AI Security

Trustworthy ML

Sangdon Park

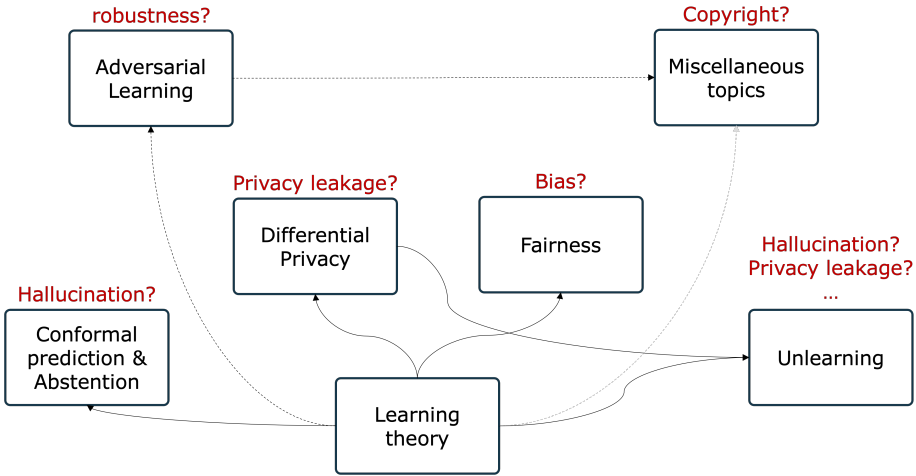
POSTECH

May 28, 2026

Outline

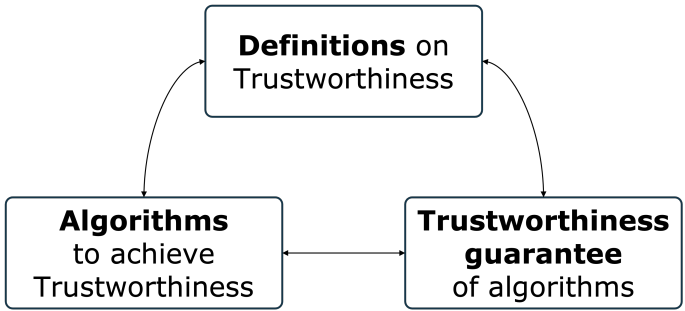
1. Introduction
2. Conformal Prediction
   - Offline Conformal Prediction
   - Online Conformal Prediction with Full Feedback
   - Online Conformal Prediction with Partial Feedback
3. Conformal Abstention
   - Offline Conformal Abstention
   - Online Conformal Abstention Full Feedback
   - Online Conformal Abstention Partial Feedback

Trustworthy Machine Learning: Overview



- Check out our Trustworthy ML class (AIGS703L / CSED703L)

Trustworthy Machine Learning: Key Components



Hallucination in AI Security

- Goal: Learning to minimize hallucination under adversarial attacks

$$\min_{f} \max_{\delta} \mathbb{E}_x \, \texttt{hallucination}(f(x+\delta))$$

  - Learner's Goal: minimize hallucination
  - Adversary's Goal: maximize hallucination

Hallucination in AI Security

- Goal: Learning to minimize hallucination under adversarial attacks

$$\min_{f} \max_{\delta} \mathbb{E}_x \, \texttt{hallucination}(f(x+\delta))$$

  - Learner's Goal: minimize hallucination
  - Adversary's Goal: maximize hallucination

- Example in Agentic AI Security:
  - Goal: Learning to minimize a task failure rate and harmfulness under adversarial attacks

$$\min_{\pi} \mathbb{E}_x \left[\texttt{failure}(\pi(x+\delta))\right] \quad \text{subj. to} \quad \mathbb{E}_x \left[\texttt{harmfulness}(\pi(x+\delta))\right] \leq \alpha \quad \forall \delta$$

Hallucination in AI Security

- Goal: Learning to minimize hallucination under adversarial attacks

$$\min_{f} \max_{\delta} \mathbb{E}_x \, \texttt{hallucination}(f(x+\delta))$$

  - Learner's Goal: minimize hallucination
  - Adversary's Goal: maximize hallucination

- Example in Agentic AI Security:
  - Goal: Learning to minimize a task failure rate and harmfulness under adversarial attacks

$$\min_{\pi} \mathbb{E}_x \left[\texttt{failure}(\pi(x+\delta))\right] \quad \text{subj. to} \quad \mathbb{E}_x \left[\texttt{harmfulness}(\pi(x+\delta))\right] \leq \alpha \quad \forall \delta$$

- Research questions:
  1. How to control a metric (*e.g.*, hallucination) without adversaries?
  2. How to control a metric (*e.g.*, hallucination) with adversaries? – open question

Outline

1. Introduction
2. Conformal Prediction
   - Offline Conformal Prediction
   - Online Conformal Prediction with Full Feedback
   - Online Conformal Prediction with Partial Feedback
3. Conformal Abstention
   - Offline Conformal Abstention
   - Online Conformal Abstention Full Feedback
   - Online Conformal Abstention Partial Feedback

Outline

1. Introduction
2. Conformal Prediction
   - Offline Conformal Prediction
   - Online Conformal Prediction with Full Feedback
   - Online Conformal Prediction with Partial Feedback
3. Conformal Abstention
   - Offline Conformal Abstention
   - Online Conformal Abstention Full Feedback
   - Online Conformal Abstention Partial Feedback

Conformal Prediction

Vladimir Vovk

FOLLOW

Department of Computer Science, Royal Holloway, University of London
Verified email at rhul.ac.uk - Homepage

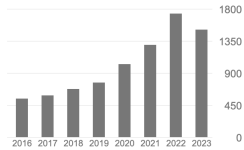
Machine Learning Conformal Prediction Foundations of Probability Foundations of Statistics
Mathematical Finance

GET MY OWN PROFILE

TITLE	CITED BY	YEAR
Algorithmic learning in a random world V Vovk, A Gammerman, G Shafer Springer	1435	2005
Ridge regression learning algorithm in dual variables C Saunders, A Gammerman, V Vovk	1044	1998
Aggregating strategies V Vovk Proceedings of 3rd Annu. Workshop on Comput. Learning Theory, 371-383	956	1990
A Tutorial on Conformal Prediction. G Shafer, V Vovk Journal of Machine Learning Research 9 (3)	853	2008

Cited by VIEW ALL

	All	Since 2018
Citations	14499	7052
h-index	58	41
i10-index	140	91



Conformal Prediction

Vladimir Vovk

Algorithmic learning in a random world

Authors	Vladimir Vovk, Alexander Gammerman, Glenn Shafer
Publication date	2005/3/1
Volume	29
Publisher	Springer
Description	This book is about conformal prediction, an approach to prediction that originated in machine learning in the late 1990s. The main feature of conformal prediction is the principled treatment of the reliability of predictions. The prediction algorithms described—conformal predictors—are provably valid in the sense that they evaluate the reliability of their own predictions in a way that is neither over-pessimistic nor over-optimistic (the latter being especially dangerous). The approach is still flexible enough to incorporate most of the existing powerful methods of machine learning. The book covers both key conformal predictors and the mathematical analysis of their properties. Algorithmic Learning in a Random World contains, in addition to proofs of validity, results about the efficiency of conformal predictors. The only assumption required for validity is that of" randomness"(the prediction algorithm is presented with …
Total citations	Cited by 1435

Scholar articles	Algorithmic learning in a random world V Vovk, A Gammerman, G Shafer - 2005 Cited by 1435 Related articles All 10 versions

> ...we are *hedging* the prediction — we are adding to it a statement about how strongly we believe it.
>
> – Vovk et al., 2005

Motivation

Conventional prediction:

$$\hat{y}\left(\text{  }\right) = \widehat{\text{Bulldog}}$$

Conformal prediction:

$$\hat{C}\left(\text{  }\right) = \left\{\begin{matrix}\text{Toy terrier} \\ \text{Bulldog} \\ \text{poodle}\end{matrix}\right\}$$

Motivation

Conventional prediction:

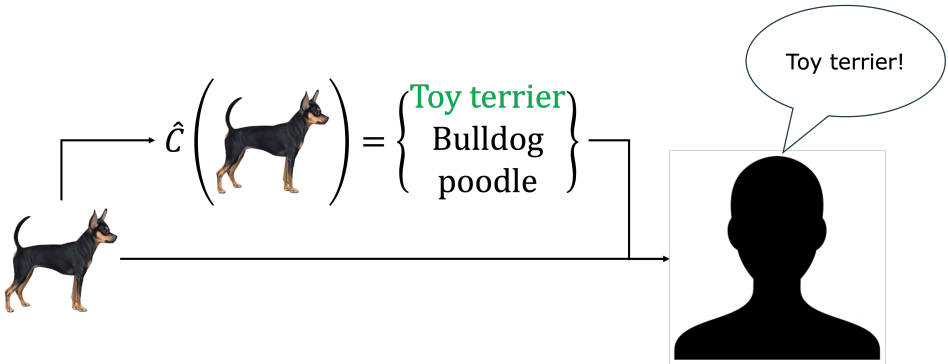
$$f : \mathcal{X} \mapsto \mathcal{Y}$$

Conformal prediction:

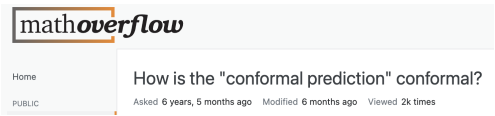
$$C : \mathcal{X} \mapsto 2^{\mathcal{Y}}$$

- Conventional prediction is a "point" prediction.
- Conformal prediction is a set-valued prediction.
- The set contains "likely-correct" alternative options.
  - The set size measures "uncertainty"!
- Why not confidence prediction? User-friendly?

Motivation: Decision Making



Why “Conformal”?



16

+100

Thanks for your interest. The term “conformal prediction” was suggested by Glenn Shafer, and at first I did not like it exactly for the reason that you mention: it has nothing (or very little) to do with conformal mappings in complex analysis. But then I discovered other meanings, even in maths; e.g., Wikipedia has five on its disambiguation page for “conformal”:

- Conformal film on a surface (same thickness)
- Conformal fuel tanks on military aircraft
- Conformal coating in electronics
- Conformal hypergraph, in mathematics
- Conformal software, in ASIC Software

So the word did not look taken to me anymore. The expression that we had used before Glenn proposed “conformal prediction” was even worse (“transductive confidence machine”).

Thanks to Hengrui Luo for drawing my attention to this question.

As for question (2), the answer depends on which robust predictors you have in mind. The predictors with most similar properties are the ones in classical statistics (such as the standard prediction intervals in linear regression based on Student's t distribution); the main difference is that they are parametric. There is a predictive version of tolerance intervals in nonparametric statistics, but their treatment of objects (x parts of observations (x,y), where y are labels) is limited. Upper bounds on the probability of error given by standard PAC predictors are often too high to be useful.

Share Cite Improve this answer Follow

answered Apr 13, 2017 at 7:35
Vladimir Vovk

Conformal (Prediction) Sets

Definition (conformal set)

$$C(x) \coloneqq \{y \in \mathcal{Y} \mid f(x, y) \geq q\}$$

Conformal (Prediction) Sets

Definition (conformal set)

$$C(x) \coloneqq \{y \in \mathcal{Y} \mid f(x, y) \geq q\}$$

- We are using more recent notations based on inductive conformal prediction.
  - The notations are from [Lei et al., 2018], [Vovk et al., 2005], [Tibshirani et al., 2019], and their combination.
  - Note that inductive conformal prediction [Papadopoulos et al., 2002] is an efficient variation of full conformal prediction [Vovk et al., 2005].

Conformal (Prediction) Sets

Definition (conformal set)

$$C(x) \coloneqq \{y \in \mathcal{Y} \mid f(x, y) \geq q\}$$

- We are using more recent notations based on inductive conformal prediction.
  - The notations are from [Lei et al., 2018], [Vovk et al., 2005], [Tibshirani et al., 2019], and their combination.
  - Note that inductive conformal prediction [Papadopoulos et al., 2002] is an efficient variation of full conformal prediction [Vovk et al., 2005].
- $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$: a conformity scoring function
  - Measures how well (x, y) *conforms* to a trained model f (via a proper training set)
  - $f(x, y)$ is a likelihood of x for being y

Conformal (Prediction) Sets

Definition (conformal set)

$$C(x) \coloneqq \{y \in \mathcal{Y} \mid f(x, y) \geq q\}$$

- We are using more recent notations based on inductive conformal prediction.
  - The notations are from [Lei et al., 2018], [Vovk et al., 2005], [Tibshirani et al., 2019], and their combination.
  - Note that inductive conformal prediction [Papadopoulos et al., 2002] is an efficient variation of full conformal prediction [Vovk et al., 2005].
- $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$: a conformity scoring function
  - Measures how well (x, y) *conforms* to a trained model f (via a proper training set)
  - $f(x, y)$ is a likelihood of x for being y
- q: A parameter to be chosen by an algorithm.

Conformity Scoring Functions I

Conformity scoring functions can be (almost) any model!

Example (classification)

$$f(x, y) \coloneqq f_{\text{cls}}(x, y)$$

- f_{cls}: a classification model, *e.g.*, resnet

Conformity Scoring Functions II

Conformity scoring functions can be (almost) any model!

Example (standard regression in 1-dimension)

$$f(x, y) := -|\mu(x) - y|$$

- μ: a regressor

Back to Conformal Sets

Definition (conformal sets)

$$C(x) \coloneqq \{y \in \mathcal{Y} \mid f(x, y) \geq q\}$$

- A conformity scoring function f is given.
- f is a target to measure uncertainty.
- How to choose q?

Assumption: Exchangeability

Assumption

A sequence of random variables $X_1, X_2, \ldots$ is exchangeable if for any permutation σ, the following holds:

$$\mathbb{P}\left\{X_1 = x_1, X_2 = x_2, \ldots\right\} = \mathbb{P}\left\{X_{\sigma(1)} = x_1, X_{\sigma(2)} = x_2, \ldots\right\}.$$

Assumption: Exchangeability

Assumption

A sequence of random variables $X_1, X_2, \ldots$ is exchangeable if for any permutation σ, the following holds:

$$\mathbb{P}\left\{X_1 = x_1, X_2 = x_2, \ldots\right\} = \mathbb{P}\left\{X_{\sigma(1)} = x_1, X_{\sigma(2)} = x_2, \ldots\right\}.$$

- The i.i.d. assumption implies the exchangeability assumption (why?).

A Goodness Metric: Coverage Guarantee

Definition (coverage guarantee)

$$\mathbb{P}\Big\{ Y_{n+1} \in \hat{C}(X_{n+1}) \Big\} \geq 1 - \alpha$$

A Goodness Metric: Coverage Guarantee

Definition (coverage guarantee)

$$\mathbb{P}\Big\{ Y_{n+1} \in \hat{C}(X_{n+1}) \Big\} \geq 1 - \alpha$$

- $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ for $i = 1, \ldots, n$: a training set

A Goodness Metric: Coverage Guarantee

Definition (coverage guarantee)

$$\mathbb{P}\Big\{ Y_{n+1} \in \hat{C}(X_{n+1}) \Big\} \geq 1 - \alpha$$

- $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ for $i = 1, \ldots, n$: a training set
- The probability is taken over (X_i, Y_i) for $i = 1, \ldots, n+1$.

A Goodness Metric: Coverage Guarantee

Definition (coverage guarantee)

$$\mathbb{P}\Big\{ Y_{n+1} \in \hat{C}(X_{n+1}) \Big\} \geq 1 - \alpha$$

- $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ for $i = 1, \ldots, n$: a training set
- The probability is taken over (X_i, Y_i) for $i = 1, \ldots, n+1$.
- (X_i, Y_i) for $i = 1, \ldots, n+1$: the exchangeable samples (thus the i.i.d. samples)

A Goodness Metric: Coverage Guarantee

Definition (coverage guarantee)

$$\mathbb{P}\Big\{ Y_{n+1} \in \hat{C}(X_{n+1}) \Big\} \geq 1 - \alpha$$

- $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ for $i = 1, \ldots, n$: a training set
- The probability is taken over (X_i, Y_i) for $i = 1, \ldots, n+1$.
- (X_i, Y_i) for $i = 1, \ldots, n+1$: the exchangeable samples (thus the i.i.d. samples)
- $\hat{C}$: A conformal set constructed by the training set

A Goodness Metric: Coverage Guarantee

Definition (coverage guarantee)

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

- $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ for $i = 1, \ldots, n$: a training set
- The probability is taken over (X_i, Y_i) for $i = 1, \ldots, n+1$.
- (X_i, Y_i) for $i = 1, \ldots, n+1$: the exchangeable samples (thus the i.i.d. samples)
- $\hat{C}$: A conformal set constructed by the training set
- $1 - \alpha \in (0, 1)$: A desired coverage rate

Quantile

Quantile of a Distribution

The level β quantile of a distribution F:

Definition (quantile)

$$\text{Quantile}(\beta; F) := \inf\{z \mid \mathbb{P}\{Z \leq z\} \geq \beta\}$$

- F: a distribution over the augmented real line, $\mathbb{R} \cup \{\infty\}$
- $Z \sim F$
  - allows multiple instances of the same element

Quantile

Quantile of an Empirical Distribution

The level β quatile of an empirical distribution of the values $v_{1:n}$:

Definition (quantile)

$$\text{Quantile}(\beta; v_{1:n}) := \text{Quantile}\left(\beta; \frac{1}{n}\sum_{i=1}^{n}\delta_{v_i}\right)$$

- $v_{1:n} := \{v_1, \ldots, v_n\}$: an unordered multiset
- δ_a: a δ-distribution (*i.e.*, a point mass at a)

Quantile Algorithm

Definition (quantile algorithm)

Given $(X_1, Y_1), \ldots, (X_n, Y_n)$,

$$\hat{q}_{1-\alpha} := \text{Quantile}(1 - \alpha, V_{1:n} \cup \{\infty\}),$$

where $V_i := -f(X_i, Y_i)$.

- The implementation is as simple as finding the k-th smallest value.

Coverage Guarantee of the Quantile Algorithm

Theorem ([Vovk et al., 2005, Lei et al., 2018])

Assume that (X_i, Y_i) *for* $i \in \{1, \dots, n+1\}$ *are exchangeable. For any scoring function* f *and any* $\alpha \in (0, 1)$*, denote the conformal set by*

$$\hat{C}(x) \coloneqq \left\{ y \in \mathcal{Y} \;\middle|\; -f(x, y) \leq \hat{q}_{1-\alpha} \right\}.$$

Then, we have

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

where the probability is taken over (X_i, Y_i)*.*

- This is a marginal coverage guarantee.

Power of Conformal Prediction

The coverage guarantee is drawn with minimal assumptions.

- It does not make assumptions on a distribution except for the exchangeability.
- The guarantee holds for any conformity scoring function.

Size of Conformal Sets

- Application-dependent issues
  - classification: set size
  - 1-D regression: interval length
  - multi-dimentional regression: *e.g.*, volume
- Larger set: uncertain (*e.g.*, the entire set)
- Smaller set: more certain (*e.g.*, a singleton)

Conclusion

- Conformal prediction is a powerful tool to construct a prediction set (for measuring uncertainty) with correctness guarantees.
- Conformal prediction has many applications due to its "distribution-free" and "scoring-function-free" nature.

Outline

1. Introduction
2. Conformal Prediction
   - Offline Conformal Prediction
   - Online Conformal Prediction with Full Feedback
   - Online Conformal Prediction with Partial Feedback
3. Conformal Abstention
   - Offline Conformal Abstention
   - Online Conformal Abstention Full Feedback
   - Online Conformal Abstention Partial Feedback

Motivation: Distribution Shift

- The main assumption of conformal prediction: exchangeability or i.i.d.
- In practice, this is fragile due to distribution shifts.
- Type of distribution shifts
  - Covariate shift
  - Label shift
  - ...
  - Adversarial shift

Adversarial Shift

- Learning setup: follows an online learning setup, *i.e.*,
  - there are multiple shifts over time
  - $p_t(x, y)$: a distribution at time t
  - $(x_t, y_t) \sim p_t(x, y)$: a labeled example sampled at time t

Adversarial Shift

- Learning setup: follows an online learning setup, *i.e.*,
  - there are multiple shifts over time
  - $p_t(x, y)$: a distribution at time t
  - $(x_t, y_t) \sim p_t(x, y)$: a labeled example sampled at time t
- Assumption: no restriction on shifts

Adversarial Shift

- Learning setup: follows an online learning setup, *i.e.*,
  - there are multiple shifts over time
  - $p_t(x, y)$: a distribution at time t
  - $(x_t, y_t) \sim p_t(x, y)$: a labeled example sampled at time t
- Assumption: no restriction on shifts
- Conformal prediction under distribution shift
  - [Gibbs and Candès, 2021]: provides the coverage guarantee
  - [Bastani et al., 2022]: provides the coverage guarantee for fairness

Adaptive Conformal Prediction

Can we learn conformal sets under distribution shift?

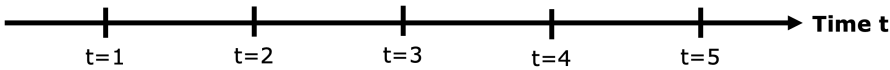
Setup:

- $\mathcal{X}$: example space
- $\mathcal{Y}$: label space
- $C_t : \mathcal{X} \to 2^{\mathcal{Y}}$: a conformal set
- A learning game between a learner and nature

for $t = 1, \ldots, T$ **do**
Learner receives an example $x_t \in \mathcal{X}$
Learner outputs a *conformal set* $C_t(x_t) \in 2^{\mathcal{Y}}$
Learner receives a true label $y_t \in \mathcal{Y}$
Learner suffers loss $\mathbb{1}(y_t \notin C_t(x_t))$
Learner update a parameter of a conformal set
end for

Adaptive Conformal Prediction

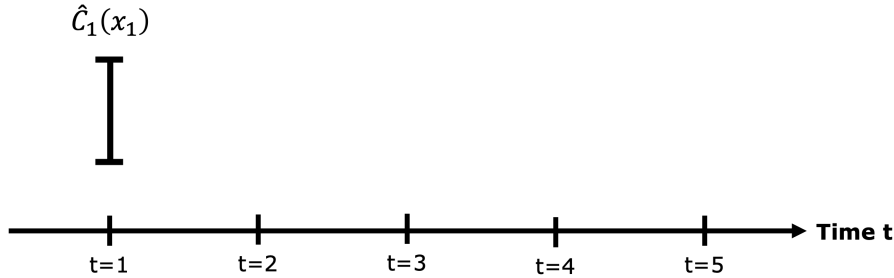
Intuition



- Adaptive conformal prediction progressively adjust a prediction set such that it covers a desired number of samples.

Adaptive Conformal Prediction

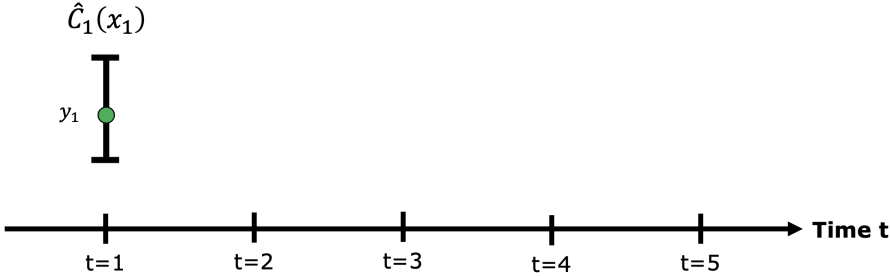
Intuition



- Adaptive conformal prediction progressively adjust a prediction set such that it covers a desired number of samples.

Adaptive Conformal Prediction

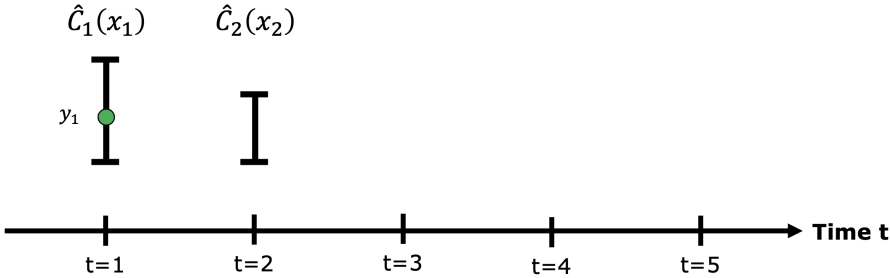
Intuition



- Adaptive conformal prediction progressively adjust a prediction set such that it covers a desired number of samples.

Adaptive Conformal Prediction

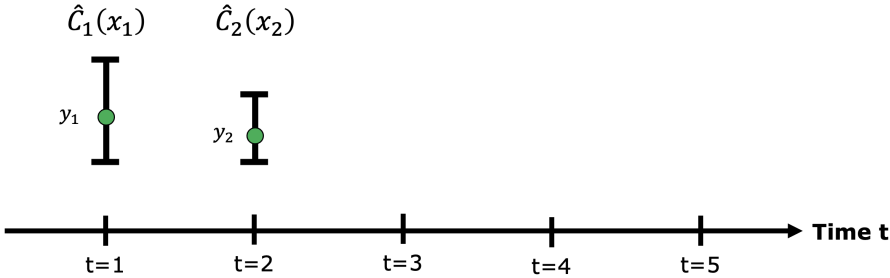
Intuition



- Adaptive conformal prediction progressively adjust a prediction set such that it covers a desired number of samples.

Adaptive Conformal Prediction

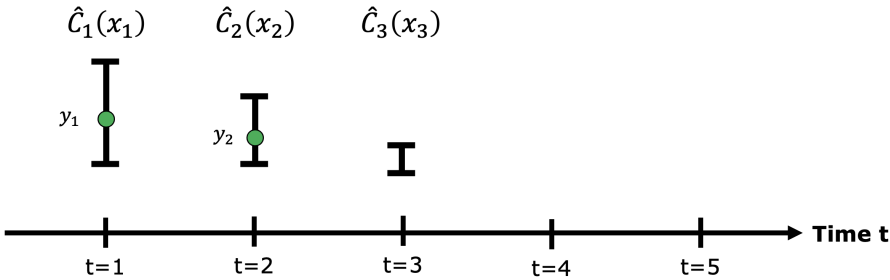
Intuition



- Adaptive conformal prediction progressively adjust a prediction set such that it covers a desired number of samples.

Adaptive Conformal Prediction

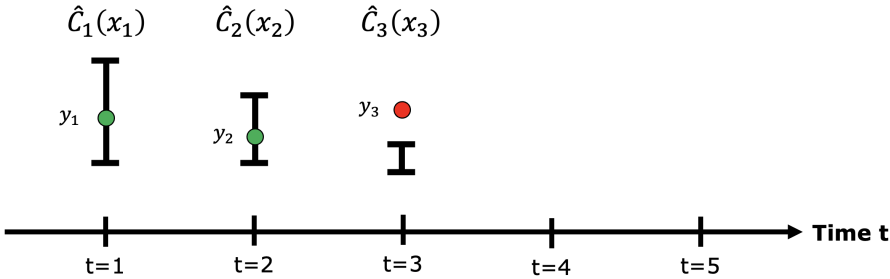
Intuition



- Adaptive conformal prediction progressively adjust a prediction set such that it covers a desired number of samples.

Adaptive Conformal Prediction

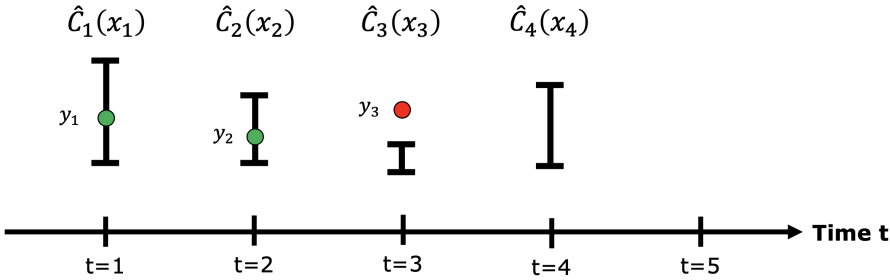
Intuition



- Adaptive conformal prediction progressively adjust a prediction set such that it covers a desired number of samples.

Adaptive Conformal Prediction

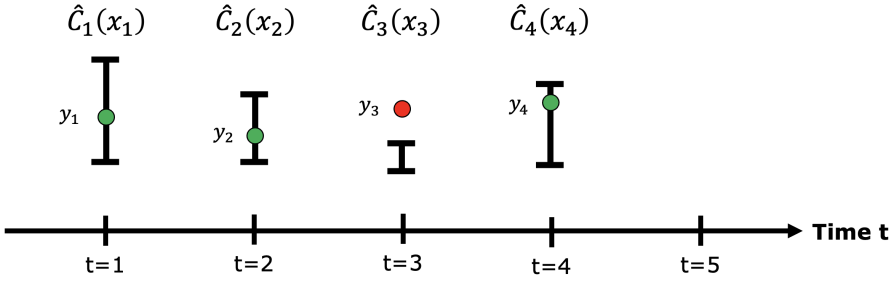
Intuition



- Adaptive conformal prediction progressively adjust a prediction set such that it covers a desired number of samples.

Adaptive Conformal Prediction

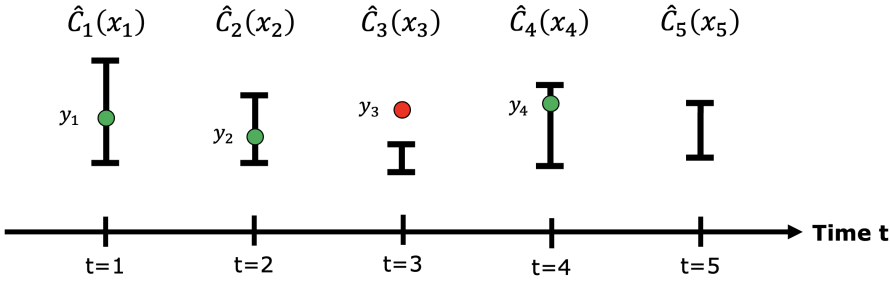
Intuition



- Adaptive conformal prediction progressively adjust a prediction set such that it covers a desired number of samples.

Adaptive Conformal Prediction

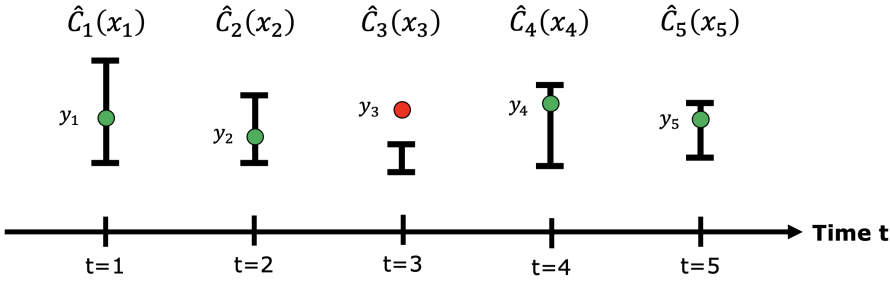
Intuition



- Adaptive conformal prediction progressively adjust a prediction set such that it covers a desired number of samples.

Adaptive Conformal Prediction

Intuition



- Adaptive conformal prediction progressively adjust a prediction set such that it covers a desired number of samples.

A Goodness Metric: "Empirical" Coverage Guarantee

Definition (empirical coverage guarantee)

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(y_t \notin \hat{C}_t(x_t) \right) - \alpha \right|$$

- $1 - \alpha$: a desired coverage rate
- T: a time horizon
- $\hat{C}_t$: a conformal set at time t constructed by an algorithm
- It is similar to the regret definition (but not exactly the same).
- We wish to bound this quantity.

A Goodness Metric: "Empirical" Coverage Guarantee

Definition (empirical coverage guarantee)

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(y_t \notin \hat{C}_t(x_t) \right) - \alpha \right|$$

- $1 - \alpha$: a desired coverage rate
- T: a time horizon
- $\hat{C}_t$: a conformal set at time t constructed by an algorithm
- It is similar to the regret definition (but not exactly the same).
- We wish to bound this quantity.
- Why not use the PAC guarantee?

A Goodness Metric: "Empirical" Coverage Guarantee

Definition (empirical coverage guarantee)

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(y_t \notin \hat{C}_t(x_t) \right) - \alpha \right|$$

- $1 - \alpha$: a desired coverage rate
- T: a time horizon
- $\hat{C}_t$: a conformal set at time t constructed by an algorithm
- It is similar to the regret definition (but not exactly the same).
- We wish to bound this quantity.
- Why not use the PAC guarantee?
  - the PAC guarantee is for the batch learning.

Algorithm

Main Ideas

- Run the batch conformal prediction (CP) for each time
- But adjust the coverage α for the batch CP to satisfy the empirical coverage guarantee.

Algorithm

Algorithm 1 A standard version of Adaptive Conformal Inference [Gibbs and Candès, 2021]

$t_1 \in \{1, \ldots, T\}$
$\alpha_{t_1} \in [0, 1]$
for $t = t_1, \ldots, T$ **do**
    $(\mathcal{D}^{(t)}_{\text{train}}, \mathcal{D}^{(t)}_{\text{cal}}) \leftarrow$ Randomly split the data $\{(x_i, y_i)\}_{i=1}^{t-1}$ and obtain non-conformity scores
    $S_t \leftarrow$ Update a scoring function using $\mathcal{D}^{(t)}_{\text{train}}$
    $q_t \leftarrow \text{Quantile}(1 - \alpha_t, \mathcal{D}^{(t)}_{\text{cal}} \cup \{\infty\})$
    Observe x_t
    Predict $\hat{C}_t(x_t)$
    Observe y_t
    Update $\alpha_{t+1} \leftarrow \alpha_t + \gamma \left(\alpha - \mathbb{1}\left(y_t \notin \hat{C}_t(x_t)\right)\right)$
end for

- A conformal set: $\hat{C}_t(x_t) \coloneqq \{y \in \mathcal{Y} \mid S_t(x_t, y) \leq q_t\}$
- Until t_1, the algorithm simply collects data.
- The algorithm is not randomized.

Coverage Bound

Theorem

For all $T \in \mathbb{N}$, $\alpha \in (0,1)$, *and* $\gamma > 0$,

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(y_t \notin \hat{C}_t(x_t)\right) - \alpha \right| \leq \frac{\max\{\alpha_1, 1-\alpha_1\} + \gamma}{T\gamma}$$

Coverage Bound

Theorem

For all $T \in \mathbb{N}$, $\alpha \in (0,1)$, *and* $\gamma > 0$,

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(y_t \notin \hat{C}_t(x_t)\right) - \alpha \right| \leq \frac{\max\{\alpha_1, 1-\alpha_1\} + \gamma}{T\gamma}$$

- The coverage decreases by $\mathcal{O}\left(\frac{1}{T}\right)$

Coverage Bound

Theorem

For all $T \in \mathbb{N}$, $\alpha \in (0,1)$, *and* $\gamma > 0$,

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(y_t \notin \hat{C}_t(x_t)\right) - \alpha \right| \leq \frac{\max\{\alpha_1, 1-\alpha_1\} + \gamma}{T\gamma}$$

- The coverage decreases by $\mathcal{O}\left(\frac{1}{T}\right)$
- This holds for any sequence $((x_1, y_1), \ldots, (x_T, y_T))$!

Coverage Bound

Theorem

For all $T \in \mathbb{N}$, $\alpha \in (0,1)$, *and* $\gamma > 0$,

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(y_t \notin \hat{C}_t(x_t)\right) - \alpha \right| \leq \frac{\max\{\alpha_1, 1-\alpha_1\} + \gamma}{T\gamma}$$

- The coverage decreases by $\mathcal{O}\left(\frac{1}{T}\right)$
- This holds for any sequence $((x_1, y_1), \ldots, (x_T, y_T))$!
  - If $\hat{C}_t(x_t) = \mathcal{Y}$, the adversary will never win without randomization.

Coverage Bound

Theorem

For all $T \in \mathbb{N}$, $\alpha \in (0, 1)$, *and* $\gamma > 0$,

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(y_t \notin \hat{C}_t(x_t) \right) - \alpha \right| \leq \frac{\max\{\alpha_1, 1 - \alpha_1\} + \gamma}{T\gamma}$$

- The coverage decreases by $\mathcal{O}\left(\frac{1}{T}\right)$
- This holds for any sequence $((x_1, y_1), \ldots, (x_T, y_T))$!
  - If $\hat{C}_t(x_t) = \mathcal{Y}$, the adversary will never win without randomization.
- Suppose $\alpha_1 = 0$, $\gamma = 0.01$, and $\varepsilon = 0.01$ (which denotes the upper bound). Then, we $T = 10,100$ observations to make the empirical coverage close to a desired coverage.

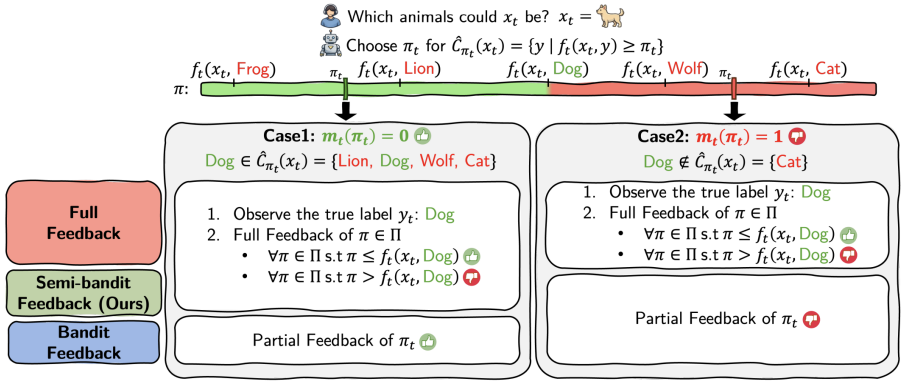
Conclusion

- Adaptive Conformal Inference [Gibbs and Candès, 2021] is the first approach to learn a conformal set under distribution shifts.
- This is an example of running a batch algorithm within an online algorithm.
  - The time and memory complexity is linear in T.
  - See a more efficient (and general) approach [Bastani et al., 2022]

Outline

1. Introduction
2. Conformal Prediction
   - Offline Conformal Prediction
   - Online Conformal Prediction with Full Feedback
   - Online Conformal Prediction with Partial Feedback
3. Conformal Abstention
   - Offline Conformal Abstention
   - Online Conformal Abstention Full Feedback
   - Online Conformal Abstention Partial Feedback

OCP with Adversarial Semi-Bandit Feedback [Yang et al., 2026]



- A novel online conformal prediction method, updated by only using partial feedback

Outline

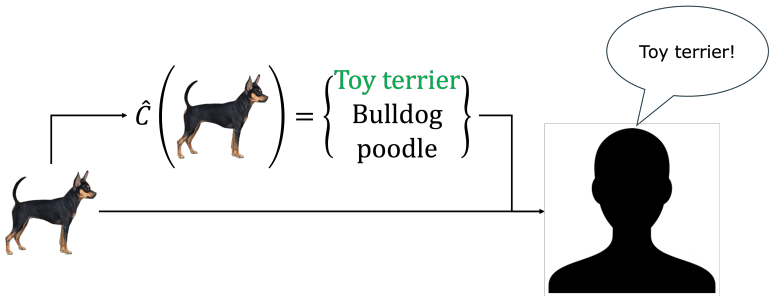
1. Introduction
2. Conformal Prediction
   - Offline Conformal Prediction
   - Online Conformal Prediction with Full Feedback
   - Online Conformal Prediction with Partial Feedback
3. Conformal Abstention
   - Offline Conformal Abstention
   - Online Conformal Abstention Full Feedback
   - Online Conformal Abstention Partial Feedback

Outline

1. Introduction
2. Conformal Prediction
   - Offline Conformal Prediction
   - Online Conformal Prediction with Full Feedback
   - Online Conformal Prediction with Partial Feedback
3. Conformal Abstention
   - Offline Conformal Abstention
   - Online Conformal Abstention Full Feedback
   - Online Conformal Abstention Partial Feedback

Motivation

- Conformal prediction is fine but requires post-processing (*i.e.*, human-in-the-loop)



- Can we use this in fully automated systems (*i.e.*, without-human-in-the-loop)?

Learning Setup

A standard supervised learning setup:

- $\mathcal{X}$: an example space
- $\mathcal{Y}$: an example space
- $\mathcal{D}$: a distribution over $\mathcal{X} \times \mathcal{Y}$
- $Z \sim \mathcal{D}^n$: a calibration set
- $\mathcal{F}$: a set of selective predictors (will introduce soon)
- loss: a false discovery rate (will introduce soon)

A Selective Classifier

Definition (a selective classifier)

$$\hat{S}(x) = \begin{cases} \hat{y}(x) & \text{if } f(x, \hat{y}(x)) \geq \tau \\ \texttt{IDK} & \text{otherwise} \end{cases}$$

- $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$: a scoring function
- $\hat{y} : \mathcal{X} \to \mathcal{Y}$: a classifier, *i.e.*,
$$\hat{y}(x) \coloneqq \arg\max_{y \in \mathcal{Y}} f(x, y)$$
- $\tau \in \mathbb{R}_{\geq 0}$: a parameter
- `IDK`: "I don't know"
- $\hat{S} : \mathcal{X} \to \mathcal{Y} \cup \{\texttt{IDK}\}$: a selective classifier

A Goodness Metric: False Discovery Rate

Definition (false discovery rate (FDR))

$$\mathbb{P}\left\{ y \neq \hat{S}(x) \,\middle|\, \hat{S}(x) \neq \texttt{IDK} \right\}$$

- the FDR is equivalent to the precision
- The following equivalence may be useful:

$$\begin{aligned}\mathbb{P}\left\{ y \neq \hat{S}(x) \,\middle|\, \hat{S}(x) \neq \texttt{IDK} \right\} &= \mathbb{P}\left\{ y \neq \hat{y}(x) \,\middle|\, \hat{S}(x) \neq \texttt{IDK} \right\} \\ &= \mathbb{P}\left\{ y \neq \hat{y}(x) \,\middle|\, f(x, \hat{y}(x)) \geq \tau \right\}\end{aligned}$$

- uncomfortable fact: the FDR is not monotonic in τ (you will see why)

Goal: Achieving a PAC-Style Guarantee

Goal

Find a PAC algorithm that returns $\hat{S}$, *i.e.*,

$$\mathbb{P}\left\{\mathbb{P}\left\{y \neq \hat{S}(x) \,\middle|\, \hat{S}(x) \neq \texttt{IDK}\right\} \leq \varepsilon\right\} \geq 1 - \delta$$

- This is an ideal goal.
- This PAC guarantee for any ε can be achievable under some condition.
  - This is related to the monotonicity of the FDR.

Assumption

Assumption (i.i.d.)

Labeled examples are independently drawn from the same (and unknown) distribution $\mathcal{D}$ over labeled examples $\mathcal{X} \times \mathcal{Y}$.

- Same as the assumption for PAC algorithms

Motivation: Direct Comparison to Conformal Prediction

Conformal Prediction

- Predictor form:

$$\hat{C}(x) = \left\{ y \in \mathcal{Y} \;\middle|\; f(x, y) \geq \tau \right\}$$

- Guarantee: a coverage guarantee

$$\mathbb{P}\left\{ y \notin \hat{C}(x) \right\} \leq \varepsilon$$

- Assumption: exchangeable or i.i.d.

Selective Prediction

- Predictor form:

$$\hat{S}(x) = \begin{cases} \hat{y}(x) & \text{if } f(x, \hat{y}(x)) \geq \tau \\ \texttt{IDK} & \text{otherwise} \end{cases}$$

- Guarantee: a false-discovery rate guarantee

$$\mathbb{P}\left\{ y \neq \hat{S}(x) \;\middle|\; \hat{S}(x) \neq \texttt{IDK} \right\} \leq \varepsilon$$

- Assumption: i.i.d.

Algorithm

Idea

- Enumerate all candidate hypotheses (*i.e.,* a set of τs), *i.e.,*

$$f(x_1, \hat{y}(x_1)), \ldots, f(x_i, \hat{y}(x_i)), \ldots, f(x_n, \hat{y}(x_n)) \quad \text{for } (x_i, \cdot) \in Z$$

- For each hypothesis, compute a binomial tail bound, *i.e.,*

$$\mathbb{P}\left\{ y \neq \hat{y}(x) \;\middle|\; f(x, \hat{y}(x)) \geq \tau_i \right\} \leq U_{\text{Binomial}}(\tau_i, \ldots) \quad \text{for } \tau_i = f(x_i, \hat{y}(x_i))$$

- Choose a hypothesis that has the binomial tail bound smaller than ε, *i.e.,*

$$U_{\text{Binomial}}(\tau_i, \ldots) \leq \varepsilon$$

- Minimize τ to maximize efficiency, *i.e.,* recall the definition of the selective classifier:

$$\hat{S}(x) = \begin{cases} \hat{y}(x) & \text{if } f(x, \hat{y}(x)) \geq \tau \\ \texttt{IDK} & \text{otherwise} \end{cases}$$

Algorithm

Algorithm 2 Selective Classifier Learning Algorithm $\mathcal{A}$ [Geifman and El-Yaniv, 2017]

procedure $\mathrm{SC}(f,\ \hat{y},\ Z,\ \varepsilon,\ \delta)$
  $Z \leftarrow \textsc{Sort}_f(Z)$ (▷) Sort Z in an increasing order of $f(x_i, \hat{y}(x_i))$
  $(\underline{i}, \bar{i}) \leftarrow (1, |Z|)$
  for $i = 1$ **to** $\lceil \log_2 |Z| \rceil$ **do**
    $\tau^{(i)} \leftarrow f(x_{\lceil (\underline{i}+\bar{i})/2 \rceil}, \hat{y}(x_{\lceil (\underline{i}+\bar{i})/2 \rceil}))$
    $Z^{(i)} \leftarrow \{(x, y) \in Z \mid f(x, \hat{y}(x)) \geq \tau^{(i)}\}$
    $k^{(i)} \leftarrow \sum_{(x,y) \in Z^{(i)}} \mathbb{1}(y \neq \hat{y}(x))$
    $U^{(i)} \leftarrow U_{\mathsf{Binom}}(k^{(i)}; |Z^{(i)}|, \delta / \lceil \log_2 |Z| \rceil)$
    if $U^{(i)} \leq \varepsilon$ **then**
      $\bar{i} \leftarrow \lceil (\underline{i} + \bar{i})/2 \rceil$
    else
      $\underline{i} \leftarrow \lceil (\underline{i} + \bar{i})/2 \rceil$
    end if
  end for
  return $\tau^{(i)}, U^{(i)}$
end procedure

FDR Guarantee

Theorem ([Geifman and El-Yaniv, 2017])

For any f and $\mathcal{D}$, we have

$$\mathbb{P}\left\{ y \neq \hat{y}(x) \;\middle|\; f(x, \hat{y}(x)) \geq \hat{\tau} \right\} \leq \hat{U}$$

with probability at least $1 - \delta$, where the probability is taken over $Z \sim \mathcal{D}^n$ and $(\hat{\tau}, \hat{U}) = \mathcal{A}(Z)$.

- $\hat{U} \leq \varepsilon$ may not be true

Conclusion

- Selective prediction could be a good alternative for conformal prediction.
- Selective prediction may not satisfy the PAC guarantee.

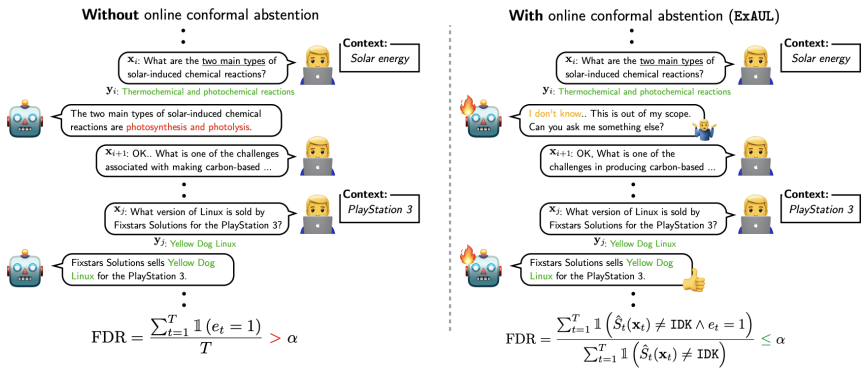
Outline

1. Introduction
2. Conformal Prediction
   - Offline Conformal Prediction
   - Online Conformal Prediction with Full Feedback
   - Online Conformal Prediction with Partial Feedback
3. Conformal Abstention
   - Offline Conformal Abstention
   - Online Conformal Abstention Full Feedback
   - Online Conformal Abstention Partial Feedback

Outline

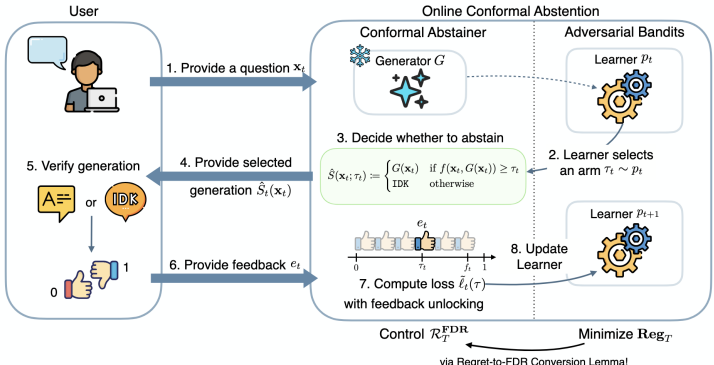
1. Introduction
2. Conformal Prediction
   - Offline Conformal Prediction
   - Online Conformal Prediction with Full Feedback
   - Online Conformal Prediction with Partial Feedback
3. Conformal Abstention
   - Offline Conformal Abstention
   - Online Conformal Abstention Full Feedback
   - Online Conformal Abstention Partial Feedback

OCA with Adversarial Bandit Feedback [Lee et al., 2026]



- A base generator does not satisfy a desired level of hallucination in terms of the false discovery rate (FDR).
- A base generator with conformal abstention provides a desired level of hallucination.

Online Learning



Learning Protocol:

1. Observe a question x_t
2. Generate an answer $G(x_t)$
3. Decide to select the answer or not, *i.e.*, $\mathcal{S}(x_t) = G(x_t)$ or IDK
4. Observe feedback e_t
5. Update a learner for conformal abstention

Conformal Abstainer (or Selective Predictor)

Conformal Abstainer

$$\hat{S}_t(x_t) = \begin{cases} G_t(x_t) & \text{if } f_t(x_t, G_t(x_t)) \geq \tau_t \\ \texttt{IDK} & \text{otherwise} \end{cases}$$

- "selective predictor" is more widely used but we will use "conformal abstainer" to highlight its guaranteeing property.
- x_t: an example or a question
- G_t: a predictor or a generator
- f_t: a scoring function, *i.e.,* whether $G_(x_t)$ conforms to x_t.
- x_t, f_t, G_t, τ_t: all can change over time – but we will fix some, *i.e.,* f, G, afterwords.
- τ_t: the main learning parameter

Goodness (or Trustworthiness) Metric

FDR metrics

$$\text{(Empirical FDR)} \quad \mathbf{FDR}_T := \frac{\sum_{t=1}^{T} \mathbb{1}(\hat{S}_t(\mathbf{x}_t) \neq \texttt{IDK}) \cdot e_t}{\sum_{t=1}^{T} \mathbb{1}(\hat{S}_t(\mathbf{x}_t) \neq \texttt{IDK})} \leq \alpha$$

$$\text{(FDR Risk)} \quad \mathcal{R}_T^{\mathbf{FDR}} := \sum_{t=1}^{T} \Big[\mathbb{1}(\hat{S}_t(\mathbf{x}_t) \neq \texttt{IDK}) \cdot e_t - \alpha \mathbb{1}(\hat{S}_t(\mathbf{x}_t) \neq \texttt{IDK}) \Big]$$

- What's the meaning of FDR?
- α: a desired FDR level
- We wish to find $\hat{S}_t$ that satisfies $\mathbf{FDR}_T \leq \alpha$
- The FDR risk considers a desired FDR level α in its definition.
- But, we need to consider a secondary metric as well (why?):

$$\mathbf{Ineff}_T := \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}(\hat{S}_t(\mathbf{x}_t) = \texttt{IDK})$$

How to Learn $\hat{S}_t$?

online learning with partial feedback

Learning $\hat{S}_t$ = Choosing the best τ_t = Choosing the best arm in "bandit problems"

- Here, we assume τ_t is a finely quantized real value between 0 and 1.
- Do you know bandit problems?

Problem Reduction

	online conformal abstention	adversarial bandits
models	conformal abstainers $\mathcal{H}$	finite arms $\mathcal{H}$
feedback	e_t	$\ell_t(\tau_t, \alpha)$
metric	$\mathbf{FDR}_T$ and $\mathbf{Ineff}_T$	$\mathbf{Reg}_T$

- We leverage adversarial bandit problems to address our problem

Goal of Adversarial Bandits

Regret

$$\mathbf{Reg}_T \coloneqq \sum_{t=1}^{T} \ell_t(\tau_t) - \min_{\tau \in \mathcal{H}} \sum_{t=1}^{T} \ell_t(\tau)$$

- T: a number of turns or interactions
- $\mathcal{H}$: a set of arms
- ℓ_t: a loss function at time t
- τ_t: an arm selected by a learner at time t
- The goal of the adversarial bandit problem is to find a learner to choose τ_t that minimizes the regret.
- How to find such learner?

Connection Between Regret and FDR?

Lemma (Regret-to-FDR Conversion Lemma)

Let $T \in \mathbb{N}$ and $\alpha \in (0,1)$. For any (x_t, y_t) sequences, leading to any loss sequences ℓ_t (following some forms) with any $a_t : \mathcal{H} \to [0,1]$ and $e_t \in [0,1]$, we have

$$\mathcal{R}_T^{\mathbf{FDR}} \leq \frac{T(1 - \mathbf{Ineff}_T) + (1+\lambda)\mathbf{Reg}_T}{\lambda} = \mathcal{O}(T/\lambda + \mathbf{Reg}_T) = \mathcal{O}(\sqrt{T}),$$

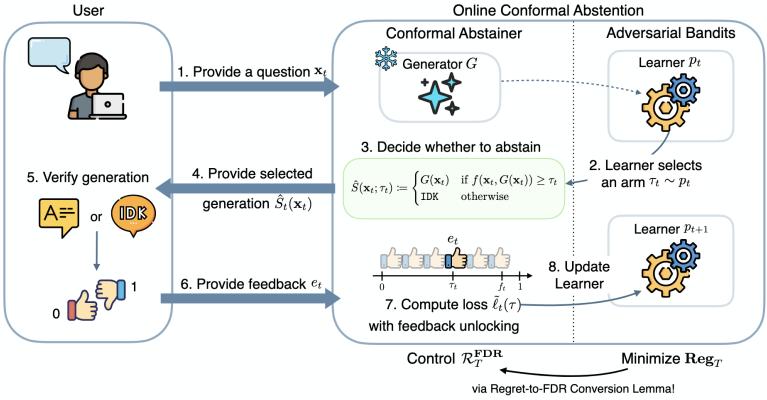
where the second equality holds if we take $\lambda = \sqrt{T}$ and $\mathbf{Reg}_T = \mathcal{O}(\sqrt{T})$.

- Takeaway messages
  1. The regret and the FDR are connected!
  2. We can use any regret minimizer for minimizing the FDR risk!
- We assume a special form of loss as follows:

$$\ell_t(\tau, \alpha) := \frac{a_t(\tau) + \lambda d_t(\tau, \alpha)}{1+\lambda} \in [0,1]$$

  - The loss balances between inefficiency $a_t(\tau) := \mathbb{1}(\hat{S}(x_t;\tau) = \texttt{IDK})$ and one-step FDR risk $d_t(\tau,\alpha) := \mathbb{1}(\hat{S}(x_t;\tau) \neq \texttt{IDK}) \cdot e_t - \alpha \mathbb{1}(\hat{S}(x_t;\tau) \neq \texttt{IDK}) + \alpha$.

Method Overview



- Flow of our method:
  1. (problem reduction) online conformal abstention (OCA) to adversarial bandits
  2. (algorithm) propose a novel bandit method that leverages a special structure of OCA
  3. (regret guarantee) prove that the proposed method achieves a bounded regret.
  4. (FDR guarantee) achieves the FDR risk bound via the regret-to-FDR conversion lemma.

Recap

Takeaway

For certified defenses, we need to know basic ingredients of learning theory.

Reference I

O. Bastani, V. Gupta, C. Jung, G. Noarov, R. Ramalingam, and A. Roth. Practical adversarial multivalid conformal prediction. *Advances in Neural Information Processing Systems*, 35: 29362–29373, 2022.

Y. Geifman and R. El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

I. Gibbs and E. Candès. Adaptive conformal inference under distribution shift, 2021.

M. Lee, Y. Jung, and S. Park. Online conformal abstention for factuality control under adversarial bandit feedback, 2026. URL `https://arxiv.org/abs/2506.14067`.

J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer, 2002.

Reference II

R. J. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas. Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32:2530–2540, 2019.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

J. Yang, K. Kim, and S. Park. Online conformal prediction with adversarial semi-bandit feedback via regret minimization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=RMWcdp5IUy`.