Parametric Scaling Law of Tuning Bias in Conformal Prediction

Hao Zeng

Department of Statistics and Data Science
Southern University of Science and Technology

July 12, 2025
ICML 2025, Vancouver, Canada

Outline

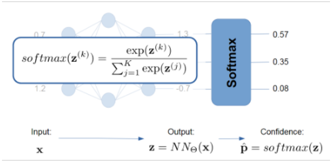
Section 1

Background: Uncertainty in AI

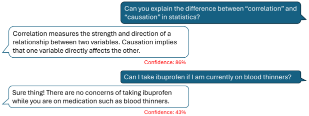
What is uncertainty

Uncertainty in artificial intelligence refers to the model's lack of certainty about its predictions. For example,

- **Classification**: Output label along with its confidence
- **Regression**: Output mean along with its variance.
- **LLM**: perplexity, verbalized confidence, ...



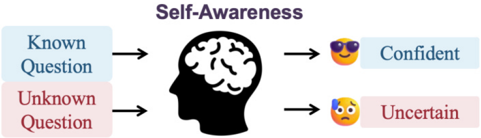
(a) Softmax confidence.



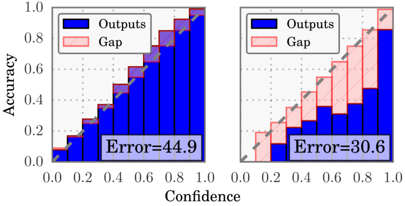
(b) Verbalized confidence.

Why we care about uncertainty?

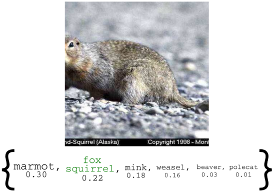
- **Awareness of knowledge boundary**: *know what I know and know what I don't know.*
  hallucination detection, model cascade, slow and deep thinking...
- **Data selection for training/labeling**: *prioritizing samples in which the model is uncertain or certain.*
  active learning, coreset selection, in-context learning...
- **Data privacy**: *identifying information leakage of sensitive data.*
  membership inference attacks, dataset inference, pretraining data detection...



How to express uncertainty?



(a) Confidence calibration



(b) Conformal prediction

Section 2

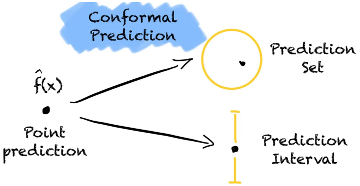
Introduction to Conformal Prediction

Conformal Prediction

Goal: For a given test input $\boldsymbol{x}$, we aim to produce a prediction set $C(\boldsymbol{x})$ containing the true label y satisfying marginal coverage rate $1-\alpha$:

$$\mathbb{P}\left(y \in C(\boldsymbol{x})\right) \geq 1-\alpha.$$

- Larger prediction sets indicate higher uncertainty in the predictions.
- Rigorous, finite-sample, for any model and dataset



Inductive Conformal Prediction

Given a calibration set $\mathscr{D}_{\text{cal}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, and a trained model f,

1. Compute non-conformity scores: $s = S(\boldsymbol{x}, y)$ for $(\boldsymbol{x}, y) \in \mathscr{D}_{\text{cal}}$
   e.g., $S(\boldsymbol{x}, y) = 1 - f_y(\boldsymbol{x})$ for classification[1]
2. Obtain the threshold $\hat{\tau}$ of the scores:

$$\hat{\tau} = \text{Quantile}\left(\{s_1, \ldots, s_n\}, \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right)$$

3. For a new test point $\boldsymbol{x}_{\text{new}}$, the prediction set is:

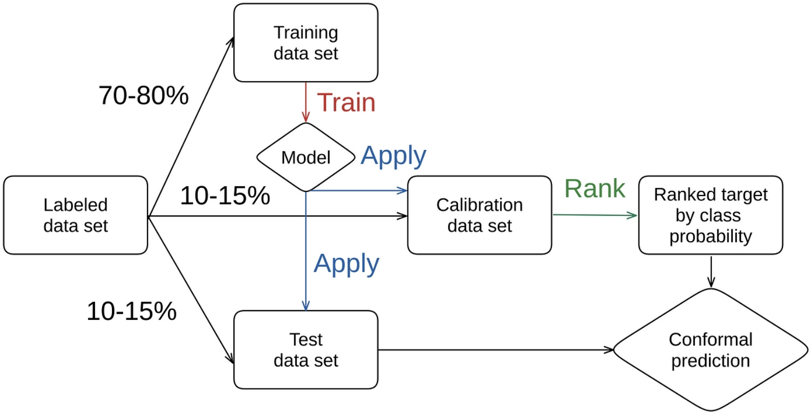
$$\mathscr{C}(\boldsymbol{x}_{\text{new}}) = \{y' \mid S(\boldsymbol{x}_{\text{new}}, y') \leq \hat{\tau}\}$$

The prediction set $\mathscr{C}(\boldsymbol{x}_{\text{new}})$ satisfies the marginal coverage if the calibration and test sets are exchangeable.

Exchangeable data $\Rightarrow$ exchangeable scores $\Rightarrow$ marginal coverage.

[1] Lei, Jing. 2014. “Classification with Confidence.” Biometrika 101 (4): 755–69.

The workflow of CP

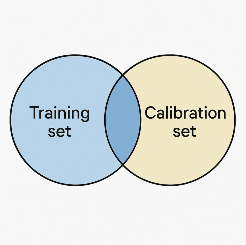


Inductive conformal prediction with APS score

Challenges of CP

If the exchangeability assumption is not satisfied?

- The overlap between the training and calibration sets



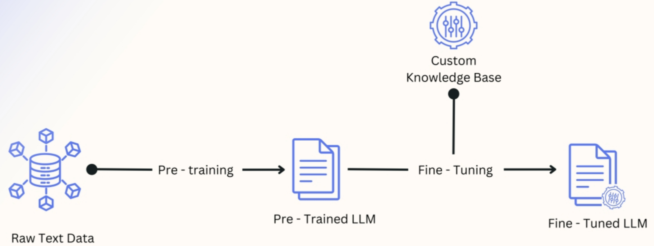
Section 3

Tuning Bias in Conformal Prediction

Parameter Tuning

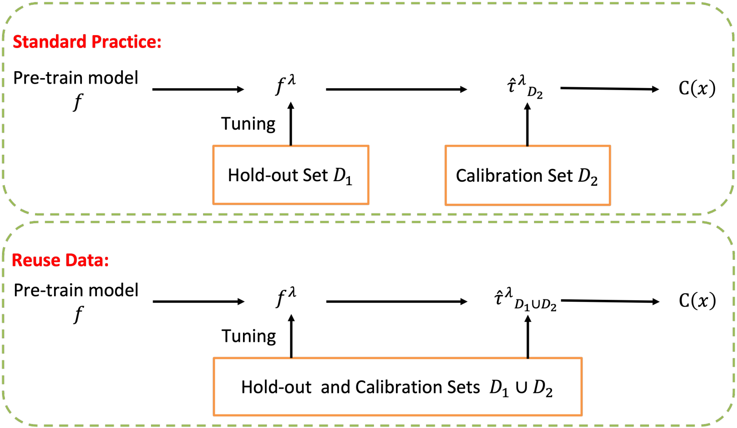
Parameter tuning with a hold-out set is common in deep learning:

- **downstream finetuning**: SFT, prompt tuning, ...
- **confidence calibration**: temperature scaling, vector scaling, ...
- **hyperparameter tuning**: early stopping, model selection, ...



The limited labeled data, when split, is often insufficient for effective tuning and CP.

Parameter Tuning in Conformal Prediction



Reusing data for tuning and conformal prediction breaks exchangeability. So, *how does this violation impact the coverage guarantee*?

Tuning Bias

Definition (Tuning Bias)

Tuning bias is the *additional* coverage gap introduced by reusing the same dataset for tuning and calibration:

$$\text{TuningBias} = \underbrace{\text{CovGap}(C_{\text{same}})}_{\text{Tune \& Calibrate on same set}} - \underbrace{\text{CovGap}(C_{\text{hold-out}})}_{\text{Tune on separate set}},$$

where $\text{CovGap}(C)$ measures the difference between the desired and achieved coverages:

$$\text{CovGap}(C) = |(1-\alpha) - \mathbb{P}(y \in C(\boldsymbol{x}))|$$

Former theoretical results imply that reusing data in parameter tuning and conformal calibration causes large tuning bias. Is it always true?

Tuning Bias is not always increased

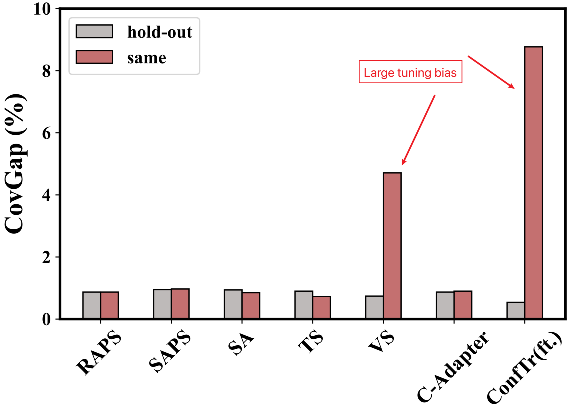
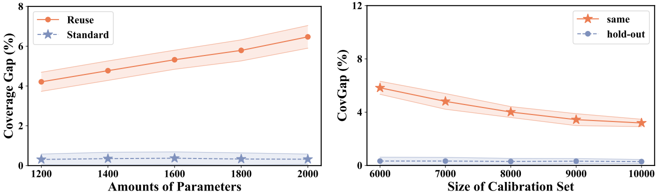


Figure: Tuning bias for various tuning methods

The bias seems related to the complexity of the parameter space being tuned.

The parametric scaling law of tuning bias



(a) Bias vs. Parameter Complexity. (b) Bias vs. Calibration Set Size.

Tuning bias **increases** with the number of tuning parameters, and **decreases** as the calibration set size grows.

A general bound on the coverage gap

Theorem (Thm. 4.1)

When reusing data for tuning, the coverage gap is bounded by:

$$CovGap(C) \leq \underbrace{\mathbb{E}\mathfrak{R}_\Lambda}_{\text{Tuning Bias Term}} + \underbrace{\varepsilon_{\alpha,n}}_{\text{Standard CP Gap}}$$

where $\varepsilon_{\alpha,n} = \lceil (n+1)(1-\alpha) \rceil / n - \alpha$, $\mathfrak{R}_\Lambda$ *is the supremum deviation of empirical probabilities from true probabilities over the entire parameter space* Λ *and* $\mathscr{T}$*:*

$$\mathfrak{R}_\Lambda := \sup_{\lambda \in \Lambda, t \in \mathscr{T}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{S^\lambda(\boldsymbol{x}_i, y_i) \leq t\} - \mathbb{P}(S^\lambda(\boldsymbol{x}, y) \leq t) \right|$$

- The tuning bias is bounded by $\mathbb{E}\mathfrak{R}_\Lambda$.
- This term depends on the complexity of the parameter space Λ.

Theoretical results of the scaling law (I)

Finite Parameter Space (e.g., RAPS with grid search)

Proposition (Simplified from Prop. 4.2)

$$TuningBias = O\left(\sqrt{\frac{\log(|\Lambda|)}{n}}\right)$$

where $|\Lambda|$ is the number of candidate parameters.

This bound shows that tuning bias grows with the logarithm of the number of parameters and decreases with calibration set size. For example, RAPS with a finite grid: Parameter space Λ is a finite grid. Then the bound $\sim \sqrt{\frac{\log(|\Lambda|)}{n}}$ is small $\Rightarrow$ Negligible tuning bias.

Theoretical results of the scaling law (II)

Infinite Parameter Space (e.g., VS, fine-tuning)

Proposition (Simplified from Prop. 4.4)

$$TuningBias = O\left(\sqrt{\frac{d}{n}}\right)$$

where d is the dimension of the parameter space (related to VC-dimension).

For example, confidence calibration like TS and VS with infinite parameter space: TS tunes only one parameter, the temperature ($d = 1$), and its bound $\sim \sqrt{\frac{1}{n}}$ and VS tunes $2K$ parameters ($d = 2K$), its bound $\sim \sqrt{\frac{2K}{n}} \geq \sqrt{\frac{1}{n}}$.

How to mitigate tuning bias?

The scaling law points to two main strategies:

1. Increase the size of the calibration set
2. Reduce Parameter Space Complexity: *Regularization*, e.g, order preserving, weight sharing.

Table: Tuning bias (%) comparison on CIFAR-100 & ImageNet.

Methods	CIFAR-100 (%)	ImageNet (%)
Temperature scaling	**0.14**	**0.04**
Vector scaling	1.13	6.63
ConfTr (ft.) w/ Order-Preserving	**0.52**	**0.40**
ConfTr (ft.) w/o Order-Preserving	6.15	21.68

Take away

1. Tuning bias is not always significant when reusing data for tuning and CP: scales with parameter complexity and inversely with data size
2. Data splitting might be unnecessary when the size is sufficiently large
3. Designing a specific regularization can mitigate the tuning bias

Section 4

Future Work

Open problems of conformal prediction

- **New CP paradigms for generative models:** large language models, vision language models, diffusion models, etc.
- **Beyond exchangeability:** distribution shift, open-vocabulary tasks, etc.
- **Conditional CP:** class-conditional CP, group-conditional CP, etc.

References

The works in this talk

1. Zeng, Hao, Kangdao Liu, Bingyi Jing, and Hongxin Wei. **Parametric Scaling Law of Tuning Bias in Conformal Prediction.** *ICML 2025*.

Other CP works from our group

1. Huang, Jianguo, Huajun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. **Conformal Prediction for Deep Classifier via Label Ranking.** *ICML 2024*.
2. Xi, Huajun, Jianguo Huang, Kangdao Liu, Lei Feng, and Hongxin Wei. **Does Confidence Calibration Improve Conformal Prediction?** *TMLR*.
3. Xi, Huajun, Kangdao Liu, Hao Zeng, Wenguang Sun, and Hongxin Wei. **Robust Online Conformal Prediction under Uniform Label Noise.** *Under review*.
4. Zhou, Xuanning, Hao Zeng, Xiaobo Xia, Bingyi Jing, and Hongxin Wei. **Semi-Supervised Conformal Prediction with Unlabeled Nonconformity Score.** *Under review*.

Thank You!

Code in https://github.com/ml-stat-Sustech/Parametric-Scaling-Law-CP-Tuning