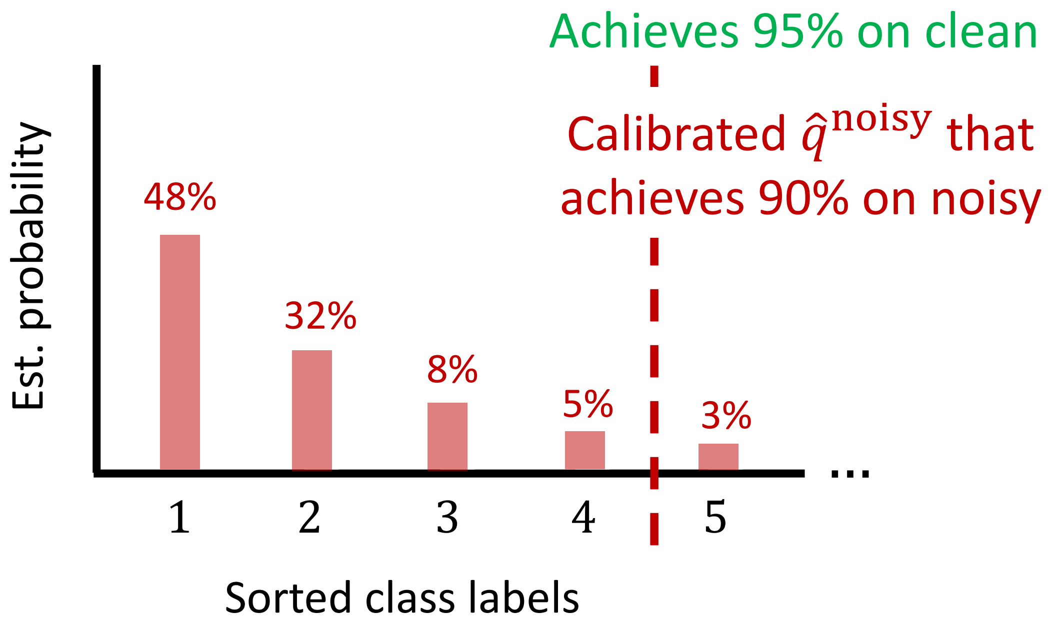
- Given a classifier $\hat{\pi}_y(x)$ that estimates the conditional class probabilities e.g., output of the softmax layer of a NNet
- Calibrate the threshold $\hat{q}^{\text{noisy}}$ on the noisy calibration data

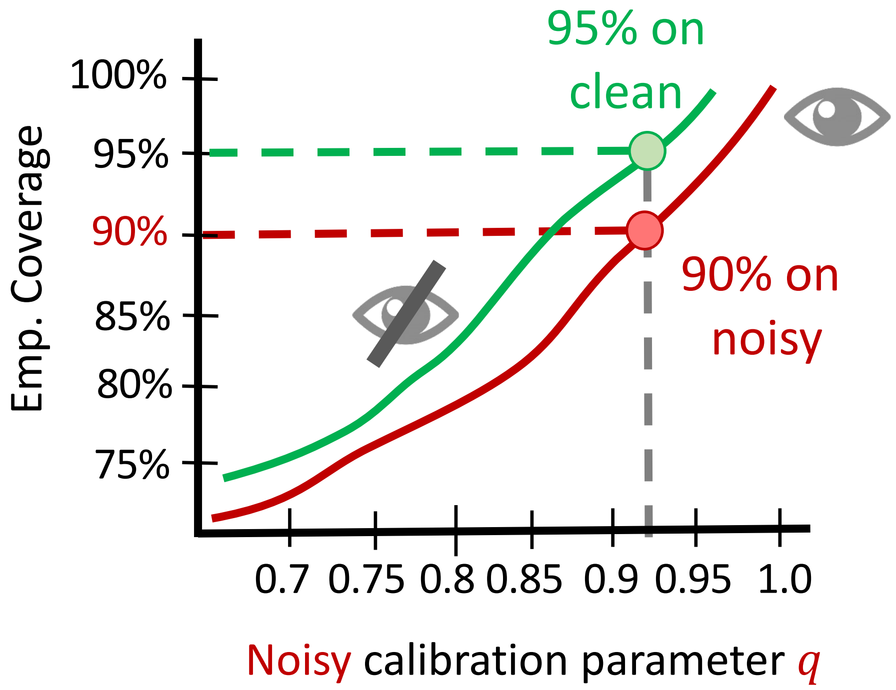




Conformal APS is robust to dispersive noise

- Given a classifier $\hat{\pi}_y(x)$ that estimates the conditional class probabilities e.g., output of the softmax layer of a NNet
- Calibrate the threshold $\hat{q}^{\text{noisy}}$ on the noisy calibration data
- **Assumption**: the classifier ranks the classes in the same order as the oracle $\mathbb{P}(\tilde{Y} \mid X)$



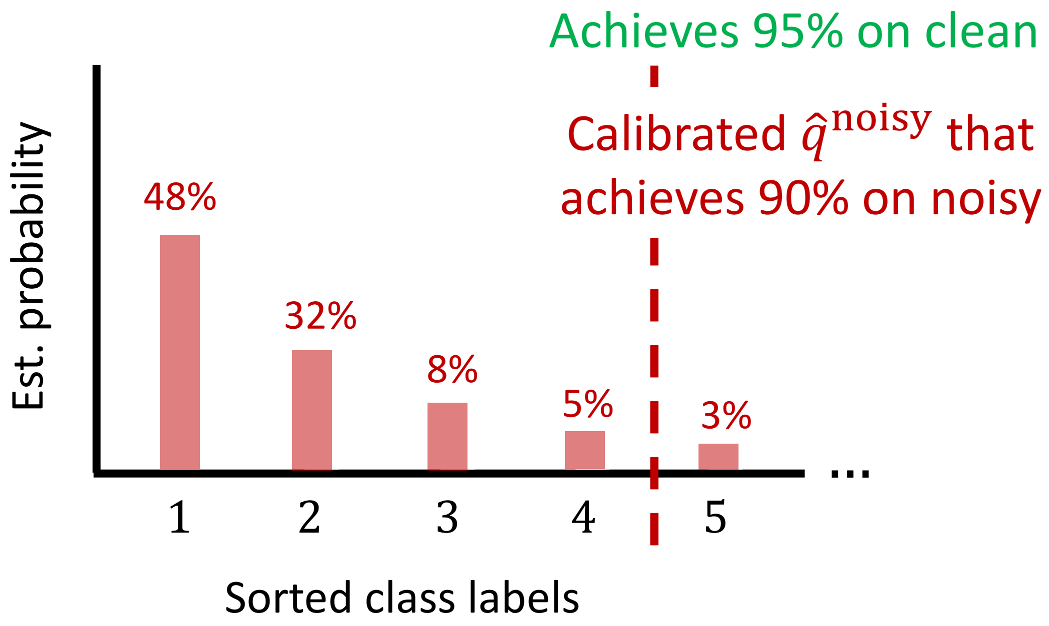


Robustness under dispersive noise

Assume a random flip noise model. If the classifier ranks the classes in the same order as the oracle $\mathbb{P}(\tilde{Y} \mid X)$, then

$$\mathbb{P}\left[Y_{\text{test}} \in C^{\text{noisy}}(X_{\text{test}})\right] \geq 1 - \alpha$$

- See paper for upper bound



Remark

+ Relatively weak assumptions on the data

– Strong assumptions on the classifier (correct rankings)

General noise setting

- **The key requirement for general noise robustness (intuition)**:
  the noise should (a) push the class probabilities closer to uniform while (b) preserving the class-probability ordering for all $x \in \mathcal{X}$
- Formally, assume for all $i, j \in \{1, \ldots, k\}$

(a) $\left|\mathbb{P}[\tilde{Y} = i \mid X = x] - \frac{1}{k}\right| \leq \left|\mathbb{P}[Y = i \mid X = x] - \frac{1}{k}\right|$

(b) $\mathbb{P}[\tilde{Y} = i \mid X = x] \leq \mathbb{P}[\tilde{Y} = j \mid X = x] \Leftrightarrow \mathbb{P}[Y = i \mid X = x] \leq \mathbb{P}[Y = j \mid X = x]$

- Then,

$$\mathbb{P}\left[Y_{\text{test}} \in C^{\text{noisy}}(X_{\text{test}})\right] \geq 1 - \alpha$$

Inclusion between results

Master Theorem $\supseteq$ General classification result $\supseteq$ Random flip result

Risk control: moving beyond the miscoverage loss

Multi-label classification

- $X \in \mathcal{X}$: an image
- $Y \in \mathcal{Y}$: clean labels, e.g., {car, dog, house}
- $\tilde{Y} \in \mathcal{Y}$: noisy labels, e.g., {truck, cat, house}
- Random-flip noise model

$$\tilde{Y}[j] = \begin{cases} Y[\mathrm{j}], & \text{w.p. } 1-\varepsilon, \\ 1 - Y[\mathrm{j}], & \text{otherwise} \end{cases}$$

- Varying #objects across different images
- **High dim.** $\boldsymbol{Y}$

→ want less stringent notion of error than miscoverage $= 1\left[Y_{\text{test}} \notin C^{\text{noisy}}(X_{\text{test}})\right]$

[Angelopoulos & Bates et al. '21; Angelopoulos et al. '21, '22]

Credit: DALL-E 2

Conformal risk control: prediction sets with controlled risk

[Angelopoulos et al. ’21; Angelopoulos et al. ’21, ‘22]

- **Goal (multi-label class.)**: construct prediction sets with a controlled *false negative rate*

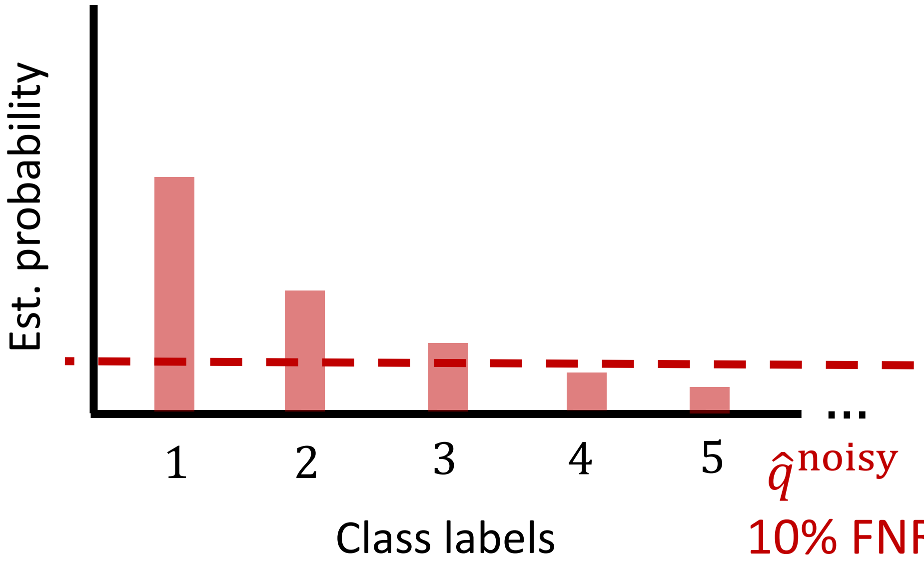
$$\underbrace{\mathbb{E}\Big[\overbrace{L^{\text{FNP}}\Big(Y_{\text{test}}, C^{\text{noisy}}(X_{\text{test}})\Big)}^{\text{Loss}}\Big]}_{\text{Risk}} \leq \alpha \quad \text{(e.g., 10\%)}$$

false negative proportion (FNP) loss:

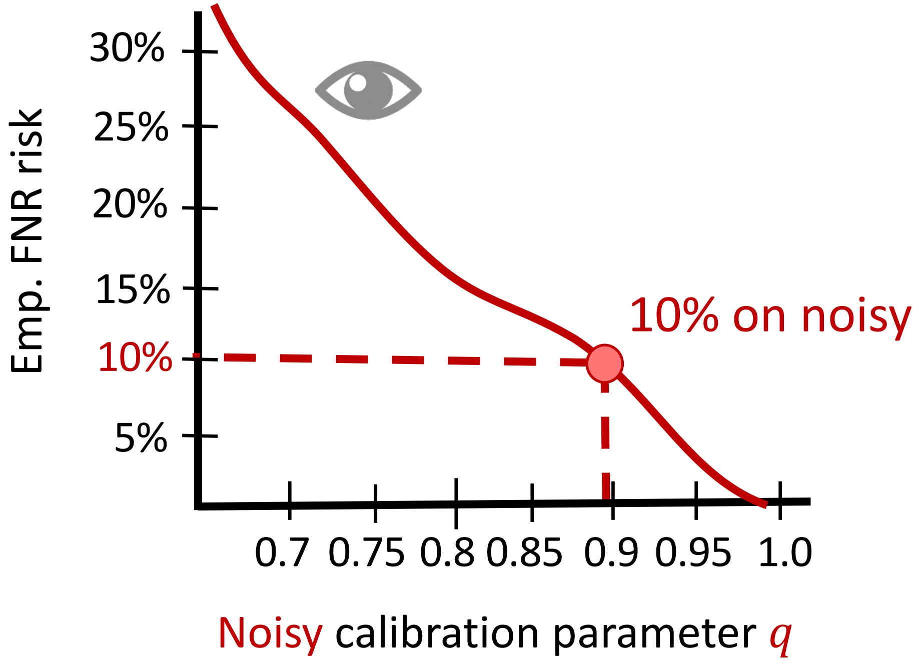
$$L^{\text{FNP}}\Big(y, C^{\text{noisy}}(x)\Big) = 1 - \frac{|y \cap C^{\text{noisy}}(x)|}{|y|} = 1 - \frac{\text{\# of lables covered}}{\text{total \# of labels}}$$

Conformal risk control: FNR for multi-label classification

- Given a classifier $\hat{\pi}_y(x)$ that estimates the conditional class probabilities
- Set function: $C^{\text{noisy}}(x, q) = \{y : \hat{\pi}_y(x) \geq 1 - q\}$ [Angelopoulos et al. '21]
- Calibrate the threshold $\hat{q}^{\text{noisy}}$ on the noisy calibration data



$\hat{q}^{\text{noisy}}$ achieves 10% FNR on noisy



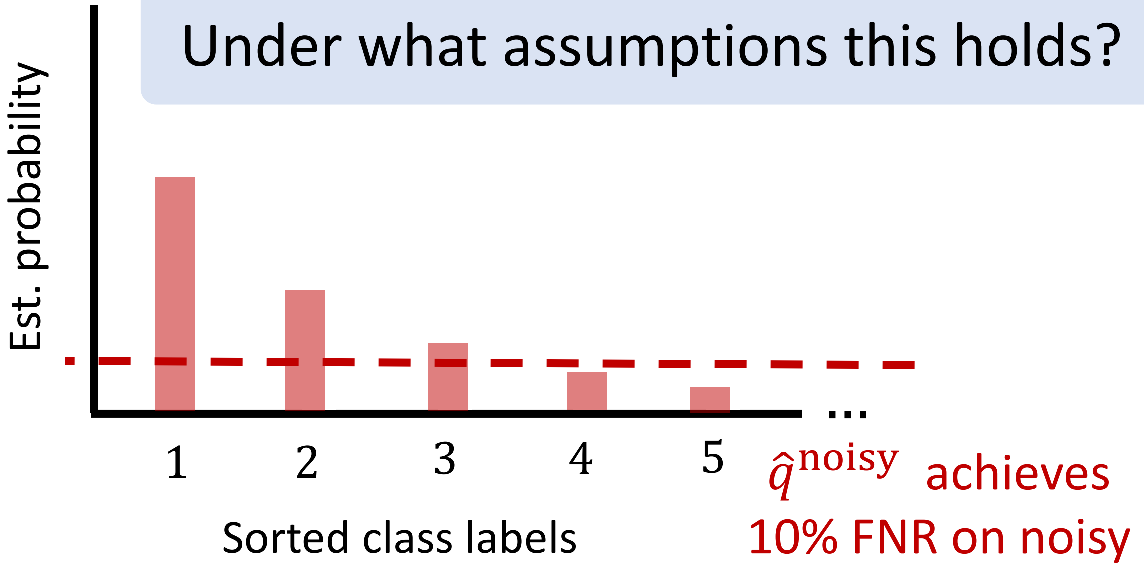
$$C^{\text{noisy}}(x, q^{\text{noisy}}) = \{1,2,3\}$$

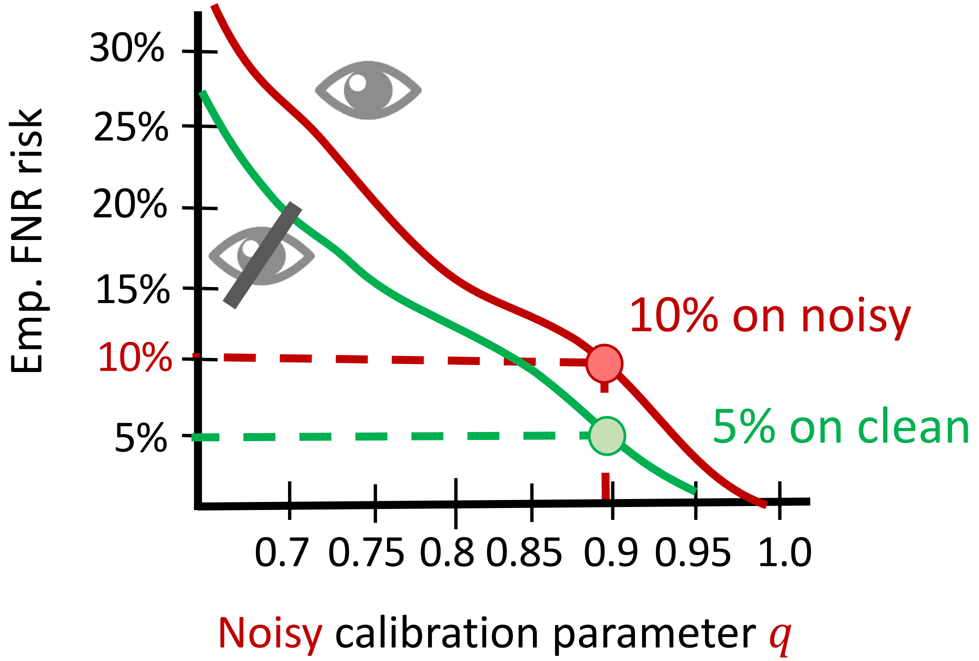
Monotonicity: $q_1 \geq q_2 \Rightarrow C(x, q_2) \subseteq C(x, q_1)$

Conformal risk control: FNR for multi-label classification

- Given a classifier $\hat{\pi}_y(x)$ that estimates the conditional class probabilities
- Set function: $C^{\text{noisy}}(x, q) = \{y : \hat{\pi}_y(x) \geq 1 - q\}$ [Angelopoulos et al. '21]
- Calibrate the threshold $\hat{q}^{\text{noisy}}$ on the noisy calibration data

Under what assumptions this holds?





$$C^{\text{noisy}}(x, q^{\text{noisy}}) = \{1,2,3\}$$

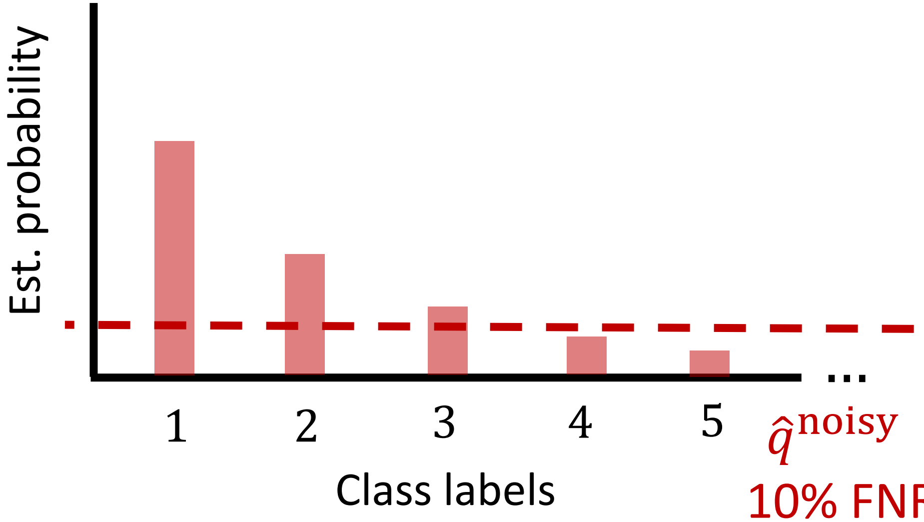
Monotonicity: $q_1 \geq q_2 \Rightarrow C(x, q_2) \subseteq C(x, q_1)$

Conformal risk control is robust to label noise

Assume a random flip noise model. Assume also that

1. The classifier ranks the classes in the same order as the oracle $\mathbb{P}(\tilde{Y} = y \mid X = x)$
2. The clean labels are conditionally independent: $Y[i] \perp Y[j] \mid X = x$ for all pairs (i, j)

$$\Longrightarrow \mathbb{E}\left[L^{\mathrm{FNP}}\left(Y_{\mathrm{test}}, C^{\mathrm{noisy}}(X_{\mathrm{test}})\right)\right] \leq \alpha$$



$\hat{q}^{\mathrm{noisy}}$ achieves 10% FNR on noisy

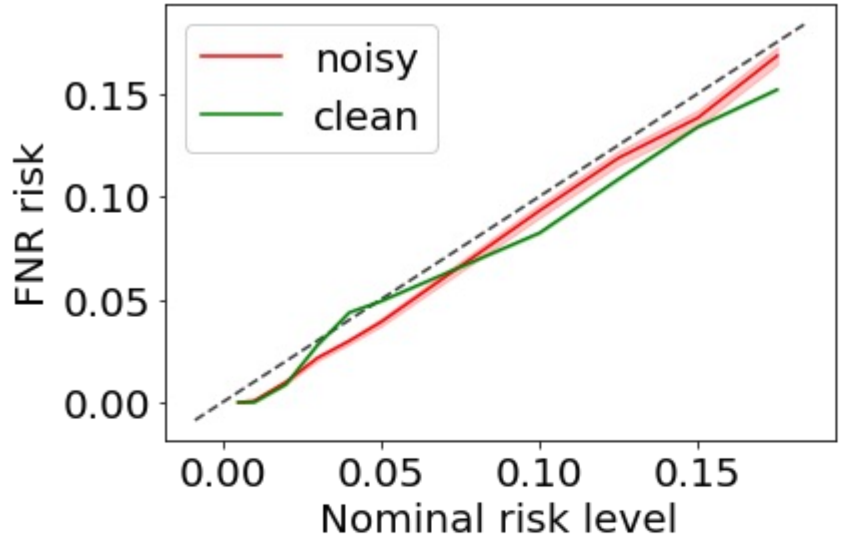
$C^{\mathrm{noisy}}(x, q^{\mathrm{noisy}}) = \{1,2,3\}$

Remark

Robustness can be guaranteed even if
1. the noise does not have the same magnitude across all labels
2. the labels are dependent

Experiment: MS COCO image data [Lin et al. ’14]

- Task: classify the objects in an image ($K = 80$ classes)
- Clean COCO : clean Y are original labels
- Noisy COCO : we collected 117 noisy $\tilde{Y}$ from single annotators (calibration set)
- NNet classifier (TResNet) [Ridnik et al. ’20]



- Exact control on noisy labels (not surprising)
- Valid control on clean labels

Conclusion, open questions, and uncovered topics

Takwaway: *accurate model + dispersive noise = conservative coverage*

Caution: there are cases where conformal **would not** obtain valid coverage (adv. noise)

Uncovered topics

- Segmentation problems
- Online, time-varying settings with drifting dist.
  → adaptive conformal inference (coverage) [Gibbs & Candes '21,'22]
  rolling risk control (FNR risk) [Feldman et al. '22]

Next step?

– Design conformity scores that are robust to label noise

Thank you!