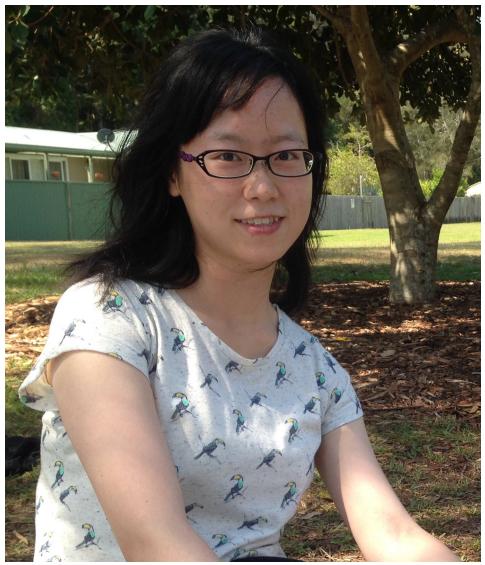
Adaptive conformal classification with noisy labels

Matteo Sesia

University of Southern California
Department of Data Sciences and Operations

September 9, 2024

Collaborators



Rachel Wang
(U. Sydney)

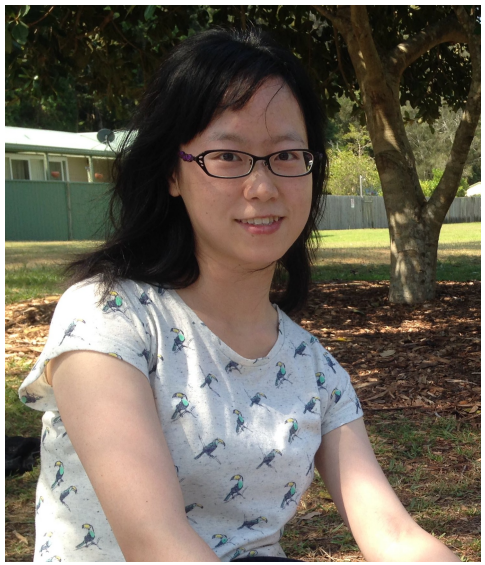
Xin Tong (USC)

Teresa Bortolotti
(Politecnico di Milano)

Reference:

"Adaptive conformal classification with noisy labels"
arxiv.org/abs/2309.05092
September 2023, under review

Collaborators



Rachel Wang
(U. Sydney)

Xin Tong (USC)

Teresa Bortolotti
(Politecnico di Milano)

Reference:

"Adaptive conformal classification with noisy labels"
arxiv.org/abs/2309.05092
September 2023, under review

Statistics and Conformal Inference in the Age of AI

This report underscores the transformative role of machine learning and AI in scientific research, national security, and societal welfare.

However, key concerns include:

- Inadequate **uncertainty quantification**.

Statistics and Conformal Inference in the Age of AI

This report underscores the transformative role of machine learning and AI in scientific research, national security, and societal welfare.

However, key concerns include:

- Inadequate **uncertainty quantification**.
- Inaccuracies in large data sets, especially problems with **improperly labeled data**.

Statistics and Conformal Inference in the Age of AI

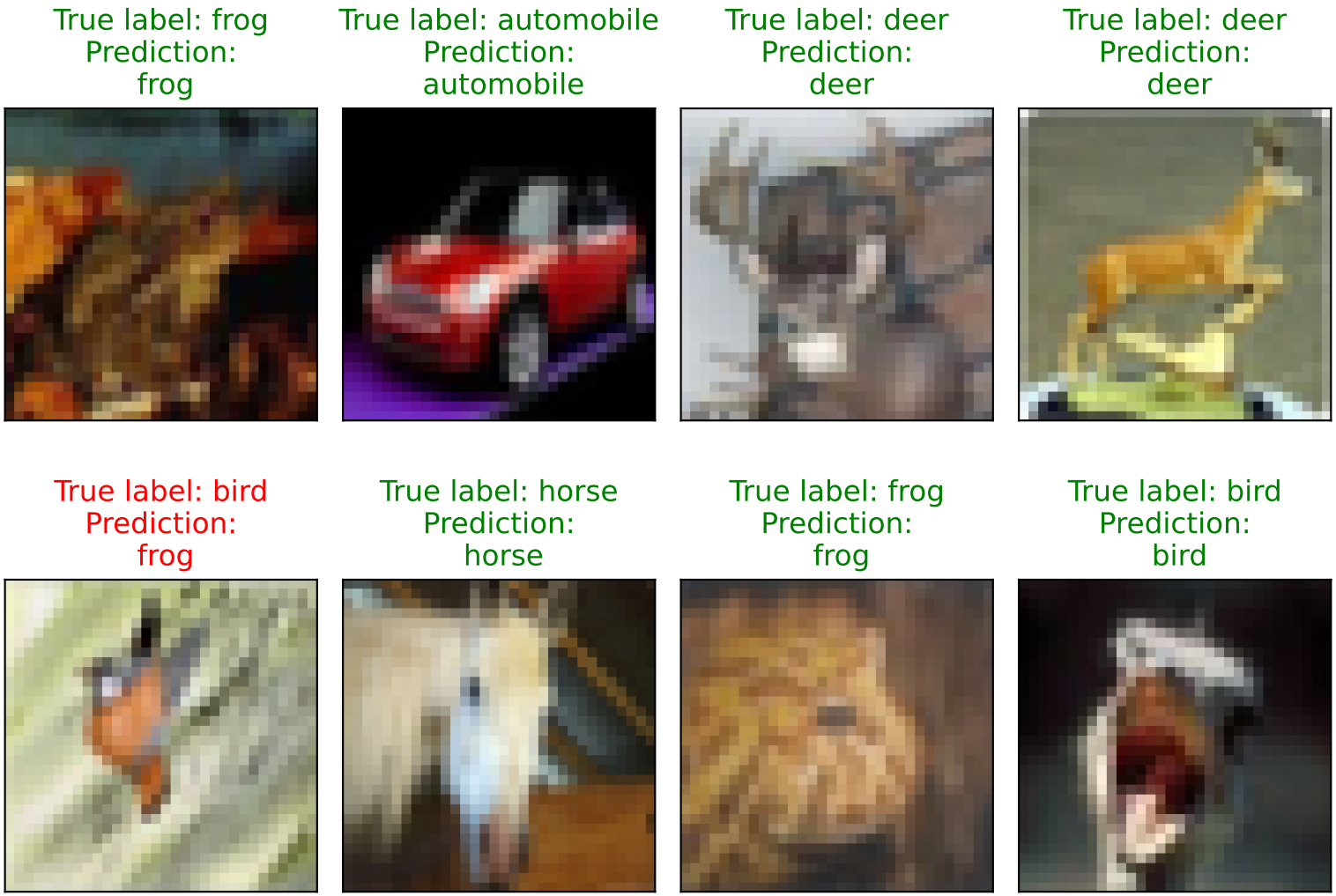
This report underscores the transformative role of machine learning and AI in scientific research, national security, and societal welfare.

However, key concerns include:

- Inadequate **uncertainty quantification**.
- Inaccuracies in large data sets, especially problems with **improperly labeled data**.
- Data privacy issues, often managed by introducing **random noise** into data.

Uncertainty Quantification (UQ) for Classification

ML models sometimes make mistakes, and are often *overconfident*.



Uncertainty Quantification (UQ) for Classification

ML models sometimes make mistakes, and are often *overconfident*.

Conformal prediction quantifies uncertainty through *prediction sets*.

True label: frog
Prediction set:
{frog}

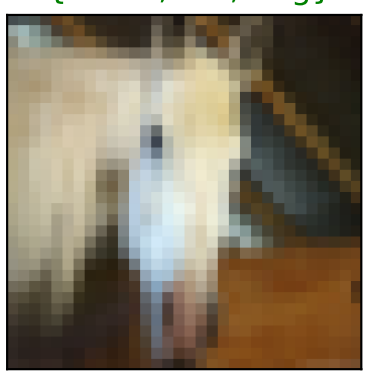
True label: automobile
Prediction set:
{automobile}

True label: deer
Prediction set:
{deer, dog}

True label: deer
Prediction set:
{deer, bird}

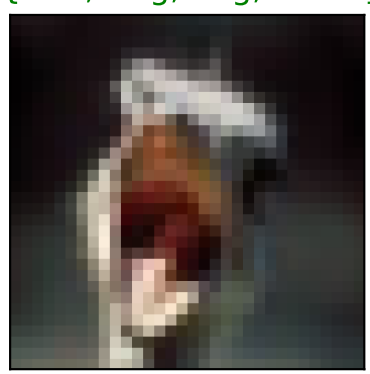
True label: bird
Prediction set:
{frog, bird, cat}

True label: horse
Prediction set:
{horse, cat, dog}



True label: frog
Prediction set:
{frog, bird, cat, deer}

True label: bird
Prediction set:
{bird, frog, dog, horse}



Some Relevant Prior Works on UQ for Classification

- Set-valued classification: Grycko (1993)
- Classification with reject option: Chow (1970), Ripley (1996), Herbei and Wegkamp (2006), Bartlett et al. (2008)

Conformal prediction for classification:

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\right] \geq 1 - \alpha$$

(Guarantees coverage for prediction sets with probability $1 - \alpha$.)

- Vovk et al. (2005)
- Nouretdinov et al. (2011)
- Papadopoulos (2014)
- Sadinle et al. (2018)
- Cauchois et al. (2020)
- Romano et al. (2020)
- Angelopoulos et al. (2021)
- Einbinder et al. (2022)

Some Relevant Prior Works on UQ for Classification

- Set-valued classification: Grycko (1993)
- Classification with reject option: Chow (1970), Ripley (1996), Herbei and Wegkamp (2006), Bartlett et al. (2008)

Conformal prediction for classification:

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\right] \geq 1 - \alpha$$

(Guarantees coverage for prediction sets with probability $1 - \alpha$.)

- Vovk et al. (2005)
- Nouretdinov et al. (2011)
- Papadopoulos (2014)
- Sadinle et al. (2018)
- Cauchois et al. (2020)
- Romano et al. (2020)
- Angelopoulos et al. (2021)
- Einbinder et al. (2022)

Typical assumption: exchangeable (or i.i.d.) data

Conformal Inference Beyond Exchangeability

Some of the key topics:

- **Testing Exchangeability**: Vovk (2021), Bates et al. (2021)
- **Time Series**: Chernozhukov et al. (2018), Xu and Xie (2021), Stankeviciute et al. (2021), Gibbs and Candès (2021, 2024), Angelopoulos et al. (2024), Zhou et al. (2024)
- **Covariate Shift**: Tibshirani et al. (2019)
- **Label Shift**: Podkopaev and Ramdas (2021), Si et al. (2023)
- **General Robustness**: Barber et al. (2022)
- **Other Applications**: S. et al. (2023), Liang et al. (2024)

Conformal Inference Beyond Exchangeability

Some of the key topics:

- **Testing Exchangeability**: Vovk (2021), Bates et al. (2021)
- **Time Series**: Chernozhukov et al. (2018), Xu and Xie (2021), Stankeviciute et al. (2021), Gibbs and Candès (2021, 2024), Angelopoulos et al. (2024), Zhou et al. (2024)
- **Covariate Shift**: Tibshirani et al. (2019)
- **Label Shift**: Podkopaev and Ramdas (2021), Si et al. (2023)
- **General Robustness**: Barber et al. (2022)
- **Other Applications**: S. et al. (2023), Liang et al. (2024)

Conformal prediction with "imperfect" data:

- **Missing Counterfactuals**: Lei et al. (2021), Yin et al. (2024)
- **Censoring**: Candès et al. (2023), Gui et al. (2024)
- **Missing Covariates**: Zaffran et al. (2023)
- **Weak Supervision**: Cauchois et al. (2022)

Conformal Inference Beyond Exchangeability

Some of the key topics:

- **Testing Exchangeability**: Vovk (2021), Bates et al. (2021)
- **Time Series**: Chernozhukov et al. (2018), Xu and Xie (2021), Stankeviciute et al. (2021), Gibbs and Candès (2021, 2024), Angelopoulos et al. (2024), Zhou et al. (2024)
- **Covariate Shift**: Tibshirani et al. (2019)
- **Label Shift**: Podkopaev and Ramdas (2021), Si et al. (2023)
- **General Robustness**: Barber et al. (2022)
- **Other Applications**: S. et al. (2023), Liang et al. (2024)

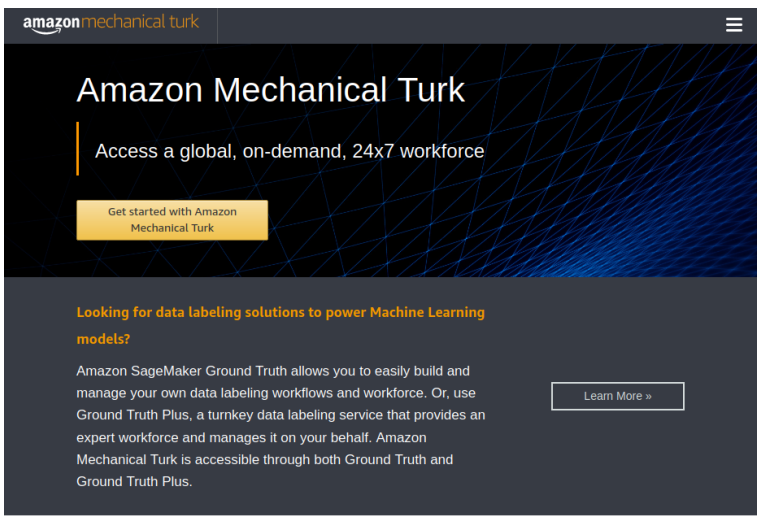
Conformal prediction with "imperfect" data:

- **Missing Counterfactuals**: Lei et al. (2021), Yin et al. (2024)
- **Censoring**: Candès et al. (2023), Gui et al. (2024)
- **Missing Covariates**: Zaffran et al. (2023)
- **Weak Supervision**: Cauchois et al. (2022)
- **Noisy Labels**: Einbinder et al. (2022), **and this talk . . .**

Why Care About Conformal Prediction with Noisy Labels?

Motivation: High-Quality Labels Aren't Always Available

- **High Cost of Labeling**:
  Manual labeling can be expensive and time-consuming
  (e.g., Snow et al. (2008); Aguinis et al. (2021))

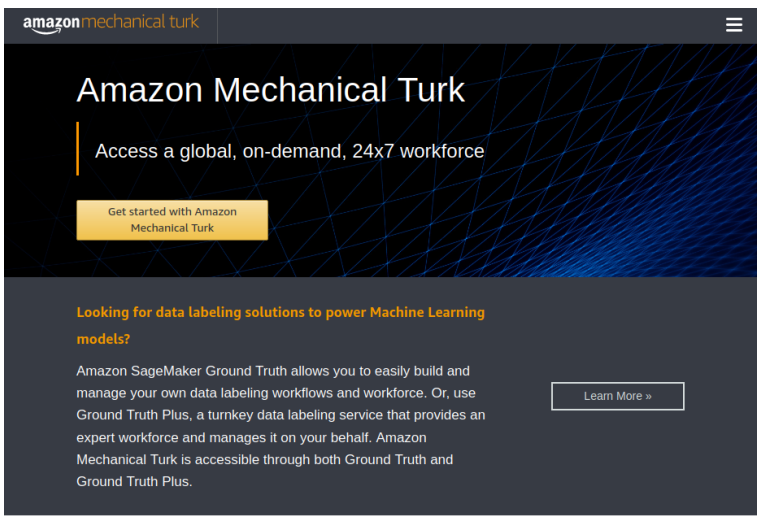


Crowdsourcing leads to noisy labels

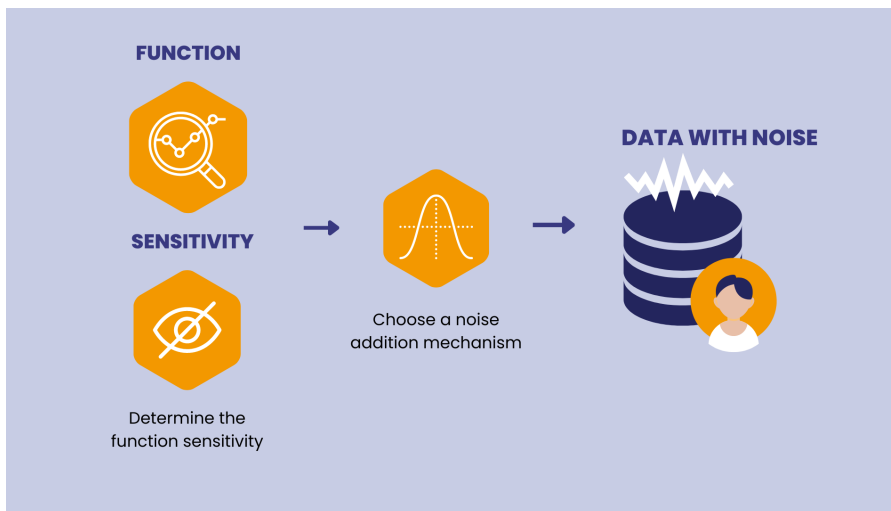
Why Care About Conformal Prediction with Noisy Labels?

Motivation: High-Quality Labels Aren't Always Available

- **High Cost of Labeling**:
  Manual labeling can be expensive and time-consuming
  (e.g., Snow et al. (2008); Aguinis et al. (2021))
- **Privacy Concerns**:
  Issues related to privacy may require *adding random noise* to the data
  (e.g., Ghazi et al. (2021))



Crowdsourcing leads to noisy labels



Adding noise to safeguard privacy

Background: Prediction from Imperfect Labels

Learning from data with noisy labels is a well-established field.

[Natarajan et al., 2013; Sukhbaatar et al., 2014; Song et al., 2022]

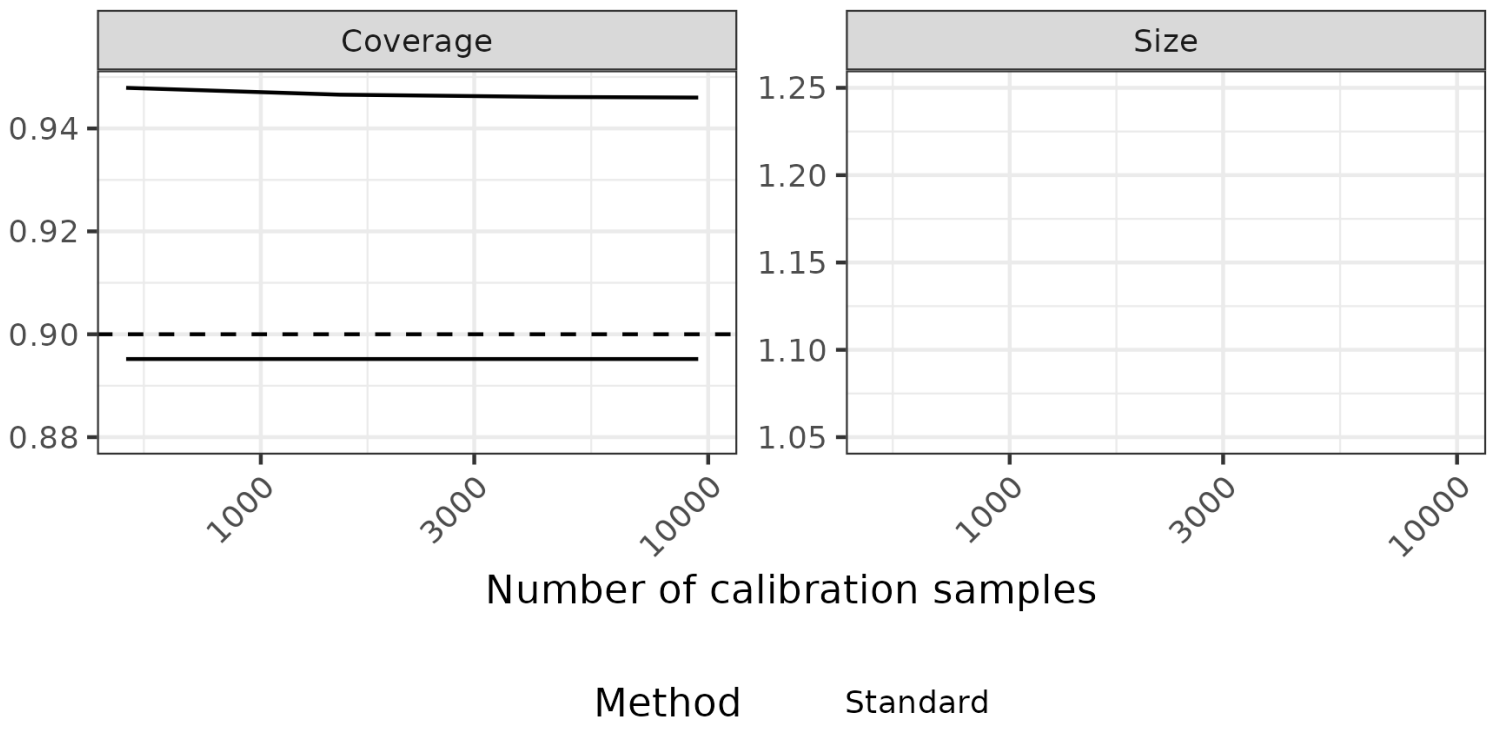
Background: Prediction from Imperfect Labels

Learning from data with noisy labels is a well-established field.

[Natarajan et al., 2013; Sukhbaatar et al., 2014; Song et al., 2022]

Robustness of conformal prediction to noisy labels:

- Conformal inference beyond exchangeability [Barber et al., 2023]
  (Worst-case bounds under Huber contamination model)



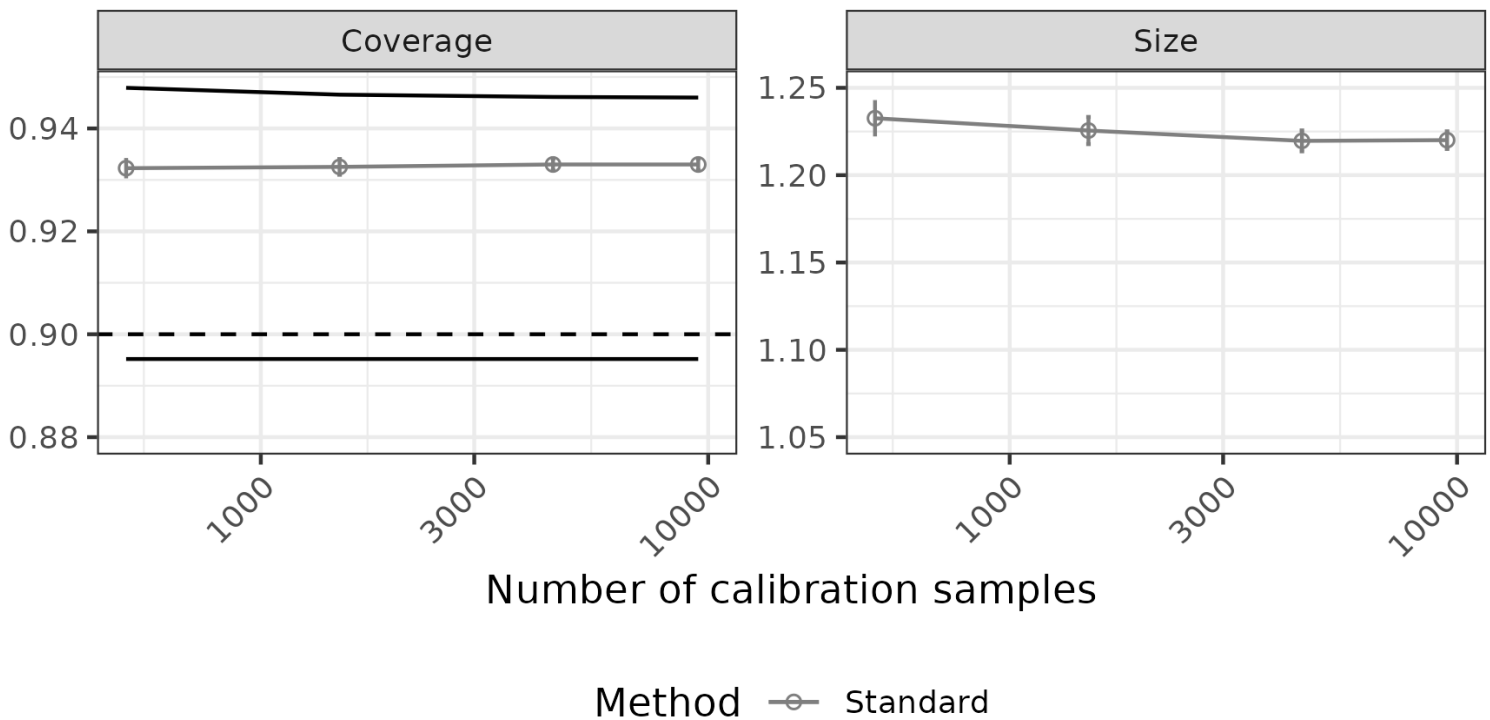
Background: Prediction from Imperfect Labels

Learning from data with noisy labels is a well-established field.

[Natarajan et al., 2013; Sukhbaatar et al., 2014; Song et al., 2022]

Robustness of conformal prediction to noisy labels:

- Conformal inference beyond exchangeability [Barber et al., 2023]
  (Worst-case bounds under Huber contamination model)



- We have some understanding of why conformal predictions tend to be "conservative" in practice. [Einbinder et al. (2022)]

Relation to Prior Works

The most closely related settings studied in the growing literature on conformal inference beyond exchangeability are:

- **Covariate Shift**: Different P_X, same $P_{Y|X}$
  Tibshirani et al. (2019)
- **Label Shift**: Different P_Y, same $P_{X|Y}$
  Podkopaev and Ramdas (2021); Si et al. (2023)
- **General Distribution Shifts**: Worst-case bounds for arbitrary shifts, including label contamination
  Barber et al. (2023)

Relation to Prior Works

The most closely related settings studied in the growing literature on conformal inference beyond exchangeability are:

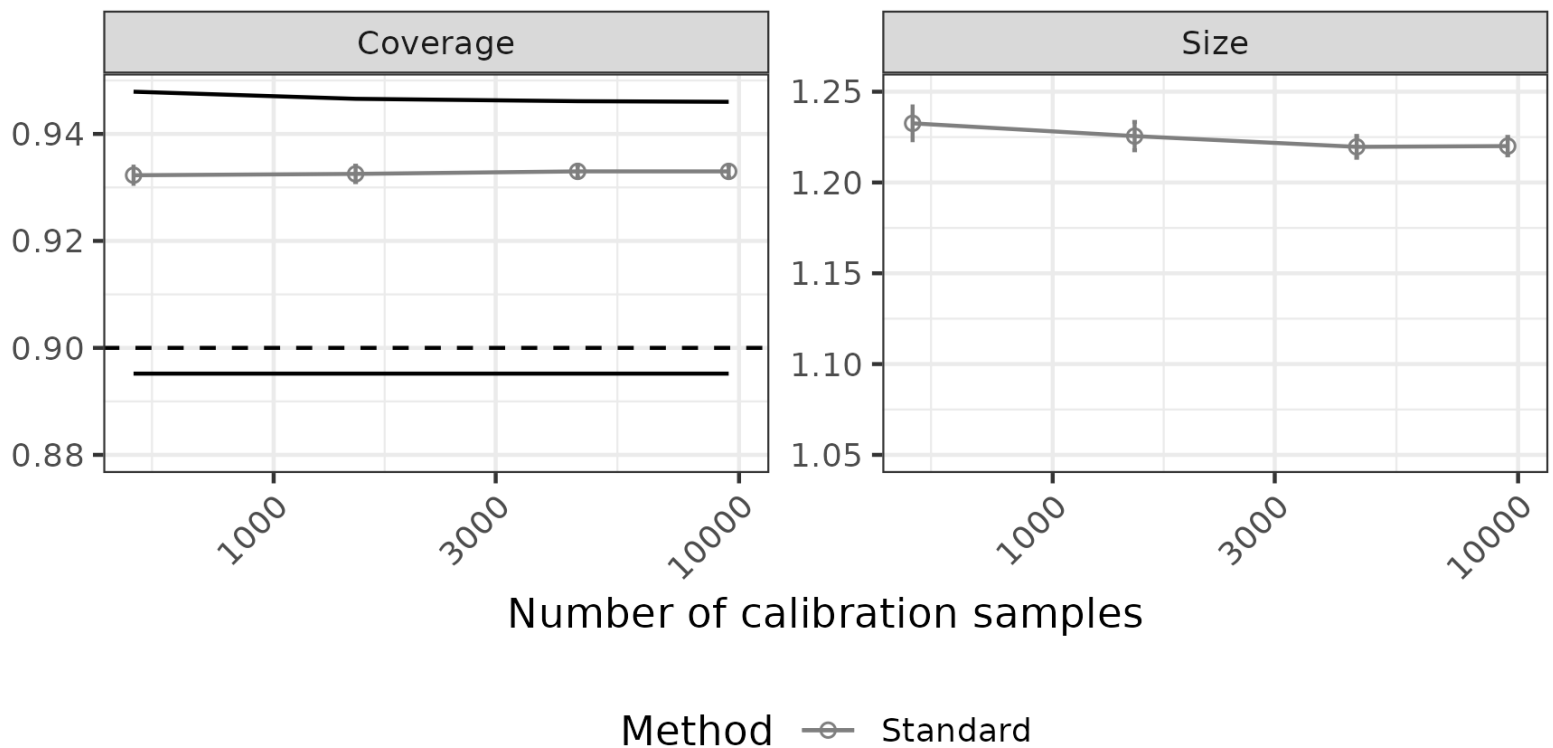
- **Covariate Shift**: Different P_X, same $P_{Y|X}$
  Tibshirani et al. (2019)
- **Label Shift**: Different P_Y, same $P_{X|Y}$
  Podkopaev and Ramdas (2021); Si et al. (2023)
- **General Distribution Shifts**: Worst-case bounds for arbitrary shifts, including label contamination
  Barber et al. (2023)

Our Approach:

- Closer to Barber et al. (2023), but **less pessimistic**.
- We focus on **methodological development**, making some additional assumptions on the label contamination process.
- We take a more general view than Einbinder et al. (2022), which did not consider how to "adapt" to label noise.

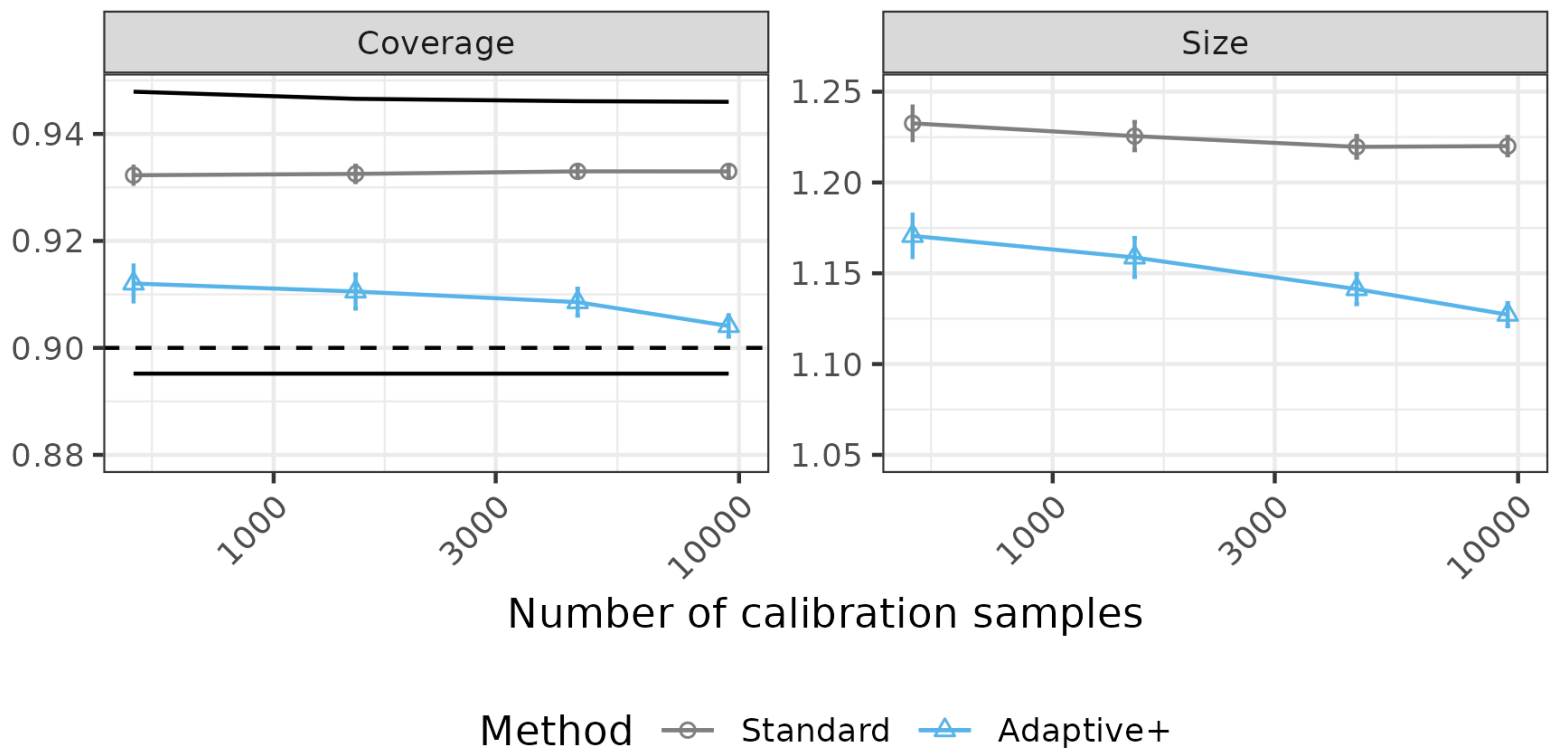
Preview of Contributions

1. We study the impacts of label noise on conformal predictions. Seeking more "actionable" insight compared to Barber et al. (2023), Einbinder et al. (2022).



Preview of Contributions

1. We study the impacts of label noise on conformal predictions. Seeking more "actionable" insight compared to Barber et al. (2023), Einbinder et al. (2022).
2. We develop an "adaptive" method to account for label noise.

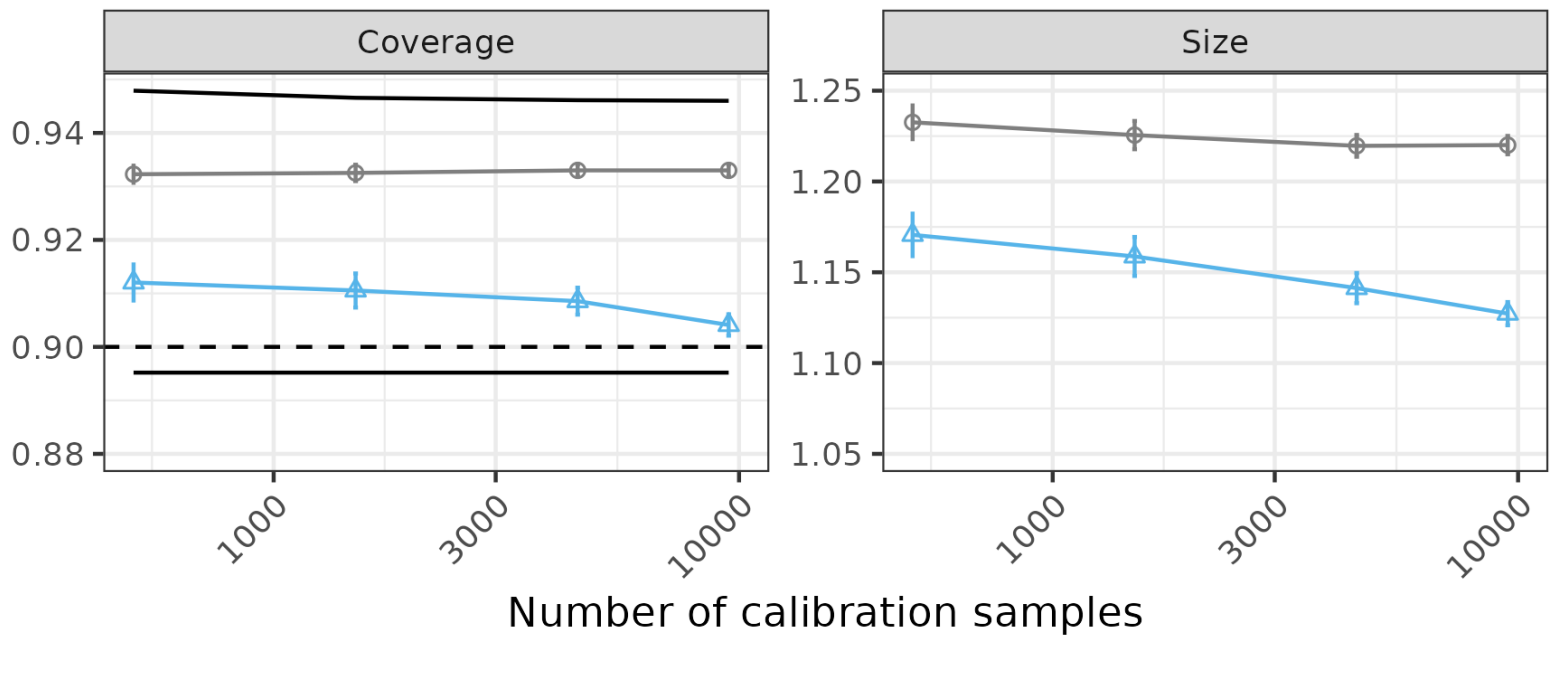


Preview of Contributions

1. We study the impacts of label noise on conformal predictions. Seeking more "actionable" insight compared to Barber et al. (2023), Einbinder et al. (2022).
2. We develop an "adaptive" method to account for label noise.

Note: the problem is **not straightforward**.

1. We see only data with "noisy" labels.
2. **We want to predict the "true" label of a new test point.**



Background: Conformal Classification

Conformal Classification [Vovk et al., 2005]

Data Setup:

- Observations: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n), (X_{n+1}, ?)$
- Features: $X \in \mathbb{R}^p$
- Label: $Y \in [K] := \{1, \ldots, K\}$

Key Assumption: $(X_i, Y_i) \overset{\text{iid}}{\sim} P$ for some unknown distribution P.

Conformal Classification [Vovk et al., 2005]

Data Setup:

- Observations: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n), (X_{n+1}, ?)$
- Features: $X \in \mathbb{R}^p$
- Label: $Y \in [K] := \{1, \ldots, K\}$

Key Assumption: $(X_i, Y_i) \overset{\text{iid}}{\sim} P$ for some unknown distribution P.

Typical Goal: prediction sets $\hat{C}_\alpha(X_{n+1})$ with *marginal coverage*:

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\right] \geq 1 - \alpha$$

at the desired confidence level $\alpha \in (0, 1)$; e.g., $\alpha = 0.1$.

Conformal Classification [Vovk et al., 2005]

Data Setup:

- Observations: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n), (X_{n+1}, ?)$
- Features: $X \in \mathbb{R}^p$
- Label: $Y \in [K] := \{1, \ldots, K\}$

Key Assumption: $(X_i, Y_i) \overset{\text{iid}}{\sim} P$ for some unknown distribution P.

Typical Goal: prediction sets $\hat{C}_\alpha(X_{n+1})$ with *marginal coverage*:

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\right] \geq 1 - \alpha$$

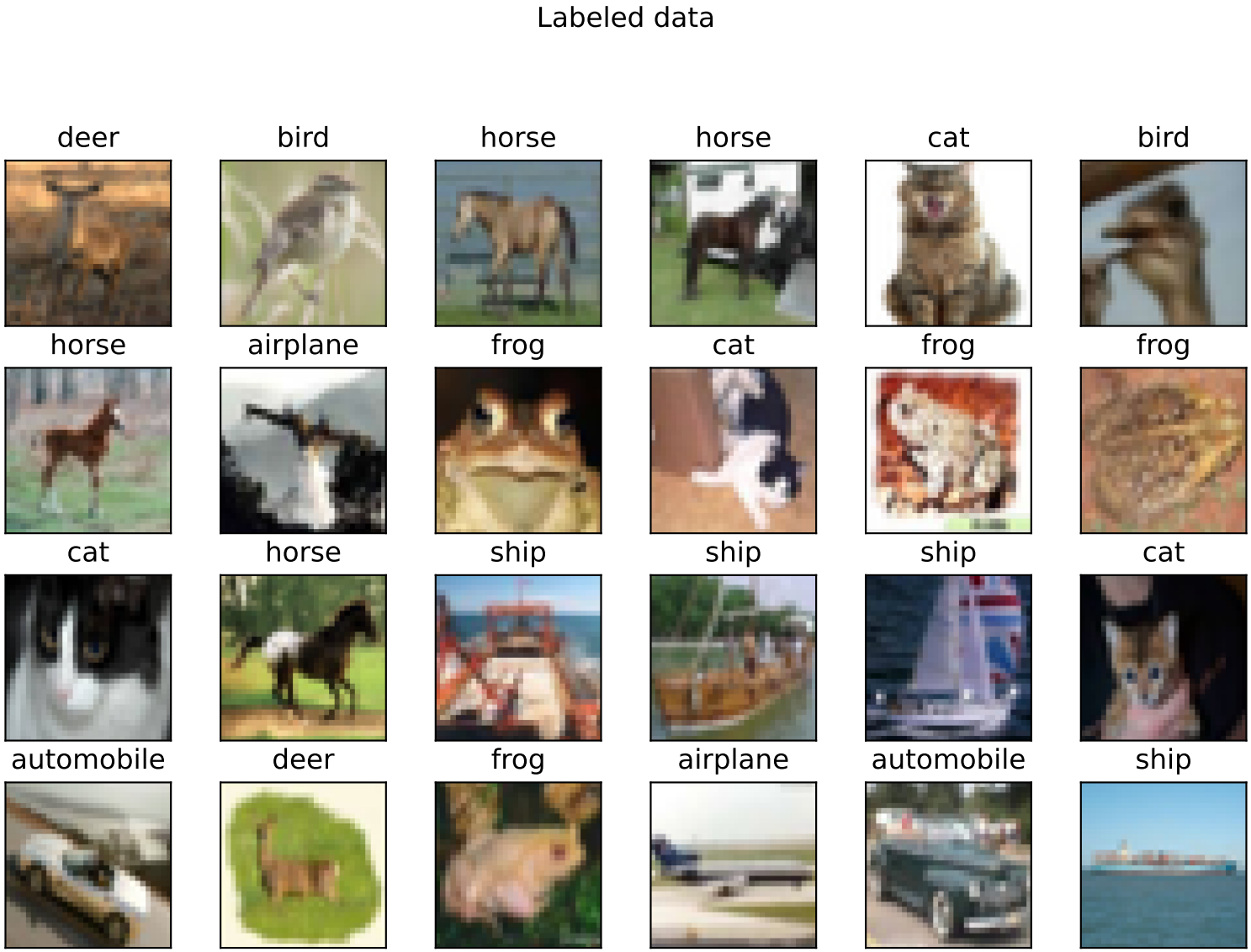
at the desired confidence level $\alpha \in (0, 1)$; e.g., $\alpha = 0.1$.

Upper Coverage Bound: Under smoothness conditions,

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\right] \leq 1 - \alpha + \mathcal{O}\left(\frac{1}{n}\right).$$

Overview of split-conformal classification[1]

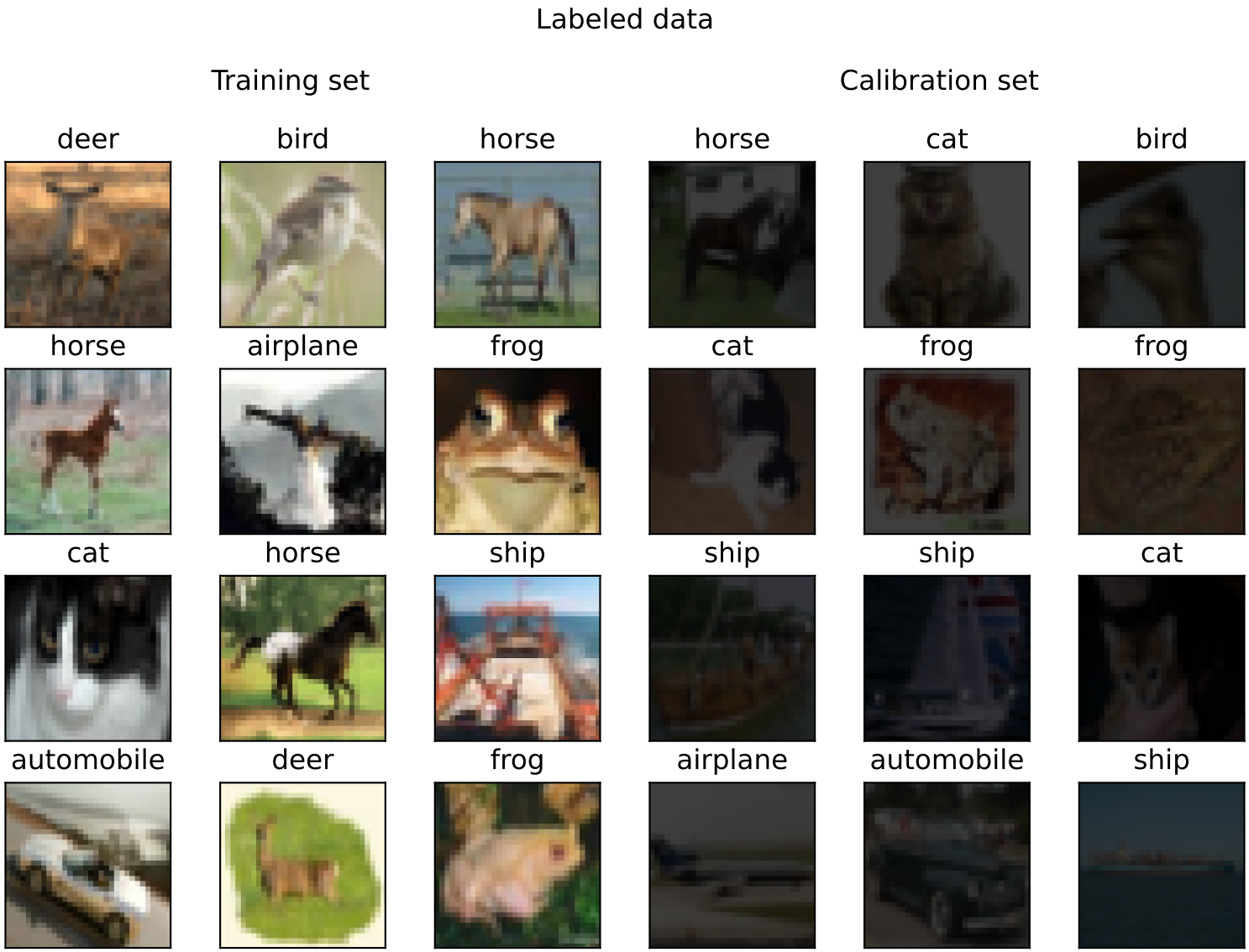
Idea: Split the data, *train* a "black-box" classifier, then *calibrate* it.



[1] Vovk et al. Algorithmic learning in a random world. Springer (2005).

Overview of split-conformal classification[1]

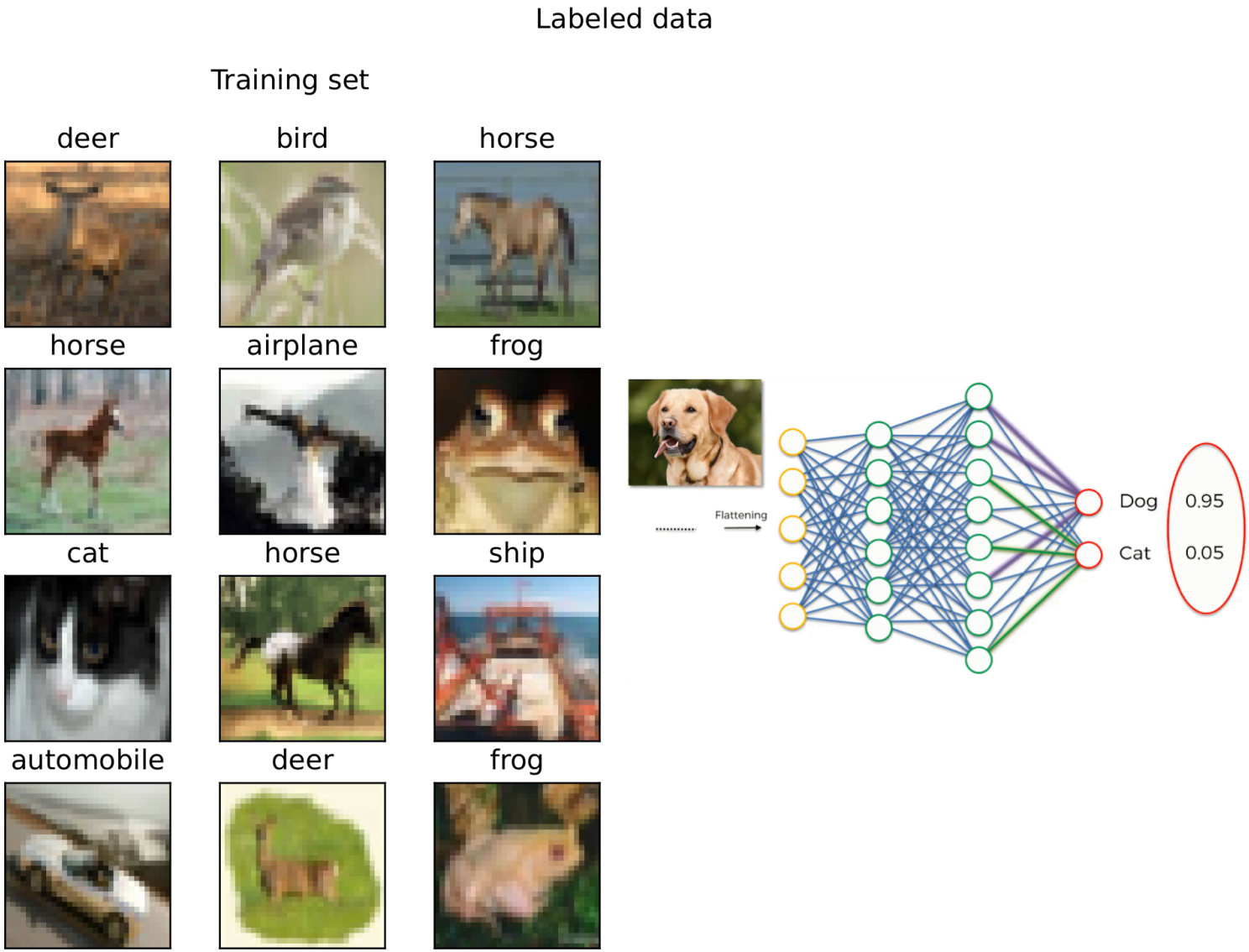
Idea: Split the data, *train* a "black-box" classifier, then *calibrate* it.



[1] Vovk et al. Algorithmic learning in a random world. Springer (2005).

Overview of split-conformal classification[1]

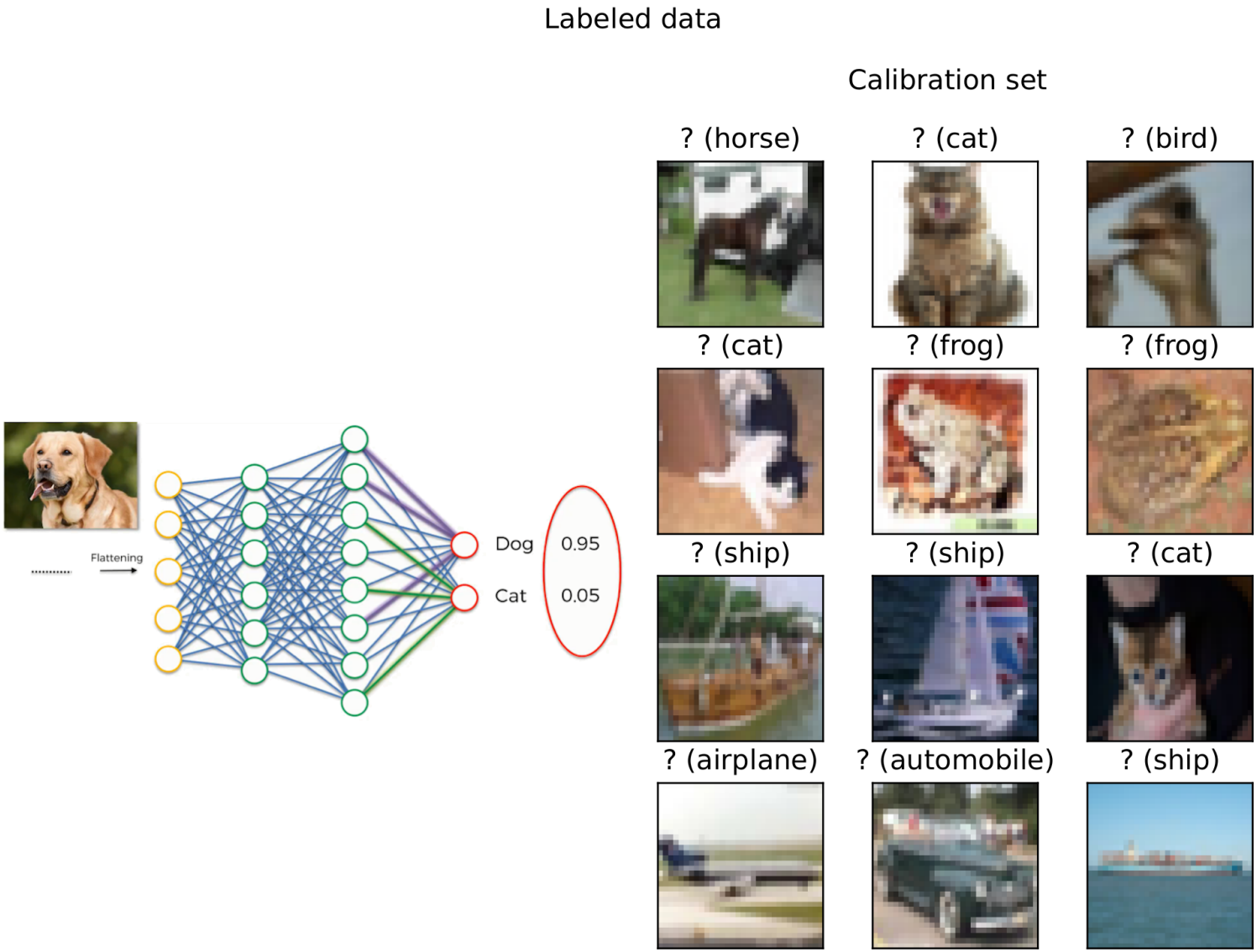
Idea: Split the data, *train* a "black-box" classifier, then *calibrate* it.



[1] Vovk et al. Algorithmic learning in a random world. Springer (2005).

Overview of split-conformal classification[1]

Idea: Split the data, *train* a "black-box" classifier, then *calibrate* it.



[1] Vovk et al. Algorithmic learning in a random world. Springer (2005).

1) Data Splitting and Model Training

Random Data Splitting

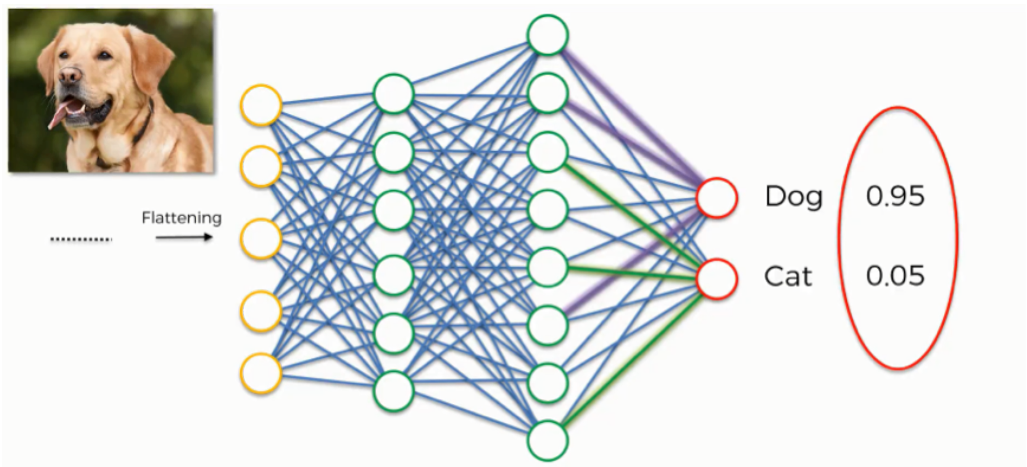
- Randomly split the data into two sets: $\mathcal{D}^{\text{train}}$ and $\mathcal{D}^{\text{cal}}$.

Model Training

- Train a classifier on $\mathcal{D}^{\text{train}}$ to estimate class probabilities:

$$\hat{\pi}_k(x) \approx \mathbb{P}[Y = k \mid X = x] \in [0, 1].$$

- Example: final soft-max layer of a deep neural network.



No Assumptions Required:

- No assumptions are needed about the accuracy of $\hat{\pi}$.
- The model may be trained using noisy labels $\tilde{Y}$.

2) Translating the fitted model into prediction sets

A classical example:

$$\mathcal{C}(x, \tau; \hat{\pi}) := \{k \in [K] : \hat{\pi}_k(x) \geq 1 - \tau\},$$

for some "threshold" $\tau \in [0, 1]$.

[2] Romano, S., and Candès. Adv. Neural Inf. Process. Syst. 33 (2020).

2) Translating the fitted model into prediction sets

A classical example:

$$\mathcal{C}(x, \tau; \hat{\pi}) := \{k \in [K] : \hat{\pi}_k(x) \geq 1 - \tau\},$$

for some "threshold" $\tau \in [0, 1]$.

In practice, we prefer to apply a slightly more complicated function $\mathcal{C}$ that can account for possible "heteroschedasticity".[2]

[2] Romano, S., and Candès. Adv. Neural Inf. Process. Syst. 33 (2020).

2) Translating the fitted model into prediction sets

A classical example:

$$\mathcal{C}(x, \tau; \hat{\pi}) := \{k \in [K] : \hat{\pi}_k(x) \geq 1 - \tau\},$$

for some "threshold" $\tau \in [0, 1]$.

In practice, we prefer to apply a slightly more complicated function $\mathcal{C}$ that can account for possible "heteroschedasticity".[2]

Definition (Prediction function)

Let $\mathcal{C}$ be a set-valued function (depending on $\hat{\pi}$), taking as input:

- $x \in \mathbb{R}^d$,
- $\tau \in [0, 1]$.

We say that $\mathcal{C}$ is a prediction function if:

1. $\mathbb{1}[k \in \mathcal{C}(X, \tau)]$ is increasing in τ.
2. $\mathcal{C}(X, \tau) = \{1, \ldots, K\}$ if $\tau = 1$.

[2] Romano, S., and Candès. Adv. Neural Inf. Process. Syst. 33 (2020).

3) Evaluating the Conformity Scores

Step 3: Computing Conformity Scores

- Use the held-out calibration samples $\mathcal{D}^{\text{cal}}$ to compute *conformity scores* $\hat{s}(X_i, k)$ for each $k \in [K]$ and $i \in \mathcal{D}^{\text{cal}}$.
- These scores can be interpreted as *generalized residuals*.

Classical Example:

$$\hat{s}(x, k) = 1 - \hat{\pi}_k(x).$$

General Case:

$$\hat{s}(x, k) = \inf \{\tau \in [0, 1] : k \in \mathcal{C}(x, \tau; \hat{\pi})\}.$$

4) Prediction Sets with Marginal Coverage

Calibrating the Prediction Sets:

1. Evaluate the conformity scores $\hat{s}(X_i, Y_i)$ for all $i \in \mathcal{D}^{\text{cal}}$.

4) Prediction Sets with Marginal Coverage

Calibrating the Prediction Sets:

1. Evaluate the conformity scores $\hat{s}(X_i, Y_i)$ for all $i \in \mathcal{D}^{\text{cal}}$.
2. Sort these scores and compute the threshold:

$$\hat{\tau}_0 = \lceil (1+|\mathcal{D}^{\text{cal}}|) \cdot (1-\alpha) \rceil \text{-th smallest value in } \{\hat{s}(X_i, Y_i)\}_{i \in \mathcal{D}^{\text{cal}}}.$$

4) Prediction Sets with Marginal Coverage

Calibrating the Prediction Sets:

1. Evaluate the conformity scores $\hat{s}(X_i, Y_i)$ for all $i \in \mathcal{D}^{\text{cal}}$.
2. Sort these scores and compute the threshold:

   $\hat{\tau}_0 = \lceil (1+|\mathcal{D}^{\text{cal}}|)\cdot(1-\alpha) \rceil$-th smallest value in $\{\hat{s}(X_i, Y_i)\}_{i \in \mathcal{D}^{\text{cal}}}$.
3. Construct the calibrated prediction set:

   $$\hat{C}(X_{n+1}) = \mathcal{C}(X_{n+1}, \hat{\tau}; \hat{\pi}),$$

   where $\hat{\tau} = (\hat{\tau}_0, \ldots, \hat{\tau}_0)$.

4) Prediction Sets with Marginal Coverage

Calibrating the Prediction Sets:

1. Evaluate the conformity scores $\hat{s}(X_i, Y_i)$ for all $i \in \mathcal{D}^{\text{cal}}$.
2. Sort these scores and compute the threshold:

   $\hat{\tau}_0 = \lceil (1+|\mathcal{D}^{\text{cal}}|)\cdot(1-\alpha) \rceil$-th smallest value in $\{\hat{s}(X_i, Y_i)\}_{i \in \mathcal{D}^{\text{cal}}}$.
3. Construct the calibrated prediction set:

   $$\hat{C}(X_{n+1}) = \mathcal{C}(X_{n+1}, \hat{\tau}; \hat{\pi}),$$

   where $\hat{\tau} = (\hat{\tau}_0, \ldots, \hat{\tau}_0)$.

Theoretical Guarantees:

- This procedure satisfies the *marginal coverage* condition:

  $$\mathbb{P}\left[Y_{n+1} \in \hat{C}(X_{n+1})\right] \geq 1 - \alpha.$$

- If the conformity scores are almost surely distinct:

  $$\mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\right] \leq 1 - \alpha + \frac{1}{|\mathcal{D}^{\text{cal}}|}.$$

Coverage Bounds Under Label Contamination

What if we cannot see the true labels?

Conformal Classification with Noisy Labels

Data: $(X_1, Y_1, \tilde{Y}_1), (X_2, Y_2, \tilde{Y}_2), \ldots, (X_n, Y_n, \tilde{Y}_n), (X_{n+1}, ?, ?)$.

- features $X \in \mathbb{R}^p$
- true label $Y \in [K] := \{1, \ldots, K\}$
- noisy label $\tilde{Y} \in [K] := \{1, \ldots, K\}$

Key assumption: $(X_i, Y_i, \tilde{Y}_i) \overset{\text{iid}}{\sim} P$, for some P.

Conformal Classification with Noisy Labels

Data: $(X_1, \quad, \tilde{Y}_1), (X_2, \quad, \tilde{Y}_2), \ldots, (X_n, \quad, \tilde{Y}_n), (X_{n+1}, ?, \ )$.

- features $X \in \mathbb{R}^p$
- true label $Y \in [K] := \{1, \ldots, K\}$
- noisy label $\tilde{Y} \in [K] := \{1, \ldots, K\}$

Key assumption: $(X_i, Y_i, \tilde{Y}_i) \overset{\text{iid}}{\sim} P$, for some P.

Question:

- Imagine applying standard conformal prediction based on the observed noisy labels $\tilde{Y}_1, \ldots, \tilde{Y}_n$.
- What happens to the coverage of Y_{n+1}?

Result: Coverage "Inflation" of Standard Prediction Sets

Theorem

For standard prediction sets $\hat{C}_\alpha$:

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\right] \geq 1 - \alpha + \mathbb{E}\left[\Delta(\hat{\tau})\right].$$

If the scores $s(X_i, \tilde{Y}_i)$ are almost surely distinct,

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\right] \leq 1 - \alpha + \frac{1}{n_{\text{cal}} + 1} + \mathbb{E}\left[\Delta(\hat{\tau})\right].$$

Coverage Inflation Factor:

$$\Delta(t) := \sum_{k=1}^{K} \left[\rho_k F_k^k(t) - \tilde{\rho}_k \tilde{F}_k^k(t)\right],$$

$$F_l^k(t) := \mathbb{P}\left[\hat{s}(X, k) \leq t \mid Y = l, \mathcal{D}^{\text{train}}\right], \quad \rho_k := \mathbb{P}[Y = k],$$
$$\tilde{F}_l^k(t) := \mathbb{P}\left[\hat{s}(X, k) \leq t \mid \tilde{Y} = l, \mathcal{D}^{\text{train}}\right], \quad \tilde{\rho}_k := \mathbb{P}[\tilde{Y} = k].$$

The Main Challenge: Estimating $\Delta(t)$

The Coverage Inflation Factor:

$$\Delta(t) := \sum_{k=1}^{K} \left[\rho_k F_k^k(t) - \tilde{\rho}_k \tilde{F}_k^k(t) \right]$$

The Difficulty:

- This factor depends on quantities that are hard to estimate, particularly:

$$F_l^k(t) := \mathbb{P}\left[\hat{s}(X, k) \leq t \mid Y = l, \mathcal{D}^{\text{train}} \right], \qquad \rho_k := \mathbb{P}[Y = k]$$

The Main Challenge: Estimating $\Delta(t)$

The Coverage Inflation Factor:

$$\Delta(t) := \sum_{k=1}^{K} \left[\rho_k F_k^k(t) - \tilde{\rho}_k \tilde{F}_k^k(t) \right]$$

The Difficulty:

- This factor depends on quantities that are hard to estimate, particularly:

$$F_l^k(t) := \mathbb{P}\left[\hat{s}(X,k) \leq t \mid Y = l, \mathcal{D}^{\text{train}}\right], \qquad \rho_k := \mathbb{P}[Y = k]$$

Assumption Needed: To make progress, we need to impose assumptions about the distribution of noisy labels.
A common starting point is:

Assumption (Conditional Independence of Label Noise)

$\tilde{Y} \perp\!\!\!\perp X \mid Y.$

Conservativeness of Standard Conformal Predictions

Corollary

Consider the setting of Theorem 1, assuming Assumption 1 holds. Additionally, suppose the CDFs of the conformity scores satisfy:

$$\max_{l \neq k} F_k^l(t) \leq F_k^k(t), \quad \forall t \in \mathbb{R},\, k \in [K].$$

Then, we have:

- $\Delta(\hat{\tau}) \geq 0$ **almost surely**.
- *The standard prediction sets* $\hat{C}_\alpha(X_{n+1})$ *are conservative:*

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\right] \geq 1 - \alpha.$$

Intuition: The condition ensures that the correct label is the most likely point prediction output by the model.

Where We Are

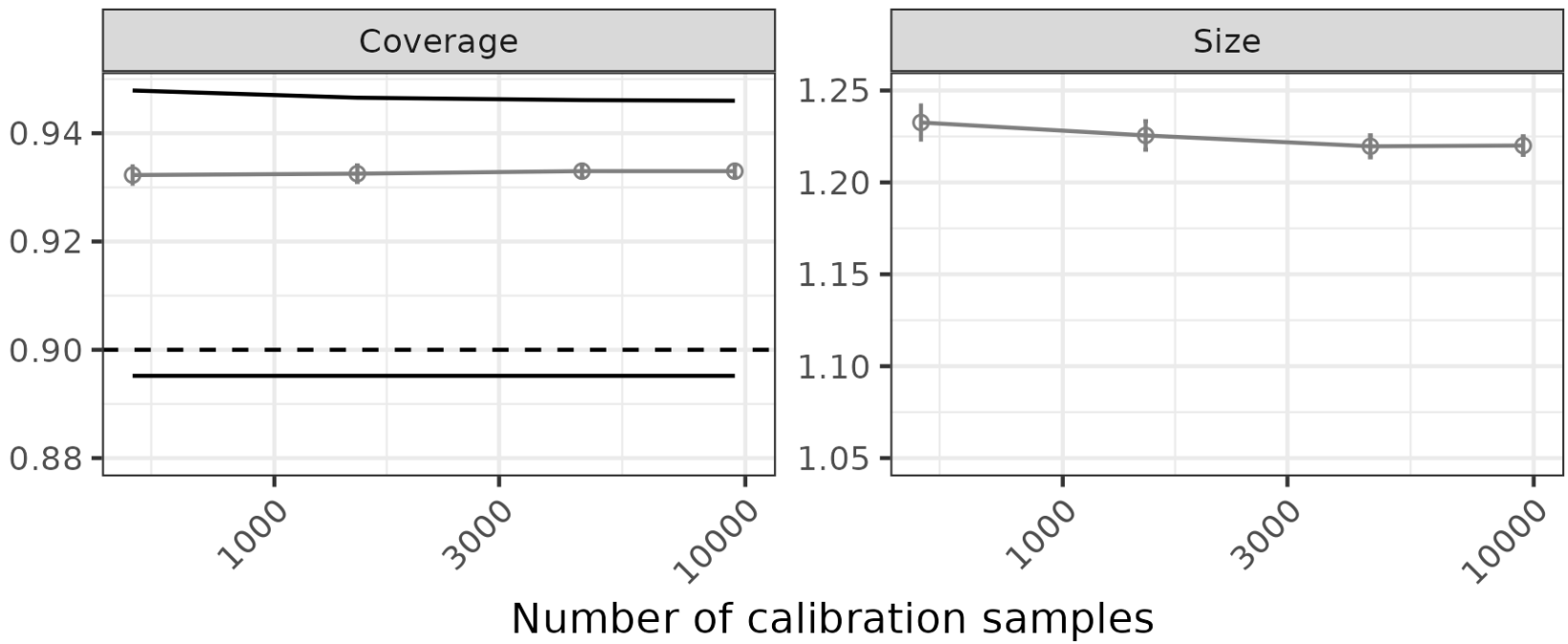
Main Challenge: Label noise often makes conformal predictions too conservative.

Our Goals:

1. **Quantify the conservativeness precisely.**
2. **Develop new adaptive methods to correct for this issue.**

Demonstration: CIFAR-10H dataset with pre-trained CNN model:

- Calibration with noisy data
- Prediction of the true labels



Adaptive prediction sets

How to correct for label noise?

Modeling the Label Noise

Assumptions:

Assumption (1)

$\tilde{Y} \perp\!\!\!\perp X \mid Y$

(The noisy label $\tilde{Y}$ is conditionally independent of the features X, given the true label Y.)

Modeling the Label Noise

Assumptions:

Assumption (1)

$\tilde{Y} \perp\!\!\!\perp X \mid Y$

(The noisy label $\tilde{Y}$ is conditionally independent of the features X, given the true label Y.)

Assumption (2)

The transition matrix $T \in [0,1]^{K \times K}$ is known, where:

$$T_{kl} := \mathbb{P}\left[\tilde{Y} = k \mid X, Y = l\right]$$

(This matrix models the probability of observing noisy label k when the true label is l.)

Plug-in Estimate for the Coverage Inflation Factor

Setup: Assume the transition matrix T is known and invertible, with inverse denoted by $W := T^{-1}$. Under Assumption 1, the coverage inflation factor $\Delta(t)$ can be written as:

$$\Delta(t) := \sum_{k=1}^{K} \sum_{l=1}^{K} W_{kl} \tilde{\rho}_l \tilde{F}_l^k(t) - \tilde{F}(t).$$

Empirical Estimates: Evaluated on calibration data:

$$\hat{F}_l^k(t) := \frac{1}{n_l} \sum_{i \in \mathcal{D}_l^{\text{cal}}} \mathbb{I}[\hat{s}(X_i, k) \leq t], \qquad \hat{\rho}_l := \frac{n_l}{n},$$

$$\hat{F}(t) := \frac{1}{n} \sum_{i \in \mathcal{D}^{\text{cal}}} \mathbb{I}[\hat{s}(X_i, \tilde{Y}_i) \leq t],$$

where $\mathcal{D}_l^{\text{cal}} = \{i \in \mathcal{D}^{\text{cal}} : \tilde{Y}_i = l\}$ and $n_l = |\mathcal{D}_l^{\text{cal}}|$.

A Novel Adaptive Calibration Algorithm

Intuition: The coverage is approximately $1-\alpha+\hat{\Delta}(\hat{\tau}(\alpha))$.
To reach $1-\alpha$, adjust the nominal level α by adding $\hat{\Delta}(\hat{\tau}(\alpha))$.

Prediction Set: $\hat{C}_\alpha(X_{n+1}) = \mathcal{C}(X, \hat{\tau}(\alpha))$.
Goal: How do we calibrate $\hat{\tau}(\alpha)$ to achieve $1-\alpha$ coverage?

A Novel Adaptive Calibration Algorithm

Intuition: The coverage is approximately $1-\alpha+\hat{\Delta}(\hat{\tau}(\alpha))$.
To reach $1-\alpha$, adjust the nominal level α by adding $\hat{\Delta}(\hat{\tau}(\alpha))$.

Prediction Set: $\hat{C}_\alpha(X_{n+1}) = \mathcal{C}(X, \hat{\tau}(\alpha))$.
Goal: How do we calibrate $\hat{\tau}(\alpha)$ to achieve $1-\alpha$ coverage?

Order Statistics: $S_{(i)}$ are the order statistics of $\{\hat{s}(X_j, \tilde{Y}_j)\}_{j\in\mathcal{D}^{\text{cal}}}$.

Standard Threshold: $\hat{\tau}(\alpha)$ is approximately given by:

$$\hat{\tau} = \begin{cases} S_{(\hat{i})}, \text{ where } \hat{i} = \min\{i \in \hat{\mathcal{I}}\}, & \text{if } \hat{\mathcal{I}} \neq \emptyset, \\ 1, & \text{if } \hat{\mathcal{I}} = \emptyset \end{cases}$$

$$\hat{\mathcal{I}} := \left\{ i \in [n] : \frac{i}{n} \geq 1-\alpha \right\}$$

A Novel Adaptive Calibration Algorithm

Intuition: The coverage is approximately $1-\alpha+\hat{\Delta}(\hat{\tau}(\alpha))$.
To reach $1-\alpha$, adjust the nominal level α by adding $\hat{\Delta}(\hat{\tau}(\alpha))$.

Prediction Set: $\hat{C}_\alpha(X_{n+1}) = \mathcal{C}(X, \hat{\tau}(\alpha))$.
Goal: How do we calibrate $\hat{\tau}(\alpha)$ to achieve $1-\alpha$ coverage?

Order Statistics: $S_{(i)}$ are the order statistics of $\{\hat{s}(X_j, \tilde{Y}_j)\}_{j\in\mathcal{D}^{\text{cal}}}$.

Intuition: Replace α with $\alpha+\hat{\Delta}(S_{(i)})$.

Standard Threshold: $\hat{\tau}(\alpha)$ is approximately given by:

$$\hat{\tau} = \begin{cases} S_{(\hat{i})}, \text{ where } \hat{i} = \min\{i \in \hat{\mathcal{I}}\}, & \text{if } \hat{\mathcal{I}} \neq \emptyset, \\ 1, & \text{if } \hat{\mathcal{I}} = \emptyset \end{cases}$$

$$\hat{\mathcal{I}} := \left\{ i \in [n] : \frac{i}{n} \geq 1-\alpha \right\}$$

A Novel Adaptive Calibration Algorithm

Intuition: The coverage is approximately $1-\alpha+\hat{\Delta}(\hat{\tau}(\alpha))$.
To reach $1-\alpha$, adjust the nominal level α by adding $\hat{\Delta}(\hat{\tau}(\alpha))$.

Prediction Set: $\hat{C}_\alpha(X_{n+1}) = \mathcal{C}(X, \hat{\tau}(\alpha))$.
Goal: How do we calibrate $\hat{\tau}(\alpha)$ to achieve $1-\alpha$ coverage?

Order Statistics: $S_{(i)}$ are the order statistics of $\{\hat{s}(X_j, \tilde{Y}_j)\}_{j\in\mathcal{D}^{\text{cal}}}$.

Intuition: Replace α with $\alpha+\hat{\Delta}(S_{(i)})$.

Adaptive Threshold: Our adaptive threshold $\hat{\tau}$ is given by:

$$\hat{\tau} = \begin{cases} S_{(\hat{i})}, \text{ where } \hat{i} = \min\{i \in \hat{\mathcal{I}}\}, & \text{if } \hat{\mathcal{I}} \neq \emptyset, \\ 1, & \text{if } \hat{\mathcal{I}} = \emptyset \end{cases}$$

$$\hat{\mathcal{I}} := \left\{ i \in [n] : \frac{i}{n} \geq 1-\alpha-\hat{\Delta}(S_{(i)}) \right\}$$

A Novel Adaptive Calibration Algorithm

Intuition: The coverage is approximately $1-\alpha+\hat{\Delta}(\hat{\tau}(\alpha))$.
To reach $1-\alpha$, adjust the nominal level α by adding $\hat{\Delta}(\hat{\tau}(\alpha))$.

Prediction Set: $\hat{C}_\alpha(X_{n+1}) = \mathcal{C}(X, \hat{\tau}(\alpha))$.
Goal: How do we calibrate $\hat{\tau}(\alpha)$ to achieve $1-\alpha$ coverage?

Order Statistics: $S_{(i)}$ are the order statistics of $\{\hat{s}(X_j, \tilde{Y}_j)\}_{j\in\mathcal{D}^{\text{cal}}}$.

Intuition: Replace α with $\alpha+\hat{\Delta}(S_{(i)})$.

Adaptive Threshold: Our adaptive threshold $\hat{\tau}$ is given by:

$$\hat{\tau} = \begin{cases} S_{(\hat{i})}, \text{ where } \hat{i} = \min\{i \in \hat{\mathcal{I}}\}, & \text{if } \hat{\mathcal{I}} \neq \emptyset, \\ 1, & \text{if } \hat{\mathcal{I}} = \emptyset \end{cases}$$

$$\hat{\mathcal{I}} := \left\{ i \in [n] : \frac{i}{n} \geq 1-\alpha-\hat{\Delta}(S_{(i)})+\delta(n) \right\}$$

- $\delta(n)$ is a finite-sample correction factor.

The Finite-Sample Correction Factor

Effect of Omitting the Correction: Suppose we apply our adaptive algorithm *without* the finite-sample correction factor:

$$\hat{\mathcal{I}} := \left\{ i \in [n] : \frac{i}{n} \geq 1 - \alpha - \hat{\Delta}(S_{(i)}) + \delta(n) \right\}$$

The Finite-Sample Correction Factor

Effect of Omitting the Correction: Suppose we apply our adaptive algorithm *without* the finite-sample correction factor:

$$\hat{\mathcal{I}} := \left\{ i \in [n] : \frac{i}{n} \geq 1 - \alpha - \hat{\Delta}(S_{(i)}) \right\}$$

Question: How low could the coverage be without the correction?

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_{\alpha}^{\text{ada}-0}(X_{n+1}) \right] \geq 1 - \alpha - \delta(n).$$

The Finite-Sample Correction Factor

Effect of Omitting the Correction: Suppose we apply our adaptive algorithm *without* the finite-sample correction factor:

$$\hat{\mathcal{I}} := \left\{ i \in [n] : \frac{i}{n} \geq 1 - \alpha - \hat{\Delta}(S_{(i)}) \right\}$$

Question: How low could the coverage be without the correction?

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_{\alpha}^{\text{ada}-0}(X_{n+1}) \right] \geq 1 - \alpha - \delta(n).$$

High-Level Result: In general, we can show that:

$$\delta(n) = \mathcal{O}\left(\frac{1}{\sqrt{n}} \right)$$

More Precisely: We will now make this precise with both

- finite-sample bounds;
- large-sample asymptotics.

Special Case: The Randomized Response Model

In the classical randomized response model (Warner, 1965), the finite-sample correction takes a simple form.

Simplified Model:

$$\mathbb{P}\left[\tilde{Y} = k \mid X, Y = l\right] = (1 - \epsilon)\mathbb{I}[k = l] + \frac{\epsilon}{K}.$$

Special Case: The Randomized Response Model

In the classical randomized response model (Warner, 1965), the finite-sample correction takes a simple form.

Simplified Model:

$$\mathbb{P}\left[\tilde{Y} = k \mid X, Y = l\right] = (1-\epsilon)\mathbb{I}[k = l] + \frac{\epsilon}{K}.$$

Result: Under this model, the finite-sample correction is given by:

$$\delta(n) = c(n) := \mathbb{E}\left[\sup_{i \in [n]} \left\{\frac{i}{n} - U_{(i)}\right\}\right],$$

where $U_{(1)}, \ldots, U_{(n)}$ are the order statistics of n i.i.d. uniform($[0, 1]$) random variables.

Special Case: The Randomized Response Model

In the classical randomized response model (Warner, 1965), the finite-sample correction takes a simple form.

Simplified Model:

$$\mathbb{P}\left[\tilde{Y} = k \mid X, Y = l\right] = (1-\epsilon)\mathbb{I}[k = l] + \frac{\epsilon}{K}.$$

Result: Under this model, the finite-sample correction is given by:

$$\delta(n) = c(n) := \mathbb{E}\left[\sup_{i \in [n]} \left\{\frac{i}{n} - U_{(i)}\right\}\right],$$

where $U_{(1)}, \ldots, U_{(n)}$ are the order statistics of n i.i.d. uniform$([0,1])$ random variables.

Key Properties:

- $c(n)$ can be accurately estimated via Monte Carlo simulation.
- By the DKWM inequality, $c(n) \leq \sqrt{\frac{\pi}{2n}}$.

General Case (Finite-Sample Empirical Process Theory)

In general, we must account for the structure of $W := T^{-1}$.

General Case (Finite-Sample Empirical Process Theory)

In general, we must account for the structure of $W := T^{-1}$.

Approach: We introduce the matrix $\bar{W}$, a $K \times K$ matrix parameterized by coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_K) \in \mathbb{R}^{K+1}$, where:

$$\bar{W}_{kl} = \beta_0 \mathbb{I}[k = l] + \frac{\beta_k}{K}.$$

Next, define the deviation matrix Ω as: $\Omega_{kl} := W_{kl} - \bar{W}_{kl}$.

General Case (Finite-Sample Empirical Process Theory)

In general, we must account for the structure of $W := T^{-1}$.

Approach: We introduce the matrix $\bar{W}$, a $K \times K$ matrix parameterized by coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_K) \in \mathbb{R}^{K+1}$, where:

$$\bar{W}_{kl} = \beta_0 \mathbb{I}[k = l] + \frac{\beta_k}{K}.$$

Next, define the deviation matrix Ω as: $\Omega_{kl} := W_{kl} - \bar{W}_{kl}$.

Finite-Sample Correction Factor: (from chaining arguments)

$$\delta(n) = \inf_{\beta} \left\{ c(n) \left(\beta_0 + \frac{\sum_{k=1}^{K} \beta_k}{K} \right) + \frac{1}{\sqrt{n}} B(K, n, \beta) \right\},$$

$$B(K, n, \beta) = 2 \min \left\{ \max_{l \in [K]} \sum_{k} |\Omega_{kl}| \sqrt{\log(Kn + 1)}, \right.$$
$$\left. 24 \max_{k,l \in [K]} |\Omega_{kl}| \frac{2 \log K + 1}{2 \log K - 1} \sqrt{2K \log K} \right\}.$$

Optimizing the Finite-Sample Constants

Goal: Optimally choose the parameters β to minimize the finite-sample correction factor $\delta(n)$:

$$\delta(n) = \inf_{\beta} \left\{ c(n) \left(\beta_0 + \frac{\sum_{k=1}^{K} \beta_k}{K} \right) + \frac{1}{\sqrt{n}} B(K, n, \beta) \right\}.$$

Approach: The optimal choice of β depends on the structure of $W := T^{-1}$. This problem can be solved using convex optimization.

Optimizing the Finite-Sample Constants

Goal: Optimally choose the parameters β to minimize the finite-sample correction factor $\delta(n)$:

$$\delta(n) = \inf_{\beta} \left\{ c(n) \left(\beta_0 + \frac{\sum_{k=1}^{K} \beta_k}{K} \right) + \frac{1}{\sqrt{n}} B(K, n, \beta) \right\}.$$

Approach: The optimal choice of β depends on the structure of $W := T^{-1}$. This problem can be solved using convex optimization.

"Worst-Case" Behavior:

$$\delta(n) \approx \min \left\{ \mathcal{O}\left(\frac{\sqrt{K \log K}}{\sqrt{n}} \right), \mathcal{O}\left(\frac{\sqrt{\log(Kn)}}{\sqrt{n}} \right) \right\}.$$

The optimization can improve the scaling with respect to K.

Optimizing the Finite-Sample Constants

Goal: Optimally choose the parameters β to minimize the finite-sample correction factor $\delta(n)$:

$$\delta(n) = \inf_{\beta} \left\{ c(n) \left(\beta_0 + \frac{\sum_{k=1}^{K} \beta_k}{K} \right) + \frac{1}{\sqrt{n}} B(K, n, \beta) \right\}.$$

Approach: The optimal choice of β depends on the structure of $W := T^{-1}$. This problem can be solved using convex optimization.

"Worst-Case" Behavior:

$$\delta(n) \approx \min \left\{ \mathcal{O}\left(\frac{\sqrt{K \log K}}{\sqrt{n}} \right), \mathcal{O}\left(\frac{\sqrt{\log(Kn)}}{\sqrt{n}} \right) \right\}.$$

The optimization can improve the scaling with respect to K.

Example: For a two-level randomized response model, we get:

$$\delta = \mathcal{O}\left(\frac{1}{\sqrt{n}} \left(1 + \sqrt{\frac{2 \log K}{K}} \right) \right).$$

Finite-Sample Coverage Guarantees

Theorem

Suppose $(X_i, Y_i, \tilde{Y}_i)$ *are i.i.d. for all* $i \in [n+1]$, *and* $\tilde{Y} \perp\!\!\!\perp X \mid Y$. *Let* $\hat{C}_\alpha(X_{n+1})$ *be our Adaptive prediction set based on the inverse* W *of the label noise model matrix* T, *using the finite-sample correction term* $\delta(n)$ *derived earlier. Then,*

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\right] \geq 1 - \alpha.$$

Finite-Sample Coverage Guarantees

Theorem

Suppose $(X_i, Y_i, \tilde{Y}_i)$ *are i.i.d. for all* $i \in [n+1]$, *and* $\tilde{Y} \perp\!\!\!\perp X \mid Y$. *Let* $\hat{C}_\alpha(X_{n+1})$ *be our Adaptive prediction set based on the inverse* W *of the label noise model matrix* T, *using the finite-sample correction term* $\delta(n)$ *derived earlier. Then,*

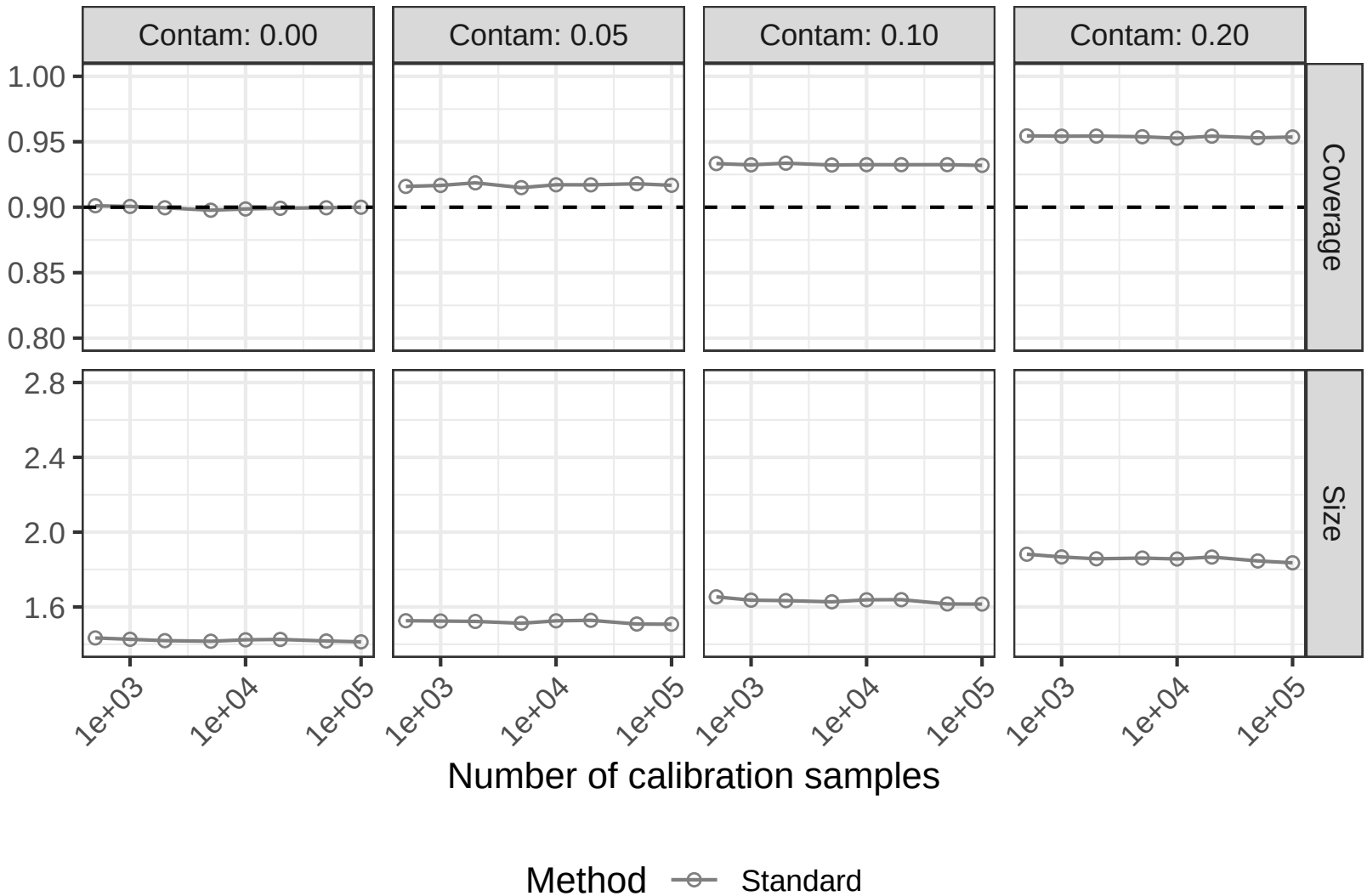
$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\right] \geq 1 - \alpha.$$

Further, under some (mild) additional technical conditions,

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\right] \leq 1 - \alpha + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

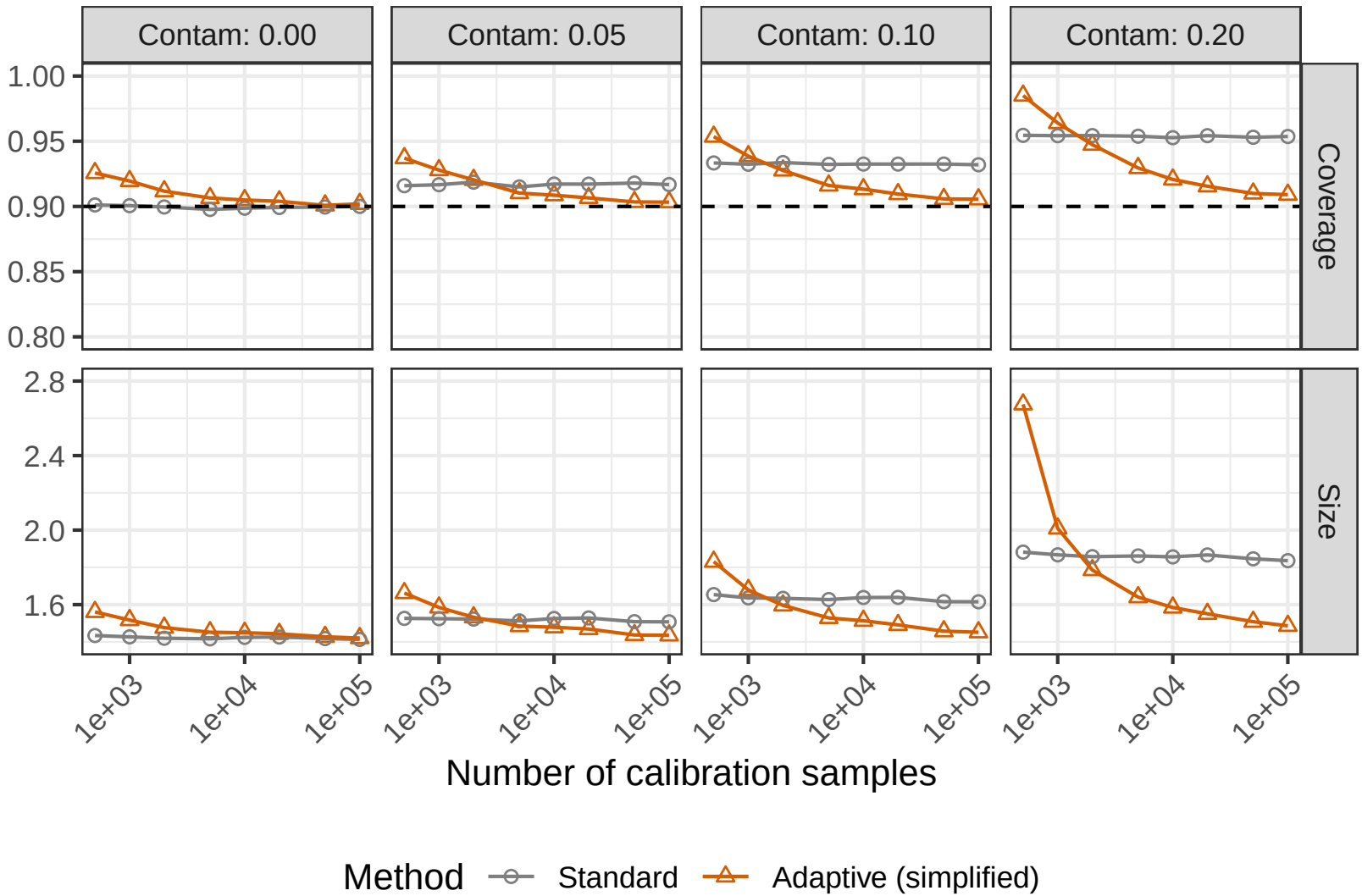
Empirical Performance on Simulated Data

- Synthetic data with 4 possible labels.
- Label noise model: two-level randomized response.
- Metrics: Average coverage and size of prediction sets.



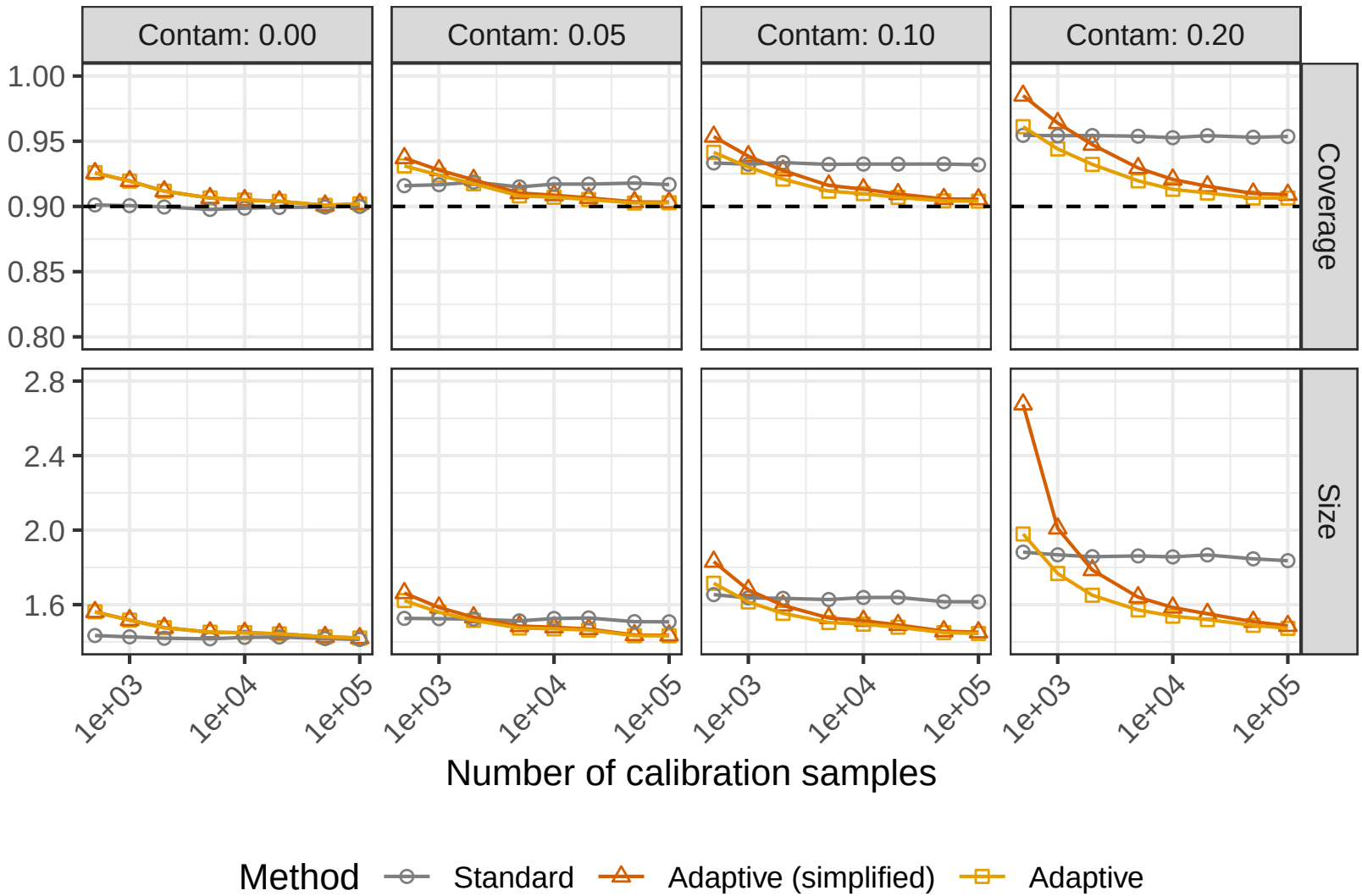
Empirical Performance on Simulated Data

- Synthetic data with 4 possible labels.
- Label noise model: two-level randomized response.
- Metrics: Average coverage and size of prediction sets.



Empirical Performance on Simulated Data

- Synthetic data with 4 possible labels.
- Label noise model: two-level randomized response.
- Metrics: Average coverage and size of prediction sets.



Boosting Power with More Optimistic Calibration

Standard conformal predictions are often conservative.

Why not take advantage of this conservativeness to boost power?

Boosting Power with More Optimistic Calibration

Standard conformal predictions are often conservative.
Why not take advantage of this conservativeness to boost power?

Optimistic Calibration: We propose an adjusted calibration rule:

$$\hat{\mathcal{I}} := \left\{ i \in [n] : \frac{i}{n} \geq 1 - \alpha - \max\left\{ \hat{\Delta}(S_{(i)}) - \delta(n), -\frac{1-\alpha}{n} \right\} \right\}.$$

Proposition

Under the setup of the previous theorem, assume also that

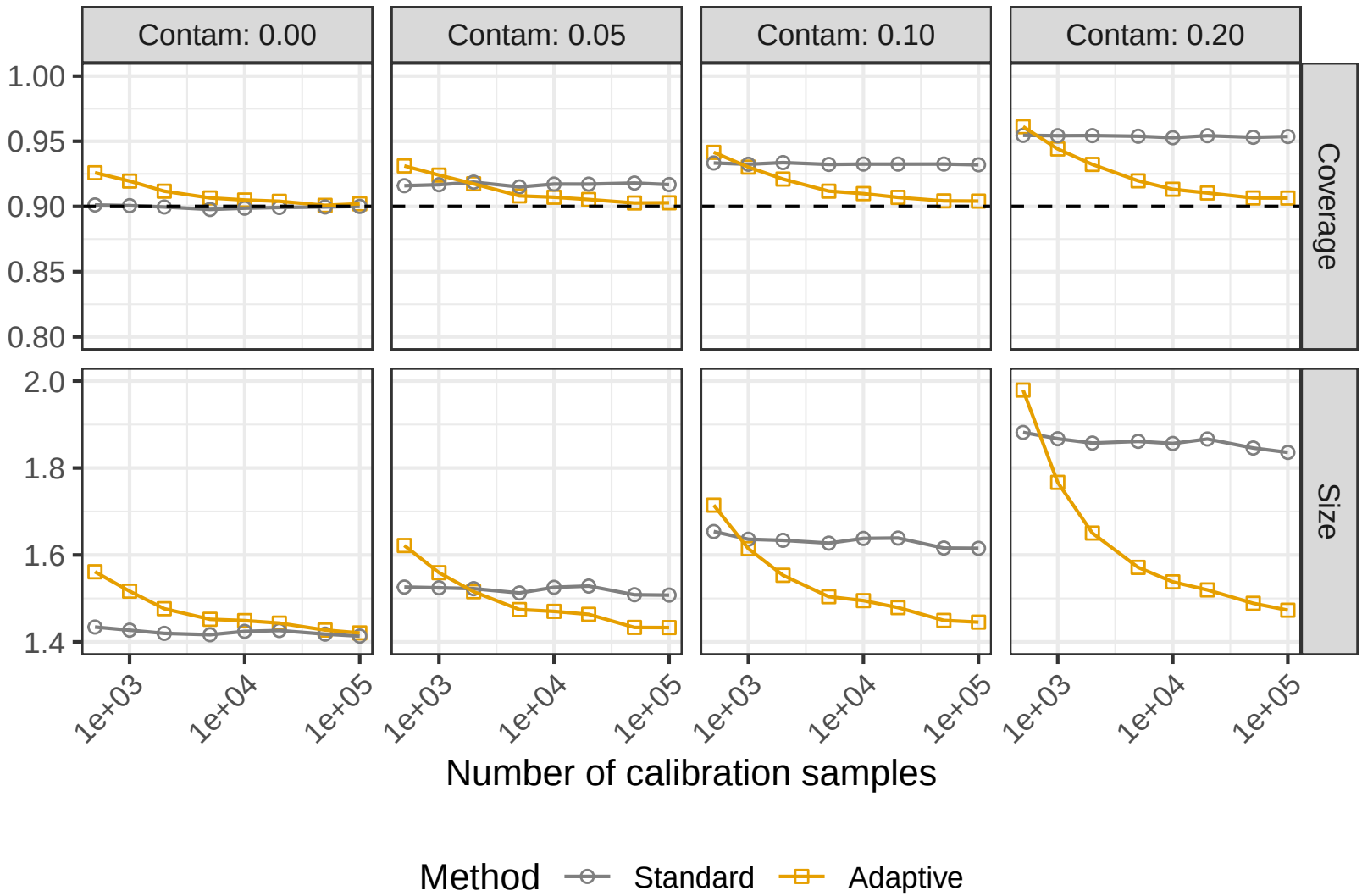
$$\inf_{t \in \mathbb{R}} \Delta(t) \geq \delta(n) - \frac{1-\alpha}{n}.$$

If $\hat{C}_\alpha(X_{n+1})$ is our Adaptive+ prediction set based on the above $\hat{\mathcal{I}}$, then $\hat{C}_\alpha(X_{n+1})$ satisfies

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1}) \right] \geq 1 - \alpha.$$

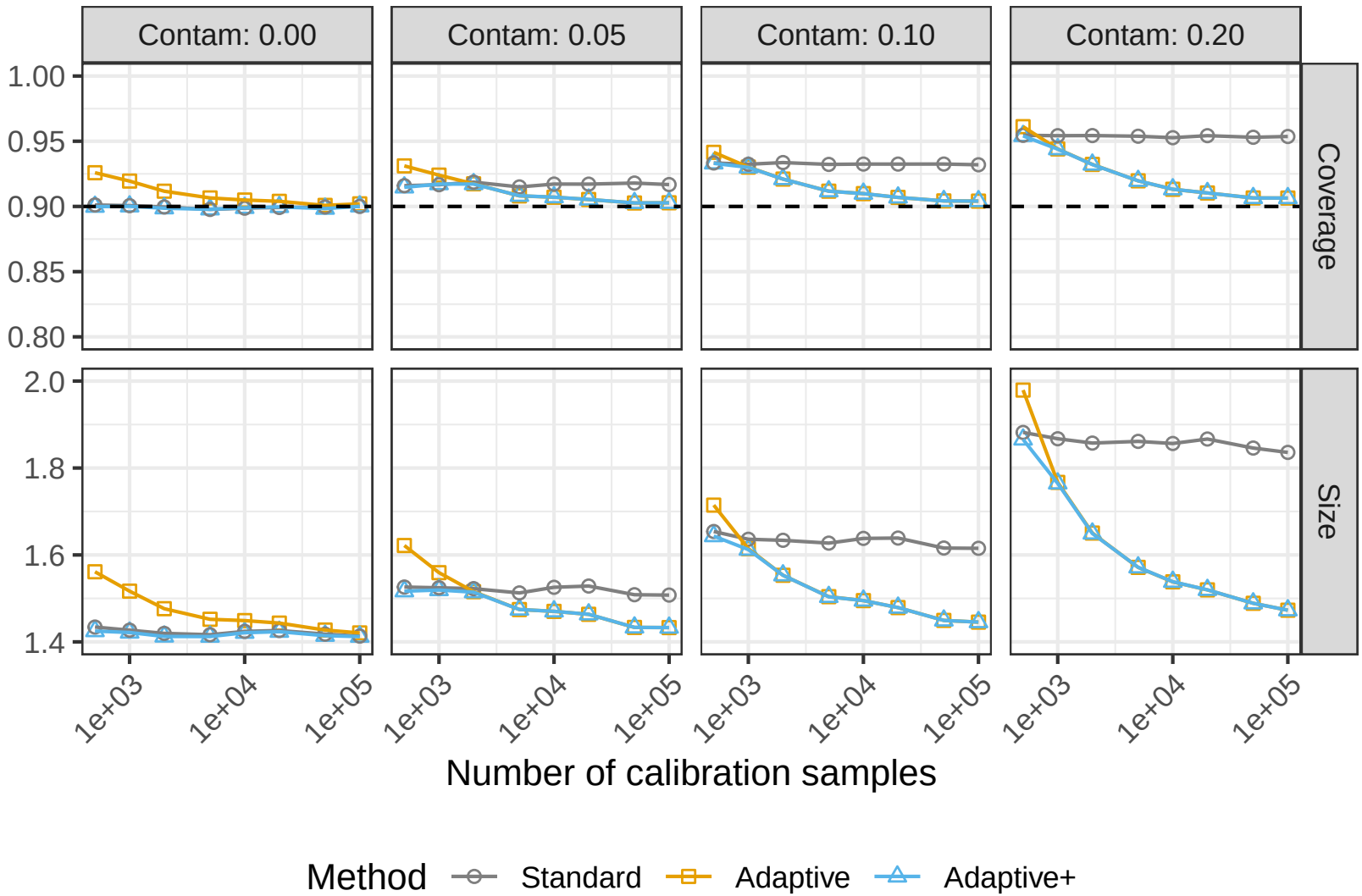
Empirical Performance on Simulated Data

- Synthetic data with 4 possible labels.
- Label noise model: two-level randomized response.
- Metrics: Average coverage and size of prediction sets.



Empirical Performance on Simulated Data

- Synthetic data with 4 possible labels.
- Label noise model: two-level randomized response.
- Metrics: Average coverage and size of prediction sets.



Large-Sample Asymptotics

Large-Sample Behavior: The large-sample behavior of our Adaptive calibration algorithm can be studied by applying Donsker's theorem to the following empirical process:

$$\hat{\Delta}(t) - \Delta(t) := \sum_{k=1}^{K} \sum_{l=1}^{K} W_{kl} \left[\hat{\rho}_l \hat{F}_l^k(t) - \tilde{\rho}_l \tilde{F}_l^k(t) \right].$$

Large-Sample Asymptotics

Large-Sample Behavior: The large-sample behavior of our Adaptive calibration algorithm can be studied by applying Donsker's theorem to the following empirical process:

$$\hat{\Delta}(t) - \Delta(t) := \sum_{k=1}^{K} \sum_{l=1}^{K} W_{kl} \left[\hat{\rho}_l \hat{F}_l^k(t) - \tilde{\rho}_l \tilde{F}_l^k(t) \right].$$

Key Result: In the large-n limit, this process converges to a Generalized Brownian Bridge, namely GBG(t), a *centered Gaussian process* with covariance function given by:

$$\mathrm{Cov}(t_1, t_2) = \mathbb{E}[g(t_1, t_2)] - \mathbb{E}[f(t_1)]\mathbb{E}[f(t_2)].$$

Large-Sample Asymptotics

Large-Sample Behavior: The large-sample behavior of our Adaptive calibration algorithm can be studied by applying Donsker's theorem to the following empirical process:

$$\hat{\Delta}(t) - \Delta(t) := \sum_{k=1}^{K} \sum_{l=1}^{K} W_{kl} \left[\hat{\rho}_l \hat{F}_l^k(t) - \tilde{\rho}_l \tilde{F}_l^k(t) \right].$$

Key Result: In the large-n limit, this process converges to a Generalized Brownian Bridge, namely GBG(t), a *centered Gaussian process* with covariance function given by:

$$\mathrm{Cov}(t_1, t_2) = \mathbb{E}[g(t_1, t_2)] - \mathbb{E}[f(t_1)]\mathbb{E}[f(t_2)].$$

Estimating the Covariance Function:

$$\hat{\mathbb{E}}[f(t)] = \sum_{k=1}^{K} \sum_{l=1}^{K} W_{kl} \hat{\rho}_l \frac{1}{n_l} \sum_{i \in \mathcal{D}_l} \mathbb{I}[\hat{s}(X_i, k) \leq t],$$

$$\hat{\mathbb{E}}[g(t_1, t_2)] = \sum_{k=1}^{K} \sum_{k'=1}^{K} \sum_{l=1}^{K} W_{kl} W_{k'l} \hat{\rho}_l \frac{1}{n_l} \sum_{i \in \mathcal{D}_l} \mathbb{I}[\hat{s}(X_i, k) \leq t_1, \hat{s}(X_i, k') \leq t_2].$$

The Asymptotic Correction Factor

Asymptotic Approximation: In the large-n limit, the correction factor $\delta(n)$ can be approximated as:

$$\delta^{\text{asymptotic}} = \hat{\mathbb{E}}\left[\sup_{t\in\mathbb{R}} \mathsf{GBG}(t)\right].$$

The Asymptotic Correction Factor

Asymptotic Approximation: In the large-n limit, the correction factor $\delta(n)$ can be approximated as:

$$\delta^{\text{asymptotic}} = \hat{\mathbb{E}}\left[\sup_{t\in\mathbb{R}} \mathrm{GBG}(t)\right].$$

Estimation Procedure:

- Replace the unknown covariance function of GBG with a plug-in empirical estimate.
- Simulate a discretized version of the Gaussian process for various step sizes h.
- Apply Richardson extrapolation to approximate the limit as $h \to 0$.

The Asymptotic Correction Factor

Asymptotic Approximation: In the large-n limit, the correction factor $\delta(n)$ can be approximated as:

$$\delta^{\text{asymptotic}} = \hat{\mathbb{E}}\left[\sup_{t\in\mathbb{R}} \mathrm{GBG}(t)\right].$$

Estimation Procedure:

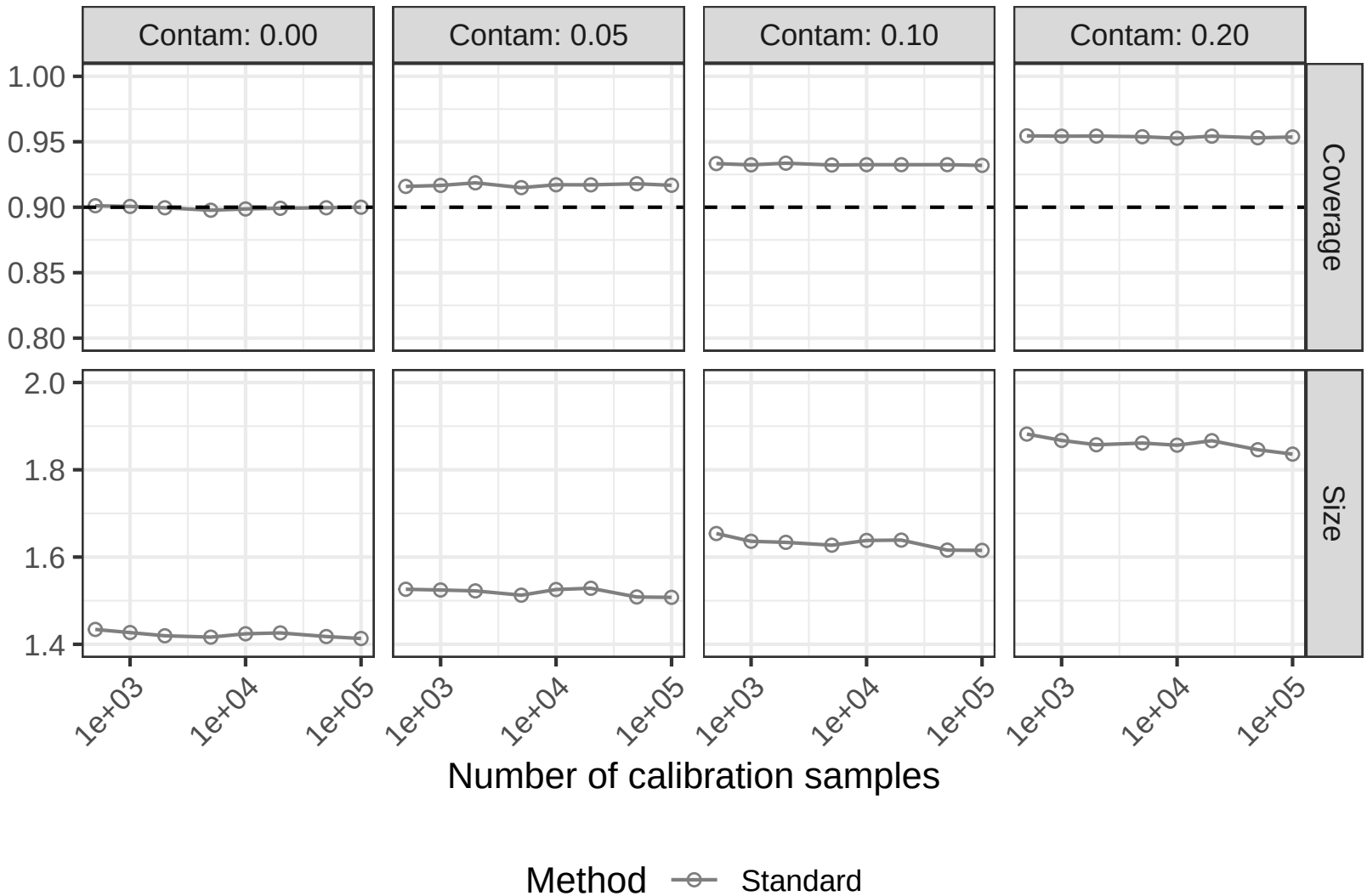
- Replace the unknown covariance function of GBG with a plug-in empirical estimate.
- Simulate a discretized version of the Gaussian process for various step sizes h.
- Apply Richardson extrapolation to approximate the limit as $h \to 0$.

Key Takeaways:

- It works well in practice.
- It agrees with the finite-sample approach in the special case of the randomized response model.
- It produces slightly tighter correction factors in general.

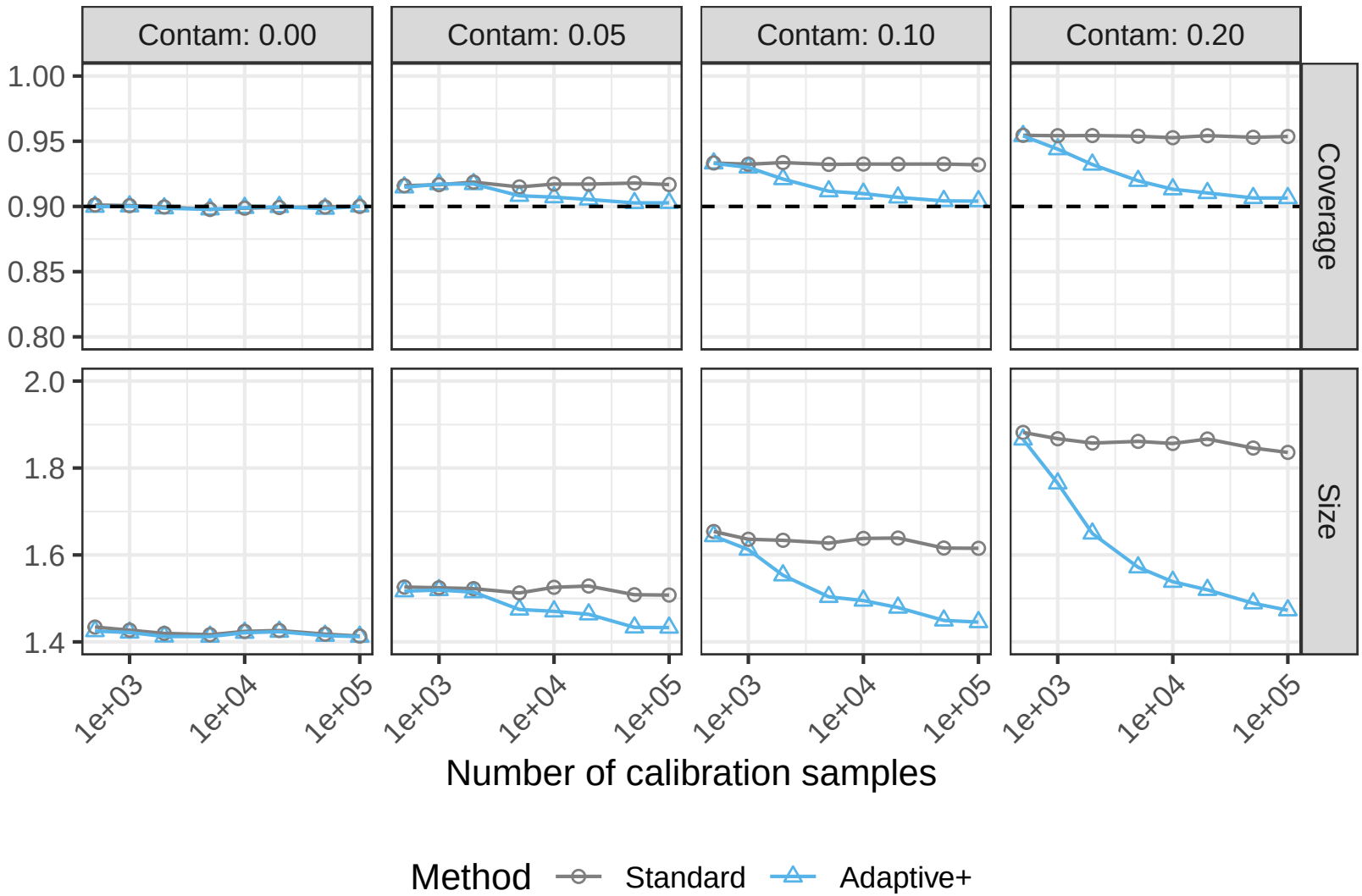
Empirical Performance on Simulated Data

- Synthetic data with 4 possible labels.
- Label noise model: two-level randomized response.
- Metrics: Average coverage and size of prediction sets.



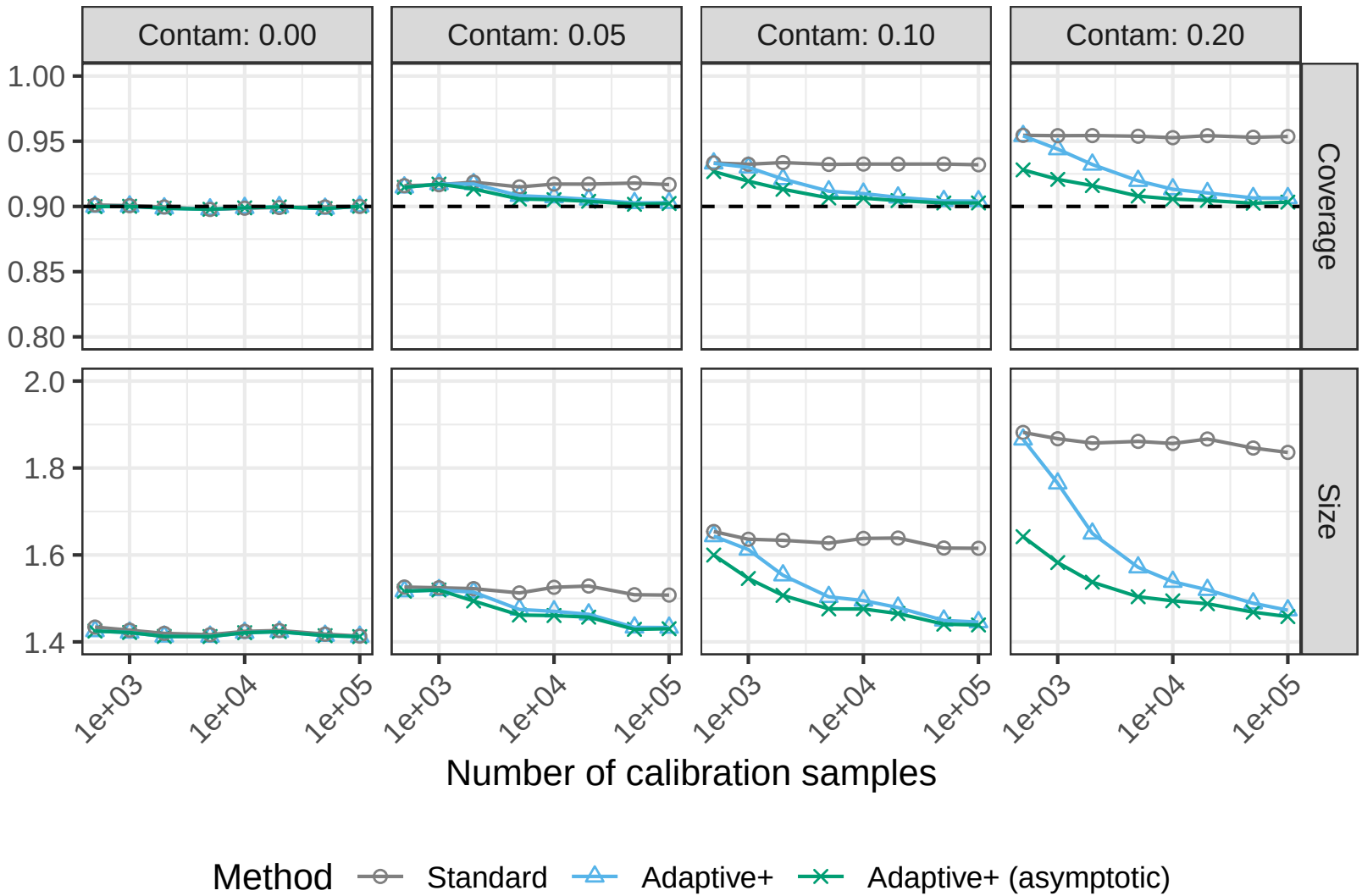
Empirical Performance on Simulated Data

- Synthetic data with 4 possible labels.
- Label noise model: two-level randomized response.
- Metrics: Average coverage and size of prediction sets.



Empirical Performance on Simulated Data

- Synthetic data with 4 possible labels.
- Label noise model: two-level randomized response.
- Metrics: Average coverage and size of prediction sets.



Methodology Extensions

Further Methodology Extensions

1) Relaxing the Assumption of a Known Label Noise Model: Instead of assuming that $W = T^{-1}$ is known, we can work with confidence bounds for its entries.

$$\mathbb{P}\left[\hat{W}_{kl}^{\text{low}} \leq W_{kl} \leq \hat{W}_{kl}^{\text{upp}},\ \forall k \neq l\right] \geq 1 - \alpha_W.$$

Further Methodology Extensions

1) Relaxing the Assumption of a Known Label Noise Model: Instead of assuming that $W = T^{-1}$ is known, we can work with confidence bounds for its entries.

$$\mathbb{P}\left[\hat{W}_{kl}^{\text{low}} \leq W_{kl} \leq \hat{W}_{kl}^{\text{upp}},\ \forall k \neq l\right] \geq 1 - \alpha_W.$$

2) Fitting the Label Noise Model: If we have access to some clean data, we can estimate W empirically.

Further Methodology Extensions

1) Relaxing the Assumption of a Known Label Noise Model: Instead of assuming that $W = T^{-1}$ is known, we can work with confidence bounds for its entries.

$$\mathbb{P}\left[\hat{W}_{kl}^{\text{low}} \leq W_{kl} \leq \hat{W}_{kl}^{\text{upp}},\ \forall k \neq l\right] \geq 1 - \alpha_W.$$

2) Fitting the Label Noise Model: If we have access to some clean data, we can estimate W empirically.

3) Adaptive Predictions with Label-Conditional Coverage: We can construct adaptive prediction sets that guarantee label-conditional coverage.

Further Methodology Extensions

1) Relaxing the Assumption of a Known Label Noise Model: Instead of assuming that $W = T^{-1}$ is known, we can work with confidence bounds for its entries.

$$\mathbb{P}\left[\hat{W}_{kl}^{\text{low}} \leq W_{kl} \leq \hat{W}_{kl}^{\text{upp}},\ \forall k \neq l\right] \geq 1 - \alpha_W.$$

2) Fitting the Label Noise Model: If we have access to some clean data, we can estimate W empirically.

3) Adaptive Predictions with Label-Conditional Coverage: We can construct adaptive prediction sets that guarantee label-conditional coverage.

4) Adaptive Predictions with Calibration-Conditional Coverage: We can construct adaptive prediction sets that guarantee calibration-conditional coverage.

Application to CIFAR-10H Data

The CIFAR-10H Dataset

Dataset Overview:

- 10,000 images across 10 classes: airplane, car, bird, cat, deer, dog, frog, horse, ship, or truck.
- Imperfect labels assigned by approximately 50 human annotators via Amazon Mechanical Turk.
- Noisy labels are correct approximately 95% of the time.
- One noisy label is selected at random for each image.

Classifier:

- A ResNet-18 convolutional neural network, trained on 50,000 CIFAR-10 images.

Potentially Mis-Specified Assumptions:

- $\tilde{Y} \perp\!\!\!\perp X \mid Y$
- Randomized response model

The CIFAR-10H Dataset: Visual Overview

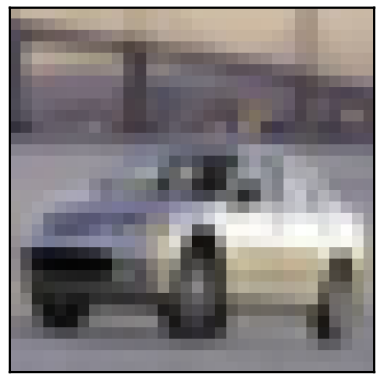
Preview of Prediction Sets for CIFAR-10H Data

True label: frog
Standard:
{frog}
Adaptive+:
{frog}

True label: automobile
Standard:
{automobile}
Adaptive+:
{automobile}

True label: deer
Standard:
{deer}
Adaptive+:
{deer}

True label: automobile
Standard:
{automobile}
Adaptive+:
{automobile}



True label: deer
Standard:
{deer, dog}
Adaptive+:
{deer}

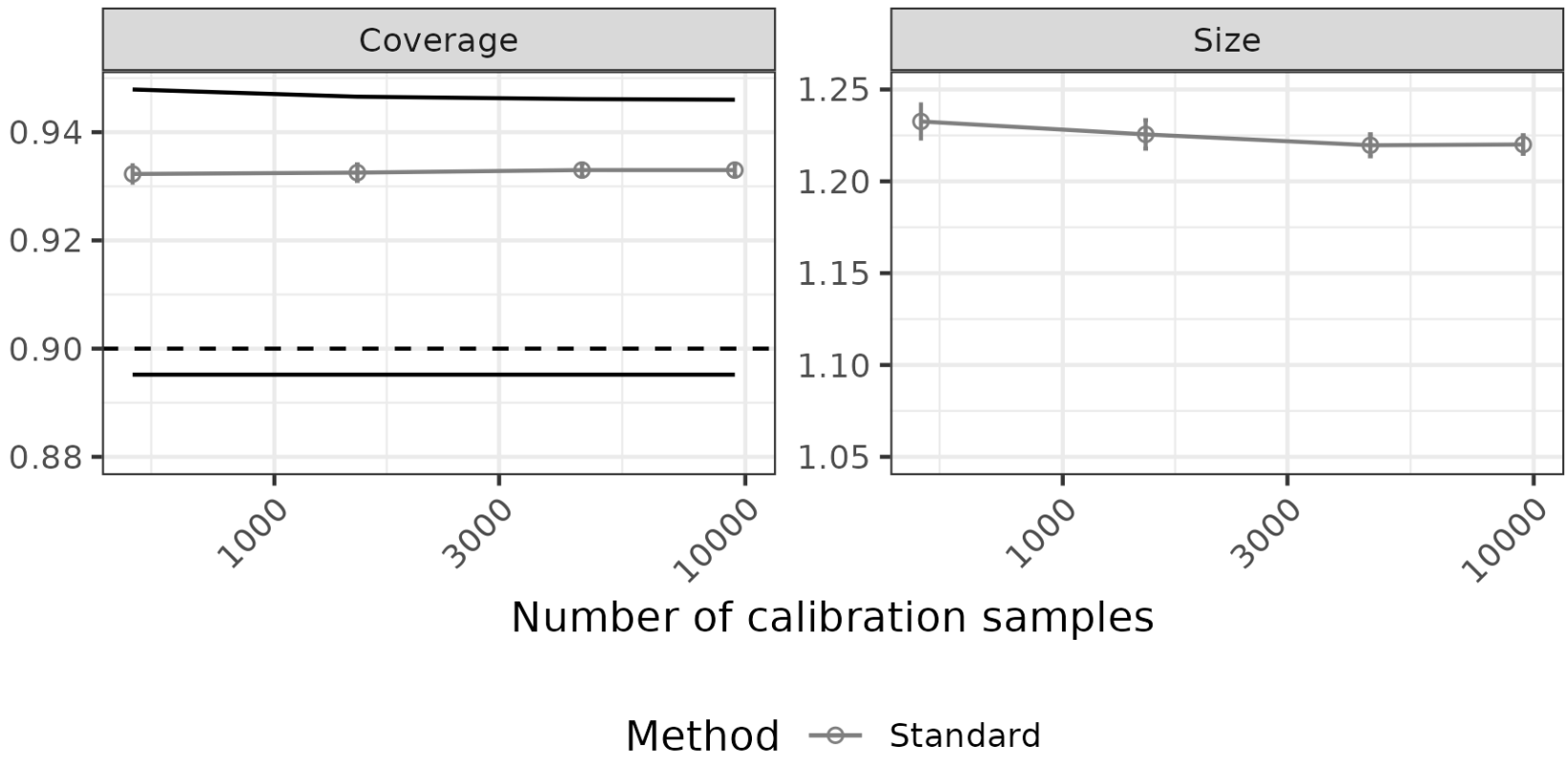
True label: bird
Standard:
{frog, bird, cat}
Adaptive+:
{bird, frog}

True label: deer
Standard:
{deer, bird}
Adaptive+:
{deer}

True label: ship
Standard:
{ship, automobile}
Adaptive+:
{ship}

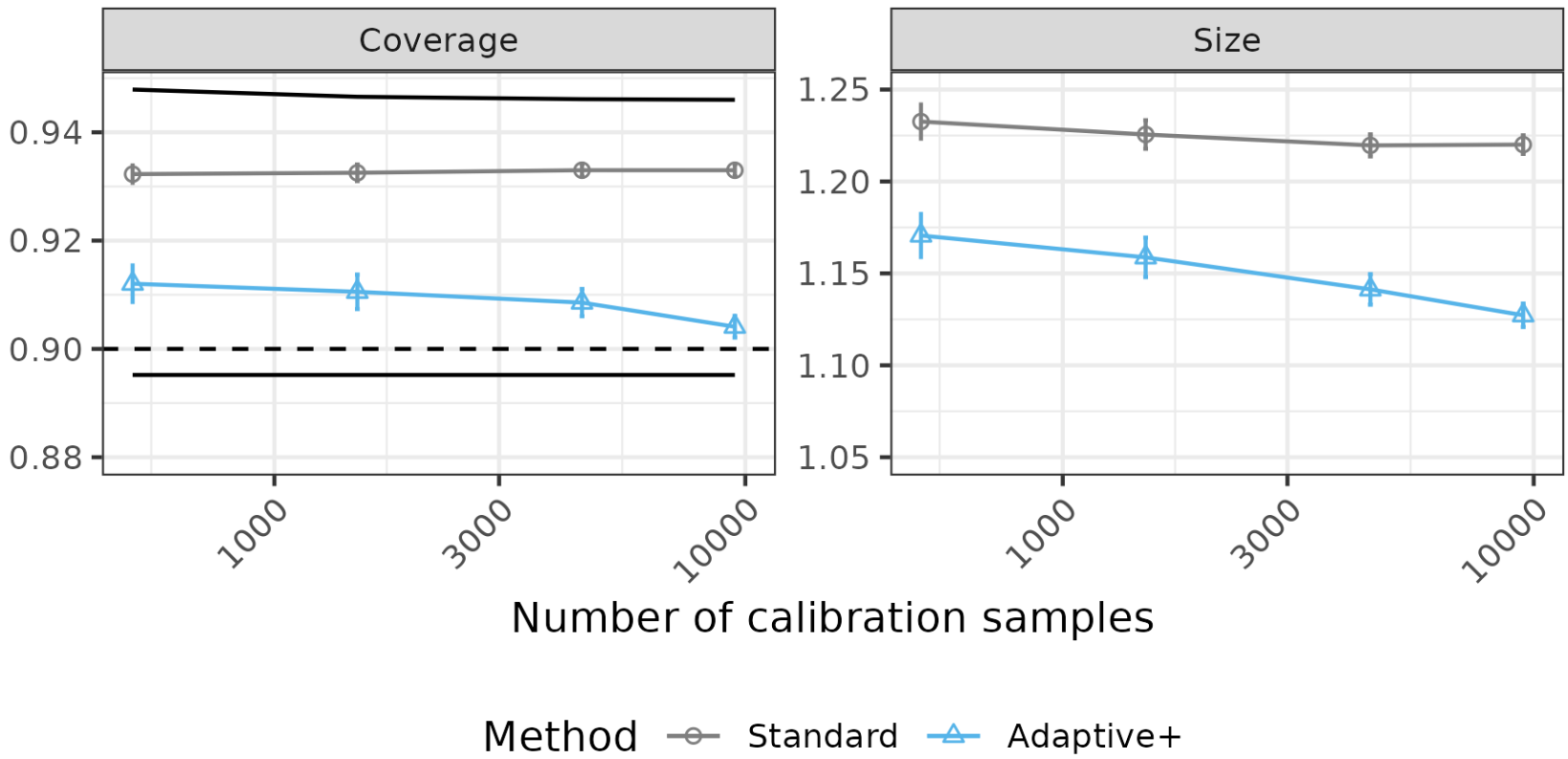
Performance on CIFAR-10H Data

Label Noise Model: Randomized response
Noise Parameter: Fitted from small clean sample
Target Coverage Level: 90%



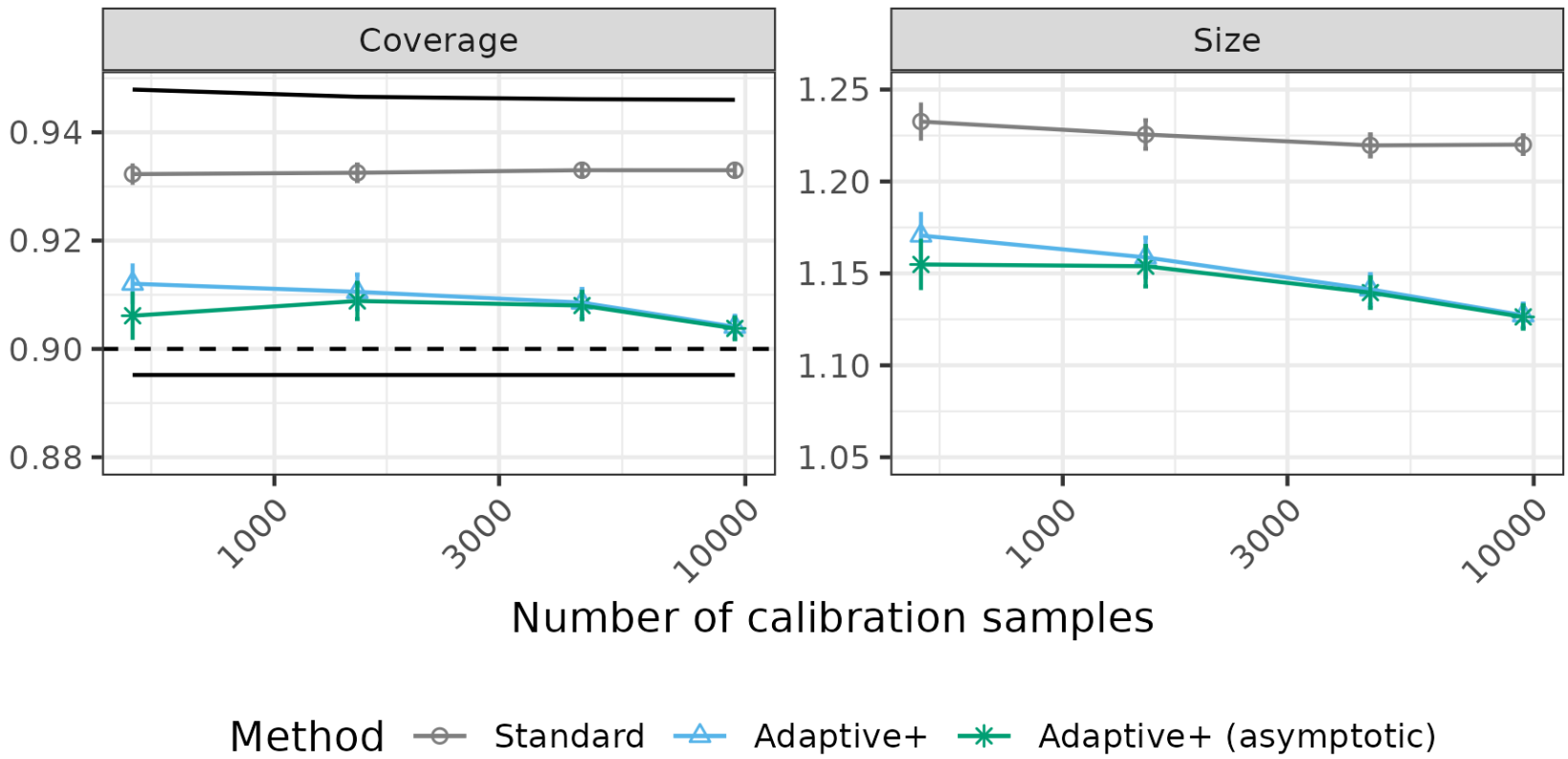
Performance on CIFAR-10H Data

Label Noise Model: Randomized response
Noise Parameter: Fitted from small clean sample
Target Coverage Level: 90%



Performance on CIFAR-10H Data

Label Noise Model: Randomized response
Noise Parameter: Fitted from small clean sample
Target Coverage Level: 90%



Conclusion

Conclusion

Some take-home messages:

- Conformal inference is growing beyond ideal exchangeability, in many directions.
- Different types of non-exchangeability require different solutions.
- Worst-case viewpoints are not always *practically* useful.
- Some modeling assumptions can be useful.

Conclusion

Some take-home messages:

- Conformal inference is growing beyond ideal exchangeability, in many directions.
- Different types of non-exchangeability require different solutions.
- Worst-case viewpoints are not always *practically* useful.
- Some modeling assumptions can be useful.

Some ideas for future work:

- Other types of data with measurement errors
- Other types of data hold-out techniques
- Other ideas? Collaborations?

Thank you!

To learn more:

"Adaptive conformal classification with noisy labels"
Matteo Sesia, Y. X. Rachel Wang, Xin Tong
arxiv.org/abs/2309.05092

A new follow-up paper (with Teresa) will be posted very soon.

Software and tutorials:

https://github.com/msesia/conformal-label-noise

Funding from: