Lecture 13: Conformal Prediction

CIS 7000: Trustworthy Machine Learning

Spring 2024

Homework 2

- Covers distribution shift and uncertainty quantification
  - Written homework focused on theoretical understanding
- Due Monday, March 18

Agenda

- Conformal prediction under distribution shift
- Composing conformal prediction sets
- Conformal structured prediction
- Uniform conformal prediction

Distribution Shift

- Given calibration data from the **source distribution** $p(x, y)$

- Want to perform well on a shifted **target distribution** $q(x, y)$:

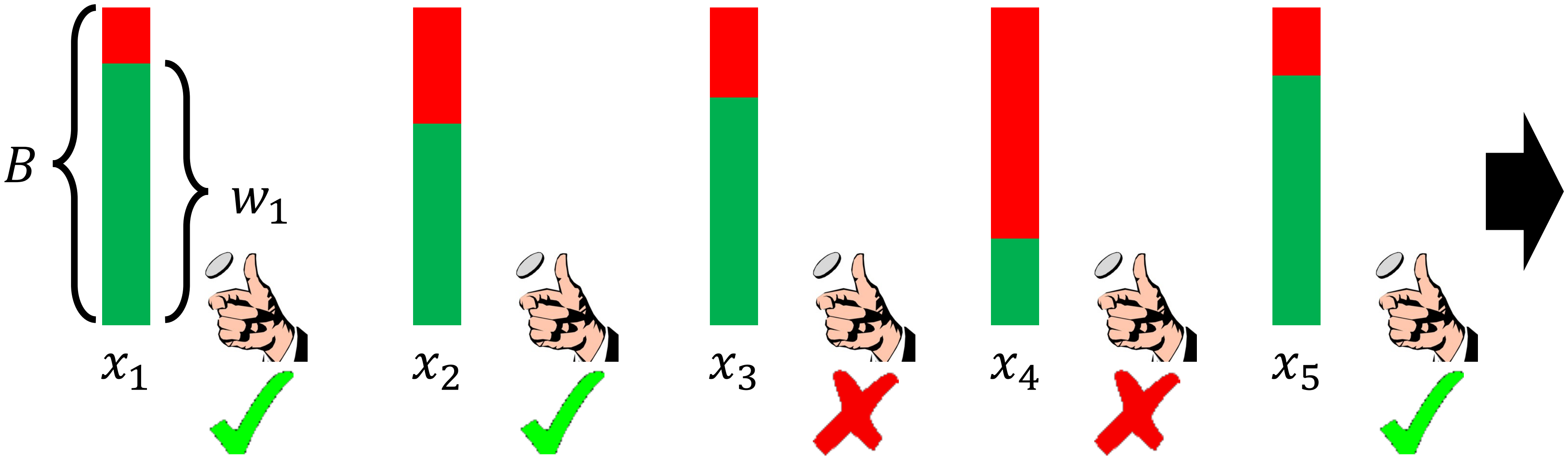
$$\Pr_{p(Z_{\text{cal}})}\left[\Pr_{q(x,y^*)}\left[y^* \in \tilde{f}_{\hat{\tau}(Z_{\text{val}})}(x)\right] \geq 1 - \epsilon\right] \geq 1 - \delta$$

- **Assumptions**
  - Given importance weight intervals $w_i \in \left[w_i^{\text{low}}, w_i^{\text{hi}}\right]$ for each $(x_i, y_i^*) \in Z_{\text{val}}$
  - Can be derived in the unsupervised domain adaptation setting under covariate shift and label shift assumptions
  - Importance weights are bounded: $w(x, y^*) \leq B$ (can be relaxed)

Case 1: Known Importance Weights

- Assume w_i is known for each $(x_i, y_i^*) \in Z_{\text{val}}$

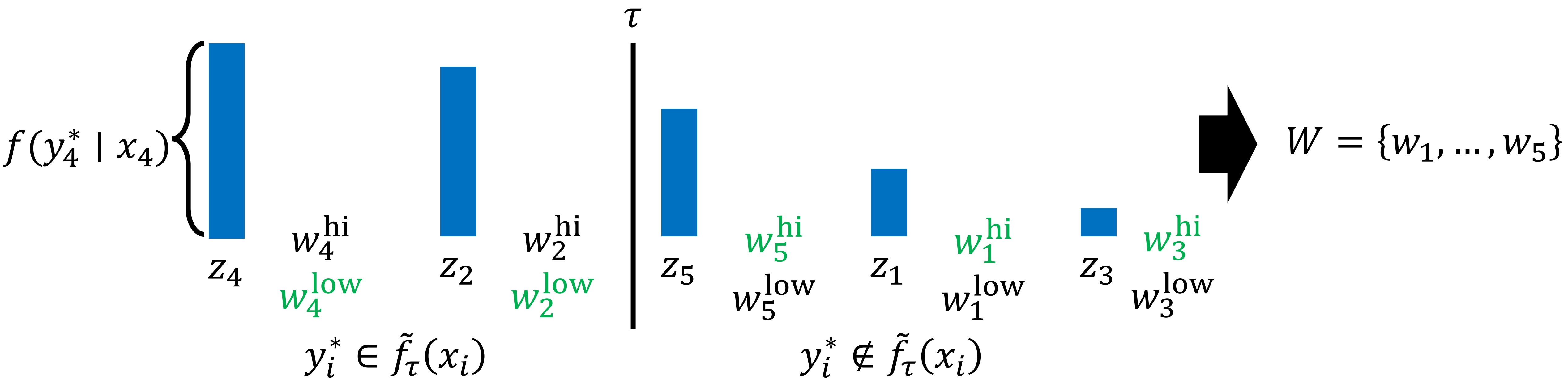
- **Algorithm**
  - **Step 1:** Use rejection sampling to convert $Z_{\text{val}} \sim p$ to $Z'_{\text{val}} \sim q$
  - **Step 2:** Construct PAC prediction set using Z'_{val}



$$Z'_{\text{val}} = \left\{ \begin{array}{l} (x_1, y_1^*), \\ (x_2, y_2^*), \\ (x_5, y_5^*) \end{array} \right\}$$

Case 2: Importance Weight Intervals

- Assume an interval $w_i \in \left[w_i^{\mathrm{low}}, w_i^{\mathrm{hi}}\right]$ is known for each $(x_i, y_i^*) \in Z_{\mathrm{val}}$
- **Algorithm**
  - **Step 1:** Choose the most conservative importance weight $w_i \in \left[w_i^{\mathrm{low}}, w_i^{\mathrm{hi}}\right]$
  - **Step 2:** Construct PAC prediction set using Z_{val} and $\{w_i\}_i$

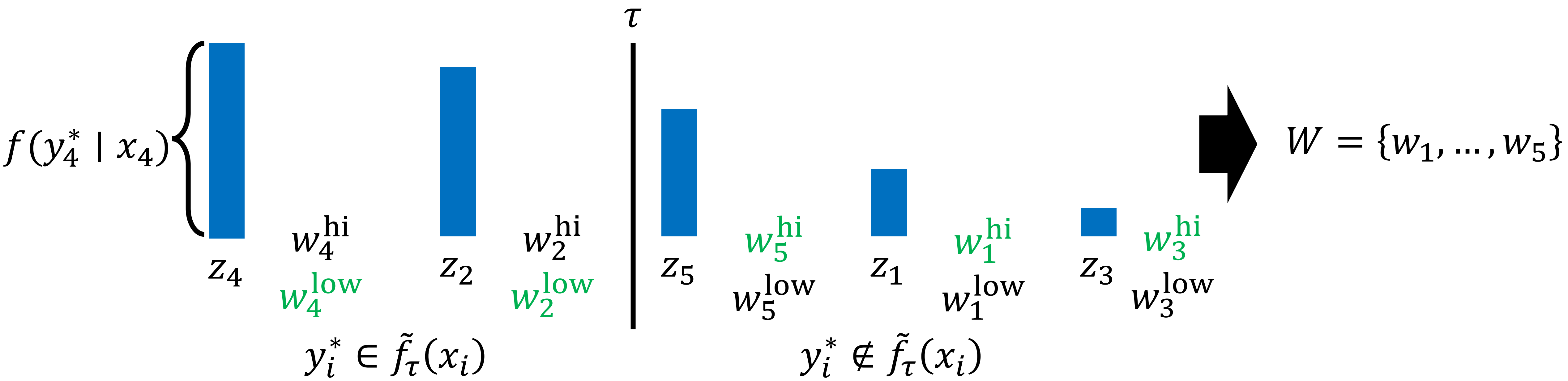


How to Compute $\boldsymbol{\tau}$?

- We have an algorithm that can evaluate a given τ
  - **Idea:** Do binary search on τ to find the best one
  - **Problem:** The algorithm is random!

Case 2: Importance Weight Intervals

- Assume an interval $w_i \in [w_i^{\mathrm{low}}, w_i^{\mathrm{hi}}]$ is known for each $(x_i, y_i^*) \in Z_{\mathrm{val}}$
- **Algorithm**
  - **Step 1:** Choose the most conservative importance weight $w_i \in [w_i^{\mathrm{low}}, w_i^{\mathrm{hi}}]$
  - **Step 2:** Construct PAC prediction set using Z_{val} and $\{w_i\}_i$

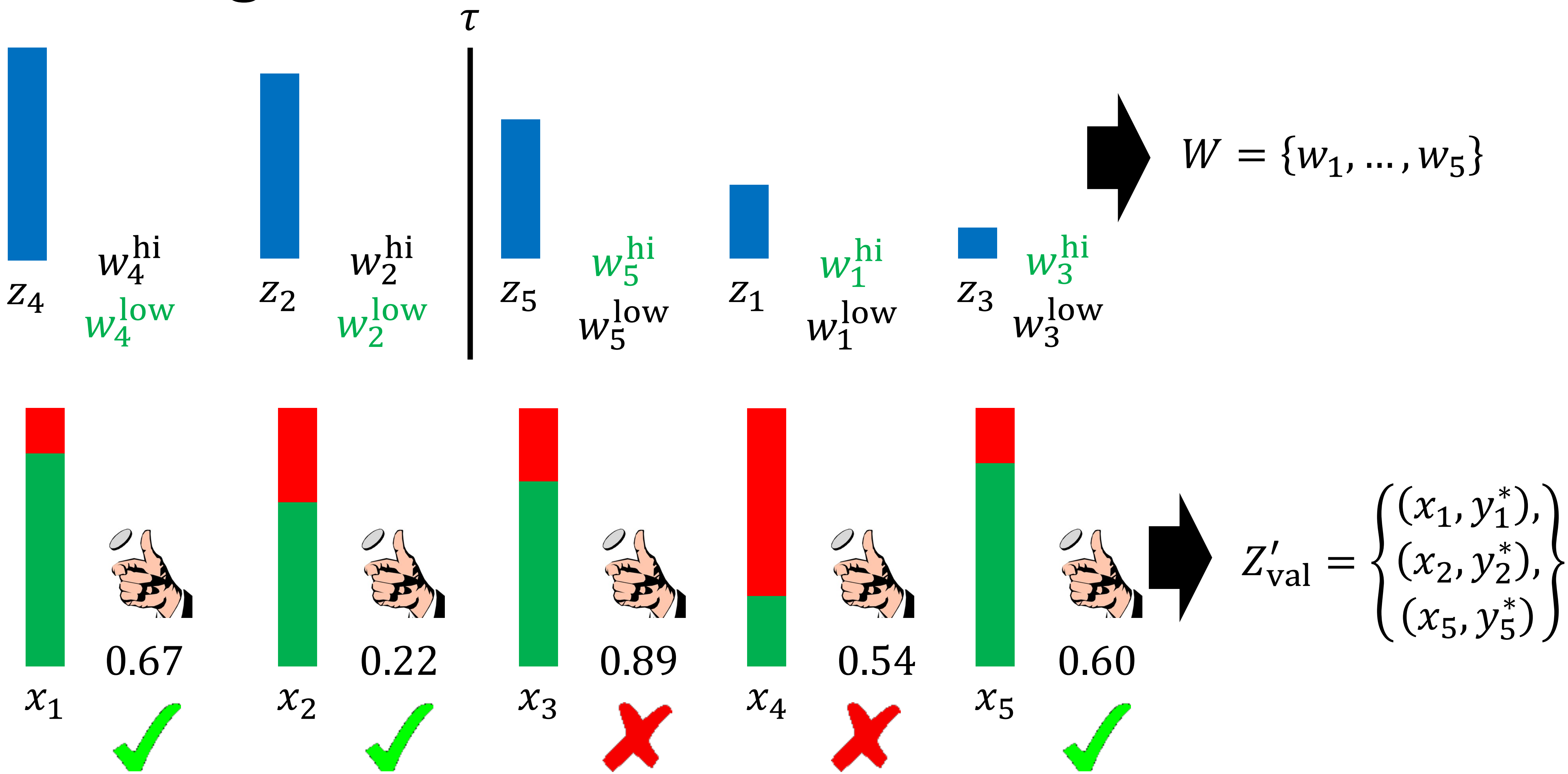


How to Compute $\boldsymbol{\tau}$?

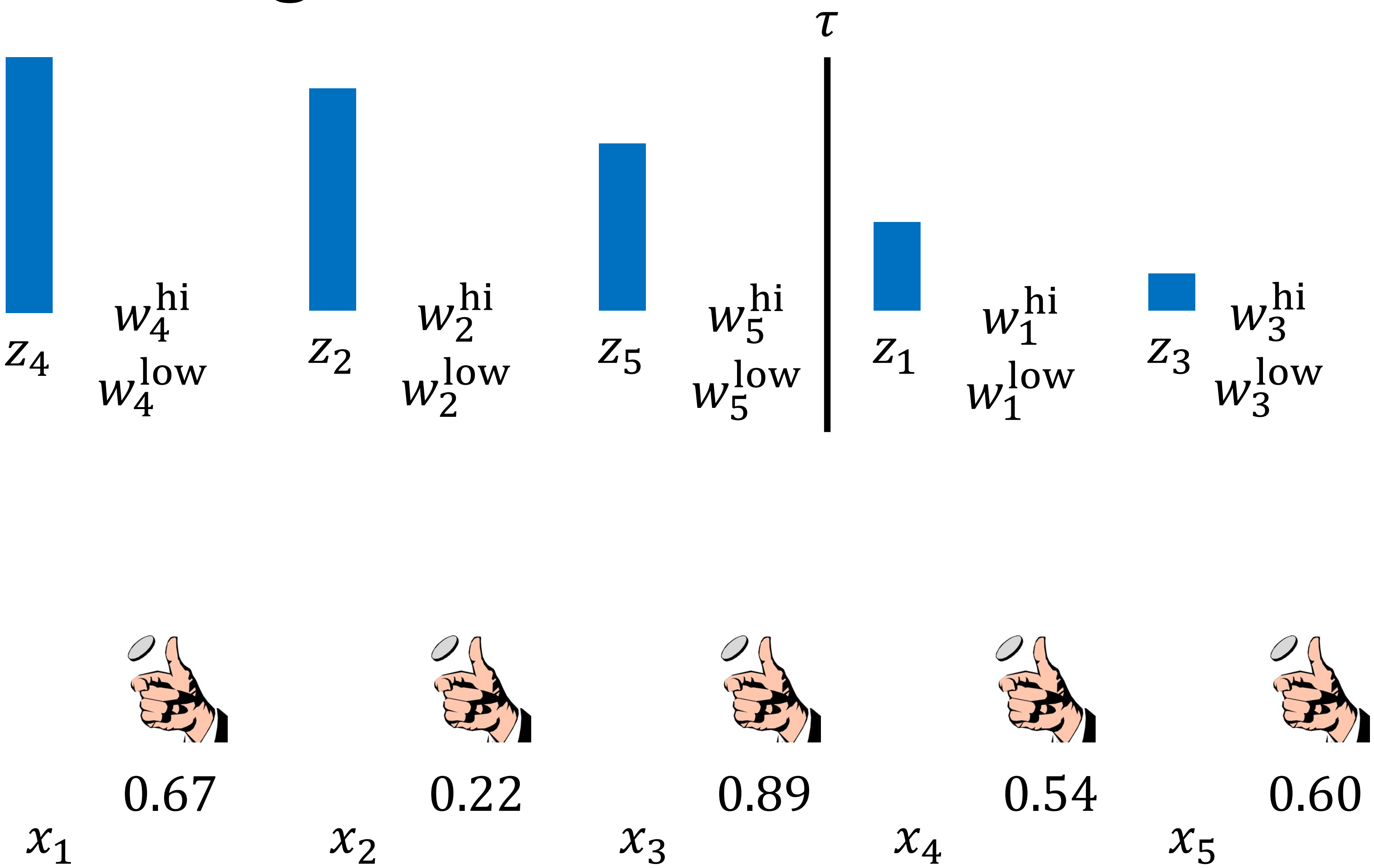
- We have an algorithm that can evaluate a given τ
  - **Idea:** Do binary search on τ to find the best one
  - **Problem:** The algorithm is random!

- **Solution:** Sample randomness **before** running binary search

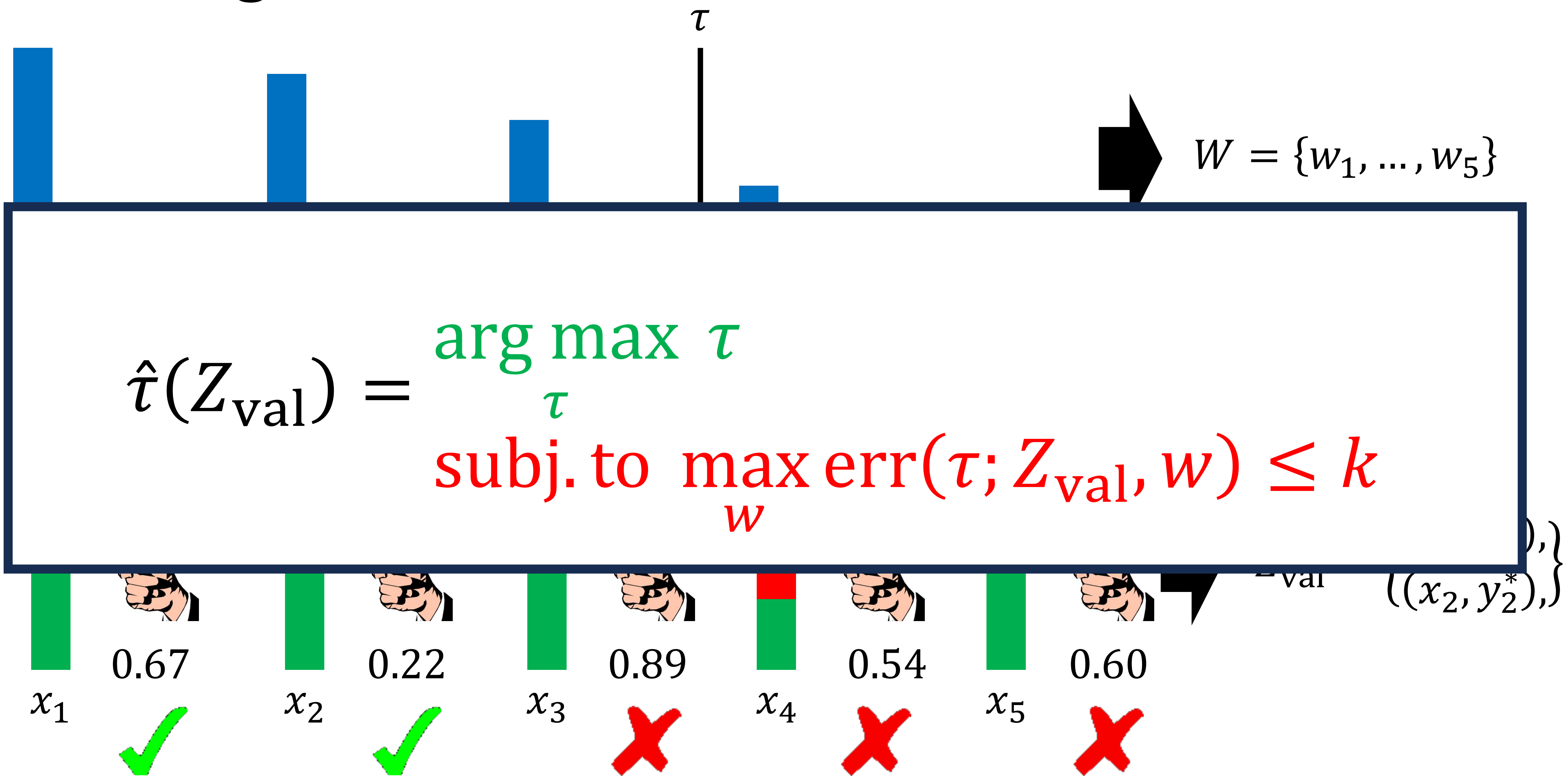
Full Algorithm



Full Algorithm



Full Algorithm



$$\hat{\tau}(Z_{\text{val}}) = \underset{\tau}{\arg\max}\ \tau$$
$$\text{subj. to } \max_{w} \text{err}(\tau; Z_{\text{val}}, w) \leq k$$

Theoretical Guarantees

- **Theorem**
  - Assume $w(x_i) \in \left[w_i^{\mathrm{low}}, w_i^{\mathrm{hi}}\right]$ for all $i \in \{1, \dots, n\}$
  - Then, $f_{\hat{\tau}(Z_{\mathrm{val}})}$ is an (ϵ, δ)-PAC prediction set with respect to q

Examples on DomainNet

Example x	$\hat{C}_{\text{CP-IID}}(x)$	$\hat{C}_{\text{RSCP-WIW}}(x)$	Example x	$\hat{C}_{\text{CP-IID}}(x)$	$\hat{C}_{\text{RSCP-WIW}}(x)$
	$\{\widehat{\text{raccoon}}\}$	$\{\text{owl}, \widehat{\text{raccoon}}\}$		$\{\text{angel}, \widehat{\text{harp}}\}$	$\{\text{angel}, \text{cello}, \widehat{\text{harp}}, \text{microphone}, \text{piano}, \text{violin}\}$
	$\{\widehat{\text{wine bottle}}\}$	$\{\text{bread}, \text{grapes}, \widehat{\text{wine bottle}}, \text{wine glass}\}$		$\{\widehat{\text{shark}}, \text{snorkel}\}$	$\{\text{dolphin}, \widehat{\text{shark}}, \text{snorkel}, \text{submarine}, \text{whale}\}$
	$\{\widehat{\text{campfire}}\}$	$\{\widehat{\text{campfire}}, \text{ocean}, \text{star}, \text{tent}\}$		$\{\text{coffee cup}, \widehat{\text{cup}}\}$	$\{\text{coffee cup}, \widehat{\text{cup}}, \text{mug}, \text{teapot}\}$
	$\{\widehat{\text{ocean}}\}$	$\{\text{hurricane}, \widehat{\text{ocean}}, \text{square}, \text{tornado}\}$		$\{\widehat{\text{brain}}\}$	$\{\widehat{\text{brain}}, \widehat{\text{fish}}, \text{lion}, \text{lollipop}, \text{sea turtle}\}$

Results

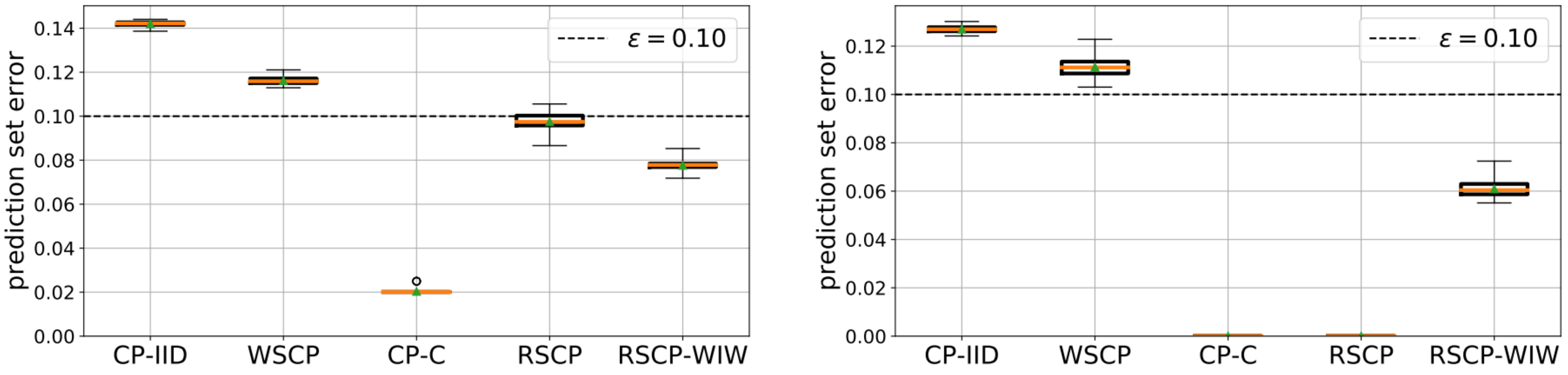


Figure 1: Error under natural rate shift by DomainNet for All $\rightarrow$ Sketch (left), and ImageNet-C synthetic perturbations to ImageNet (right), over 100 random trials, with $m = 50,000$ (for DomainNet) and $m = 20,000$ (for ImageNet), $\varepsilon = 0.1$, and $\delta = 10^{-5}$.

Obtaining IW Intervals

- How do we get importance weight intervals $w_i \in \left[w_i^{\mathrm{low}}, w_i^{\mathrm{hi}}\right]$?
  - **Covariate shift:** Need to use heuristics
  - **Label shift:** Can get exact intervals

IW Intervals for Label Shift

- Recall that $w = C^{-1}q$, where

$$C_{ij} = \mathbb{P}_P[f(x) = i, y = j]$$

IW Intervals for Label Shift

- Recall that $w = C^{-1}q$, where

$$C_{ij} = \mathbb{P}_P[f(x) = i, y = j] \approx |Z|^{-1} \sum_{(x,y)\in Z} 1(f(x) = i, y = j) = \hat{C}_{ij}$$

$$q_i = \mathbb{P}_Q[f(x) = i]$$

IW Intervals for Label Shift

- Recall that $w = C^{-1}q$, where

$$C_{ij} = \mathbb{P}_P[f(x) = i, y = j] \approx |Z|^{-1} \sum_{(x,y)\in Z} 1(f(x) = i, y = j) = \hat{C}_{ij}$$

$$q_i = \mathbb{P}_Q[f(x) = i] \approx |X|^{-1} \sum_{x\in X} 1(f(x) = i) = \hat{q}_i$$

IW Intervals for Label Shift

- **Hoeffding's inequality**
  - Let $b_1, \ldots, b_n \sim_{\text{i.i.d.}} \mathrm{Bernoulli}(\mu)$ be samples
  - Let $\hat{\mu} = n^{-1} \sum_{k=1}^{n} b_k$ be the empirical mean
  - Then, with probability at least $1 - \delta$, we have

$$|\hat{\mu} - \mu| \leq \sqrt{\frac{\log(2/\delta)}{2n}}$$

IW Intervals for Label Shift

- With probability $\geq 1 - \delta$, the following hold individually:

$$\left|\hat{C}_{ij} - C_{ij}\right| \leq \sqrt{\frac{\log(2/\delta)}{2|Z|}}$$

IW Intervals for Label Shift

- With probability $\geq 1 - \delta$, the following hold individually:

$$\left|\hat{C}_{ij} - C_{ij}\right| \leq \sqrt{\frac{\log(2/\delta)}{2|Z|}} \quad \text{and} \quad |\hat{q}_i - q_i| \leq \sqrt{\frac{\log(2/\delta)}{2|X|}}$$

- **Union bound:** If $\Pr[A_i] \geq 1 - \delta_i$ for all i, then $\Pr[\bigwedge_i A_i] \geq 1 - \sum_i \delta_i$

- With probability $\geq 1 - (d + d^2)\delta$, all of the following hold:

$$\left|\hat{C}_{ij} - C_{ij}\right| \leq \sqrt{\frac{\log(2/\delta)}{2|Z|}} \quad \text{and} \quad |\hat{q}_i - q_i| \leq \sqrt{\frac{\log(2/\delta)}{2|X|}}$$

IW Intervals for Label Shift

- With probability $\geq 1 - \delta$, the following hold individually:

$$\left|\hat{C}_{ij} - C_{ij}\right| \leq \sqrt{\frac{\log(2/\delta)}{2|Z|}} \quad \text{and} \quad |\hat{q}_i - q_i| \leq \sqrt{\frac{\log(2/\delta)}{2|X|}}$$

- **Union bound:** If $\Pr[A_i] \geq 1 - \delta_i$ for all i, then $\Pr[\wedge_i A_i] \geq 1 - \sum_i \delta_i$

- With probability $\geq 1 - (d + d^2)\delta$, all of the following hold:

$$C_{ij} \in \left[\hat{C}_{ij} - \sqrt{\frac{\log(2/\delta)}{2|Z|}}, \hat{C}_{ij} + \sqrt{\frac{\log(2/\delta)}{2|Z|}}\right] \quad \text{and} \quad q_i \in \left[\hat{q}_i - \sqrt{\frac{\log(2/\delta)}{2|X|}}, \hat{q}_i + \sqrt{\frac{\log(2/\delta)}{2|X|}}\right]$$

IW Intervals for Label Shift

- We need to bound $|\widehat{w}_i - w_i|$, where $w = C^{-1}q$ and $\widehat{w} = \hat{C}^{-1}\hat{q}$
- **Strategy:** Abstract interpretation!
  - If we have C and q, then we could compute w using Gaussian elimination
  - If we have intervals around the entries of C and q, then we can run Gaussian elimination on these intervals using abstract interpretation

IW Intervals for Label Shift

- **Recall:** Given a function $f: \mathbb{R}^d \to \mathbb{R}$, its corresponding **abstract transformer** is a function $\hat{f}: \widehat{\mathbb{R}}^d \to \widehat{\mathbb{R}}$ such that if

$$x' = f(x_1, \dots, x_d) \quad \text{and} \quad \wedge_{i=1}^{d} x_i \in \gamma(\hat{x}_i)$$

then we have

$$x' \in \gamma\left(\hat{f}(\hat{x}_1, \dots, \hat{x}_d)\right)$$

IW Intervals for Label Shift

- Let $\widehat{\mathbb{R}} = \mathbb{R} \times \mathbb{R}$ be the **interval domain**
  - $\alpha(\{r_1, \ldots, r_k\}) = \left(\min_i r_i, \max_i r_i\right) \in \widehat{\mathbb{R}}$
  - $\gamma\big((a, b)\big) = [a, b] \subseteq \mathbb{R}$

- Then, we have:
  - $(a, b) \widehat{+} (c, d) = (a + c, b + d)$
  - $(a, b) \widehat{-} (c, d) = (a - d, b - c)$
  - $(a, b) \widehat{\times} (c, d) = (a \times c, b \times d)$ (assuming everything is non-negative)
  - $(a, b) \mathbin{\widehat{\div}} (c, d) = (a \div d, b \div c)$ (assuming everything is non-negative)

IW Intervals for Label Shift

- **Step 1:** Compute $\hat{C}$ and $\hat{q}$
- **Step 2:** Use Hoeffding to obtain intervals $\hat{C}_{\min}, \hat{C}_{\max}, \hat{q}_{\min}, \hat{q}_{\max}$ such that $\hat{C}_{\min} \leq C \leq \hat{C}_{\max}$ and $\hat{q}_{\min} \leq q \leq \hat{q}_{\max}$ with high probability
  - Inequalities are interpreted elementwise
- **Step 3:** Run Gaussian elimination using abstract interpretation to obtain intervals $\widehat{w}_{\min}, \widehat{w}_{\max}$
  - By abstract interpretation guarantee, we have $\widehat{w}_{\min} \leq w \leq \widehat{w}_{\max}$
- **Step 4:** Run PAC conformal prediction with IW intervals

IW Intervals for Label Shift

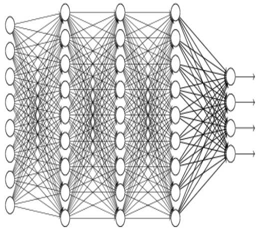
- **Theorem**: $f_{\hat{\tau}(Z_{\mathrm{val}})}$ is an $(\epsilon, 2\delta)$-PAC prediction set with respect to q
  - 2δ comes from IW intervals + PAC property (and union bound)

Agenda

- Conformal prediction under distribution shift
- Composing conformal prediction sets
- Conformal structured prediction
- Uniform conformal prediction

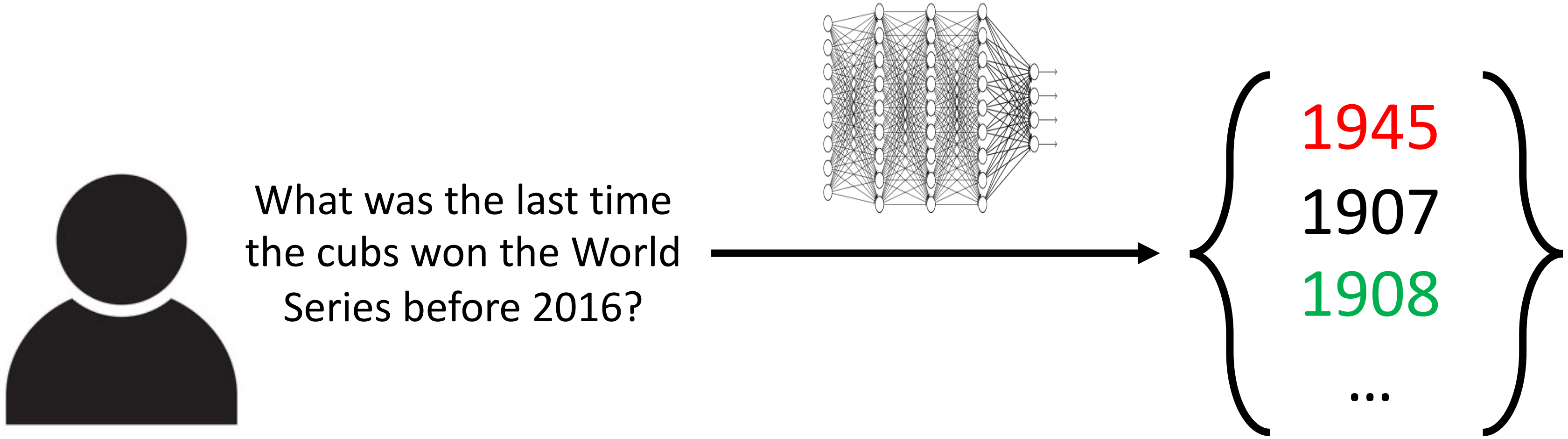
Prediction Sets for Question Answering

What was the last time the cubs won the World Series before 2016?



1945

Prediction Sets for Question Answering

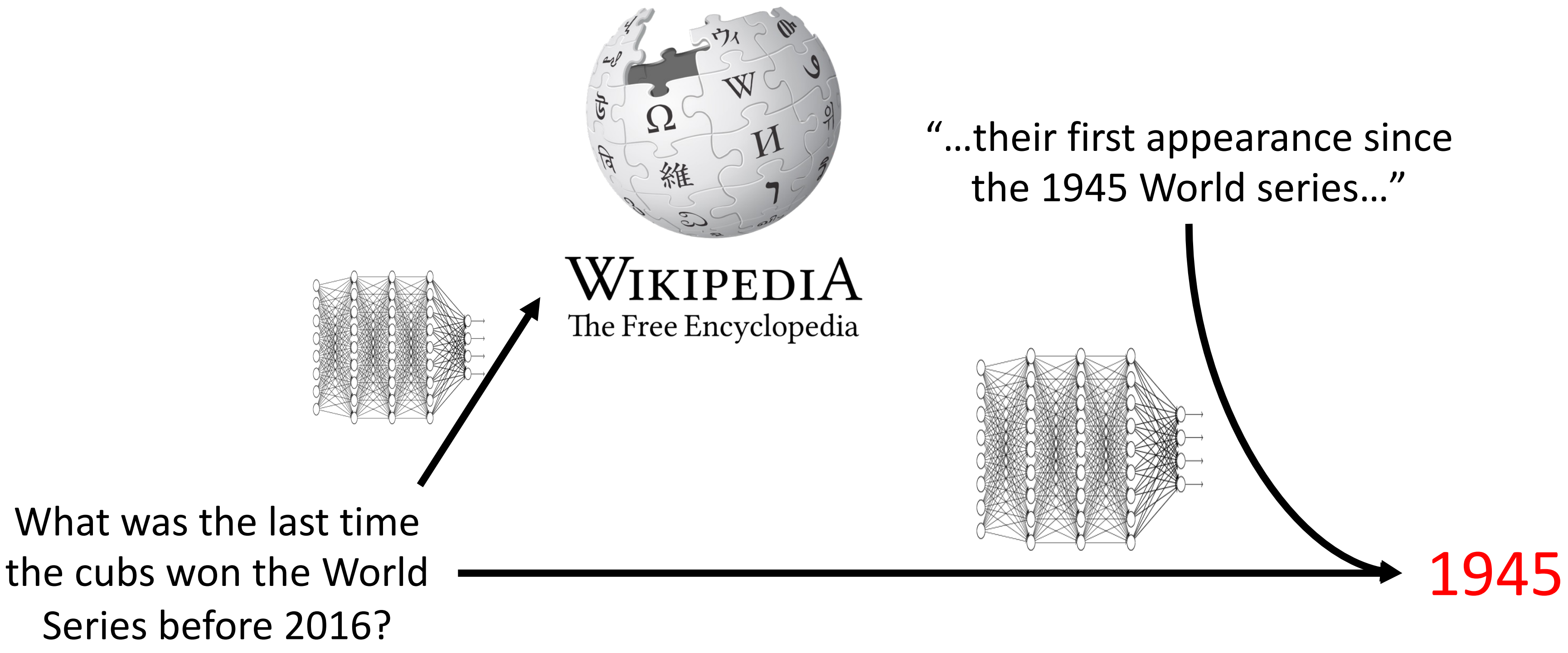


Retrieval Augmented Question Answering

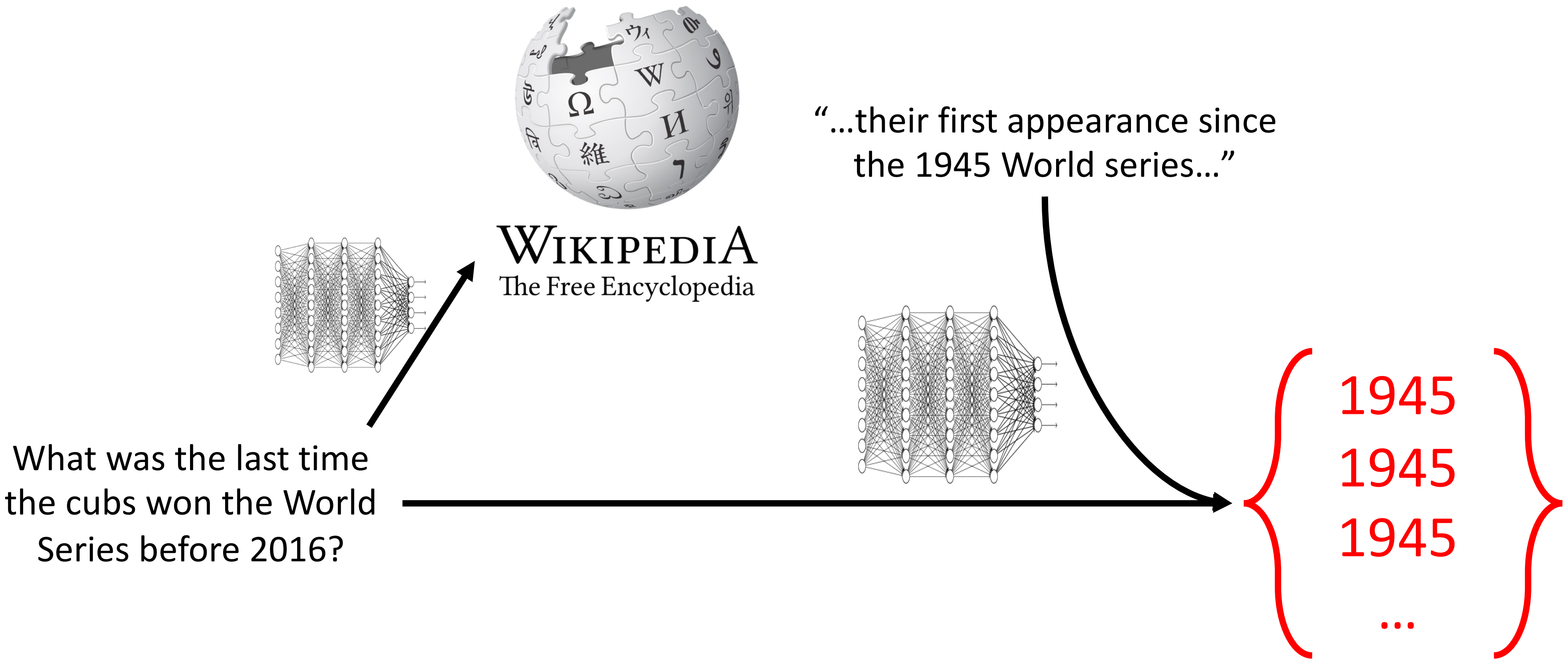
- Many applications of large language models rely on specialized sources of knowledge that are not present in the training data

- **Retrieval augmented question answering**
  - Extract relevant knowledge from knowledge base (e.g., Wikipedia)
  - Incorporate knowledge into query to generative model

Retrieval Augmented Question Answering



Retrieval Augmented Question Answering



Retrieval Augmented Question Answering

How to compose probabilistic guarantees?

$\epsilon_{\text{retrieval}}$

ϵ_{QA}

$\epsilon = \epsilon_{\text{retrieval}} + \epsilon_{\text{QA}}$

Minimizing Prediction Set Size

- **Challenge:** How to choose $\epsilon_{\mathrm{retrieval}}$ and ϵ_{QA}?

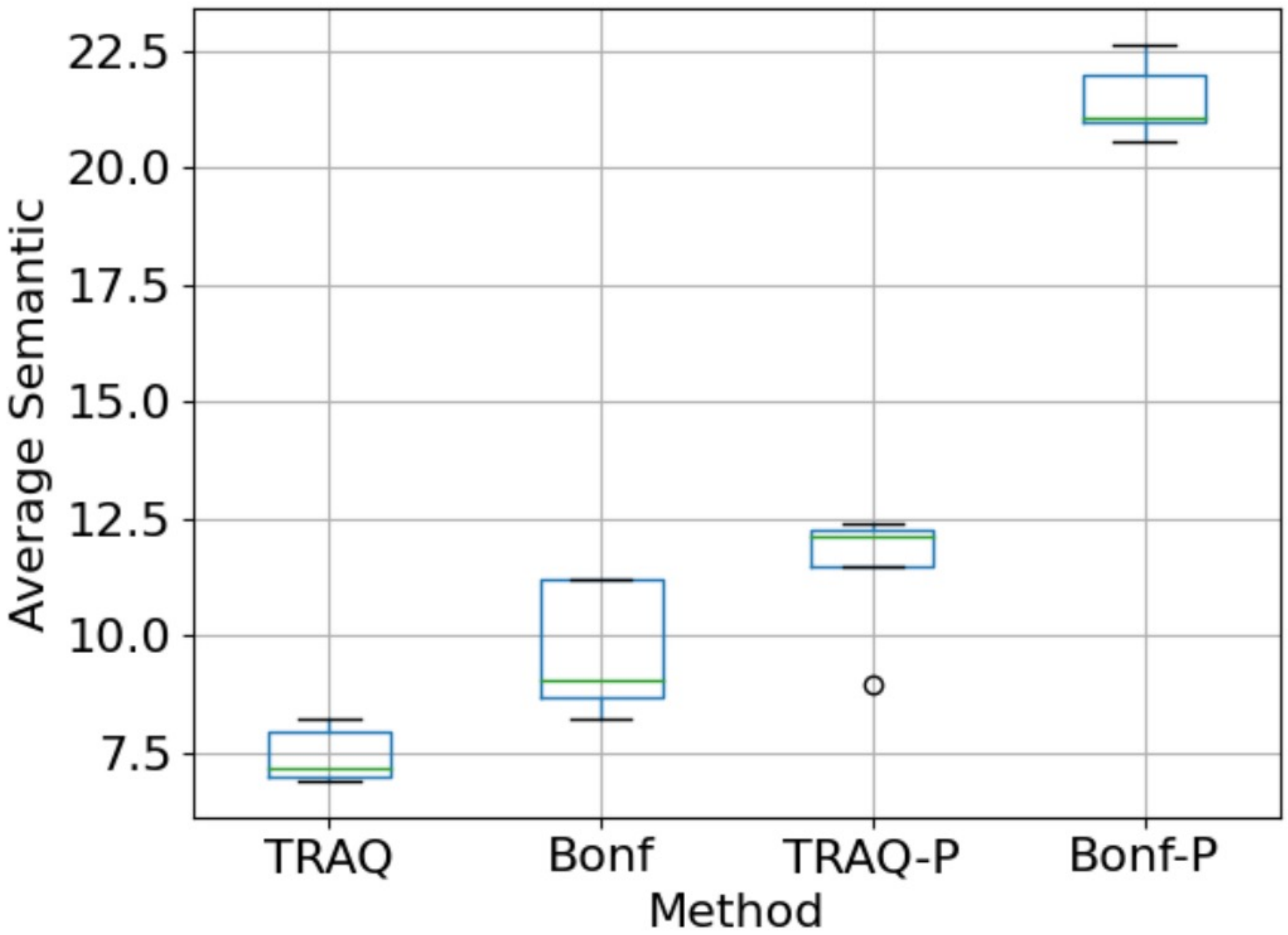
- **Solution:** Optimize them on a held-out **optimization set** Z_{opt}
  - Optimization variables are $\epsilon_{\mathrm{retrieval}}$ and ϵ_{QA}
  - Given a candidate, compute conformal prediction thresholds: $\hat{\tau}_{\mathrm{retrieval}}(Z_{\mathrm{opt}}; \epsilon_{\mathrm{retrieval}})$ and $\hat{\tau}_{\mathrm{QA}}(Z_{\mathrm{opt}}; \epsilon_{\mathrm{QA}})$
  - Objective is expected prediction set size: $\sum_{(x,y^*)\in Z_{\mathrm{opt}}} \left| C_{\hat{\tau}_{\mathrm{retrieval}}(Z_{\mathrm{opt}};\epsilon_{\mathrm{retrieval}}),\hat{\tau}_{\mathrm{QA}}(Z_{\mathrm{opt}};\epsilon_{\mathrm{QA}})}(x) \right|$

- We use Bayesian optimization to optimize $\epsilon_{\mathrm{retrieval}}$ and ϵ_{QA}

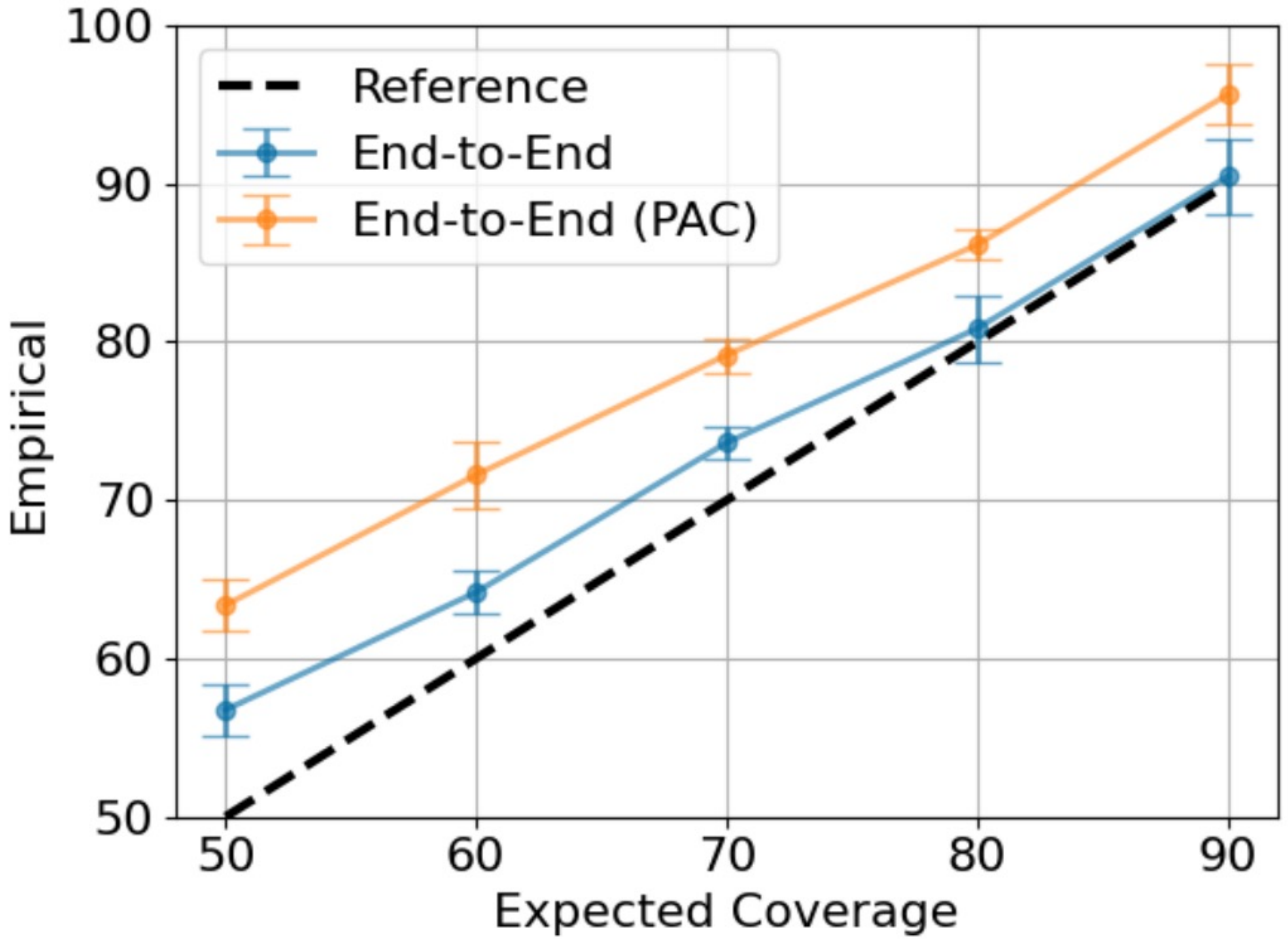
Experimental Results

- **Dataset:** SQuAD question answering dataset
  - Similar results on TriviaQA and Natural Questions
- **Model:** GPT-3.5-Turbo
  - Similar results on Llama 2 7B
- Consider both PAC prediction sets and traditional (marginal) conformal prediction
- **Baseline:** No Bayesian optimization

Prediction Set Size



Coverage

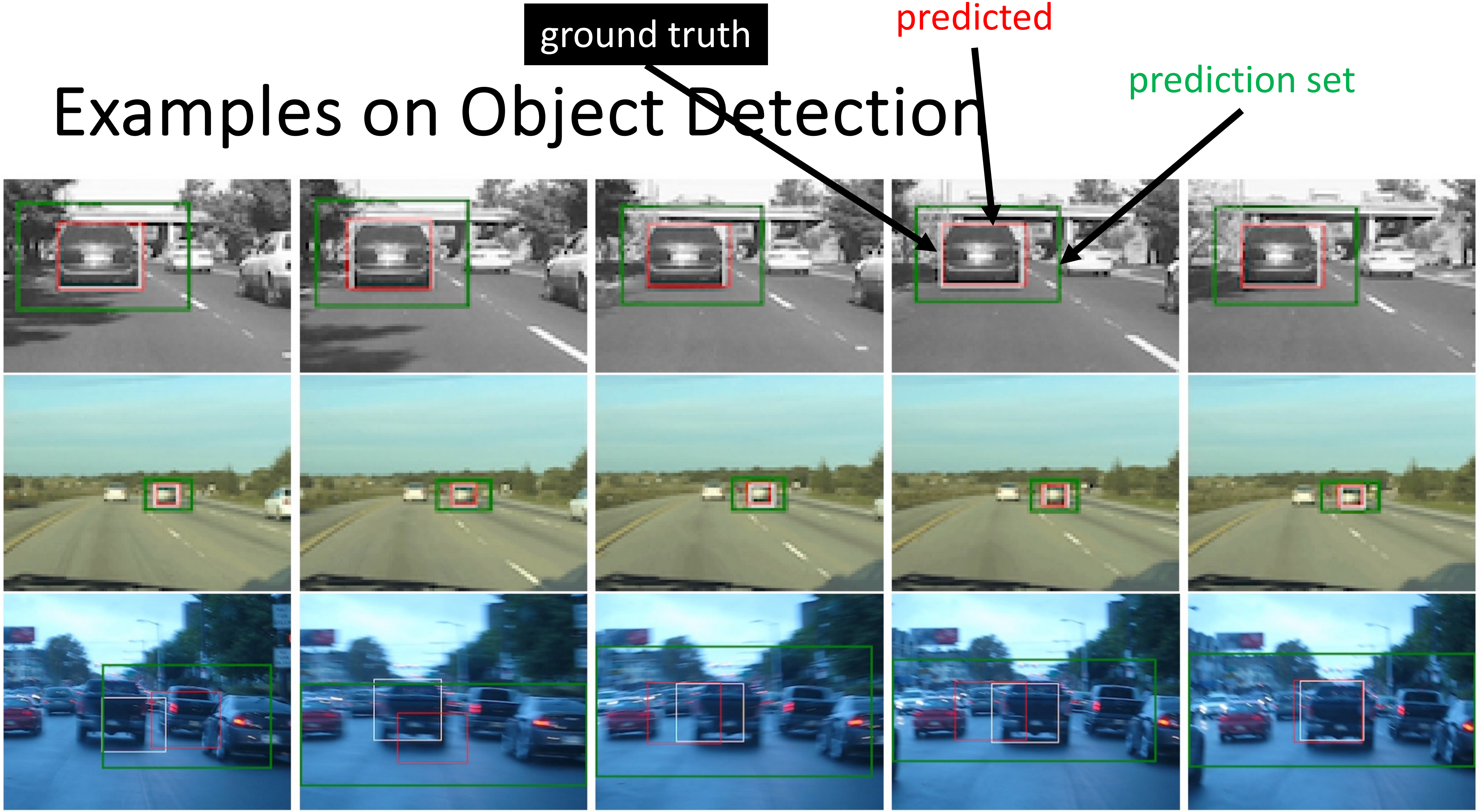

Reference
End-to-End
End-to-End (PAC)
100
90
80
70
60
50
Empirical
50
60
70
80
90
Expected Coverage

Compositional Conformal Prediction

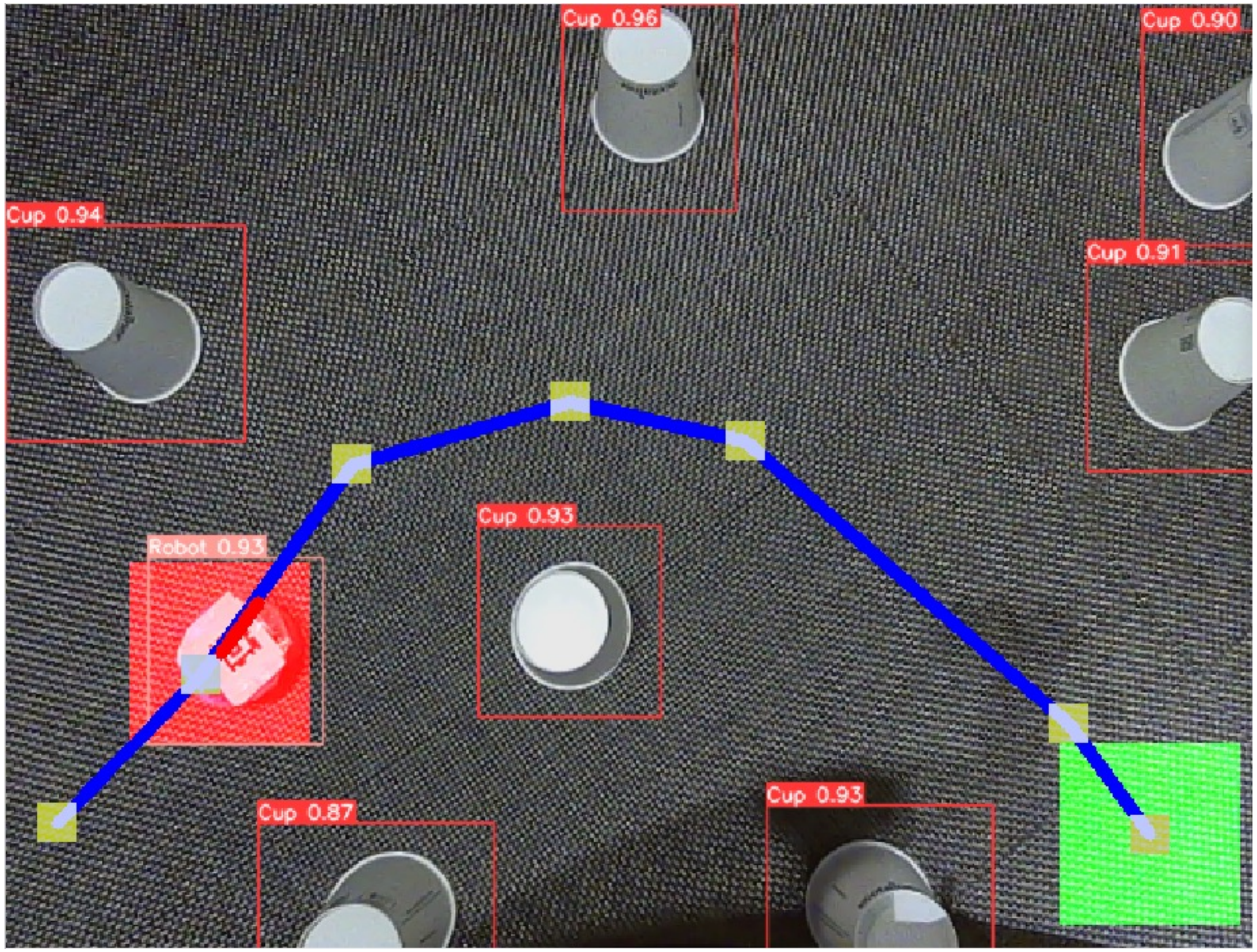
- Approach generalizes to more complex model compositions
  - For more complex data types, can use abstract interpretation to compose

- **Another example:** Object detection
  - Output is obtained by composing region proposal network, bounding box regression network, object classification network
  - Can use combination of previous techniques to obtain prediction sets

Examples on Object Detection



Prediction Sets for Safe Visual Navigation



Agenda

- Conformal prediction under distribution shift
- Composing conformal prediction sets
- Conformal structured prediction
- Uniform conformal prediction

Conformal Prediction for Code Generation

- **True program:**

```
return fib(n-1) + fib(n-2)
```

- **Generated program:**

```
return fib(n-0) + fib(n-3)
```

Conformal Prediction for Code Generation

- **True program:**

```
return fib(n-1) + fib(n-2)
```

- **Generated program:**

```
return fib(n-0) + fib(n-3)
```

$$\left\{\begin{array}{l}
\texttt{return fib(n-0) + fib(n-1)} \\
\texttt{return fib(n-0) + fib(n-2)} \\
\texttt{return fib(n-0) + fib(n-3)} \\
\texttt{return fib(n-0) + fib(n-4)} \\
\texttt{return fib(n-1) + fib(n-1)} \\
\texttt{return fib(n-1) + fib(n-2)} \\
\texttt{return fib(n-1) + fib(n-3)} \\
\texttt{return fib(n-1) + fib(n-4)} \\
\texttt{return fib(n-2) + fib(n-1)} \\
\texttt{return fib(n-2) + fib(n-2)} \\
\texttt{return fib(n-2) + fib(n-3)} \\
\texttt{...}
\end{array}\right\}$$

Challenge for Code Generation

- Code generation produces a structured output
- Naïve prediction set might contain thousands of programs!
- **Idea:** Compact representation of set of programs
  - Implicitly represent prediction set as a **partial program**
  - Partial program represents set of all programs that can be obtained by completing it in some way

Prediction Sets as Partial Programs

- **True program:**

```
return fib(n-1) + fib(n-2)
```

- **Generated program:**

```
return fib(n-0) + fib(n-3)
```

```
return fib(n-0) + fib(n-1)
return fib(n-0) + fib(n-2)
return fib(n-0) + fib(n-3)
return fib(n-0) + fib(n-4)
return fib(n-1) + fib(n-1)
return fib(n-1) + fib(n-2)
return fib(n-1) + fib(n-3)
return fib(n-1) + fib(n-4)
return fib(n-2) + fib(n-1)
return fib(n-2) + fib(n-2)
return fib(n-2) + fib(n-3)
...
```

Prediction Sets as Partial Programs

- **True program:**

```
return fib(n-1) + fib(n-2)
```

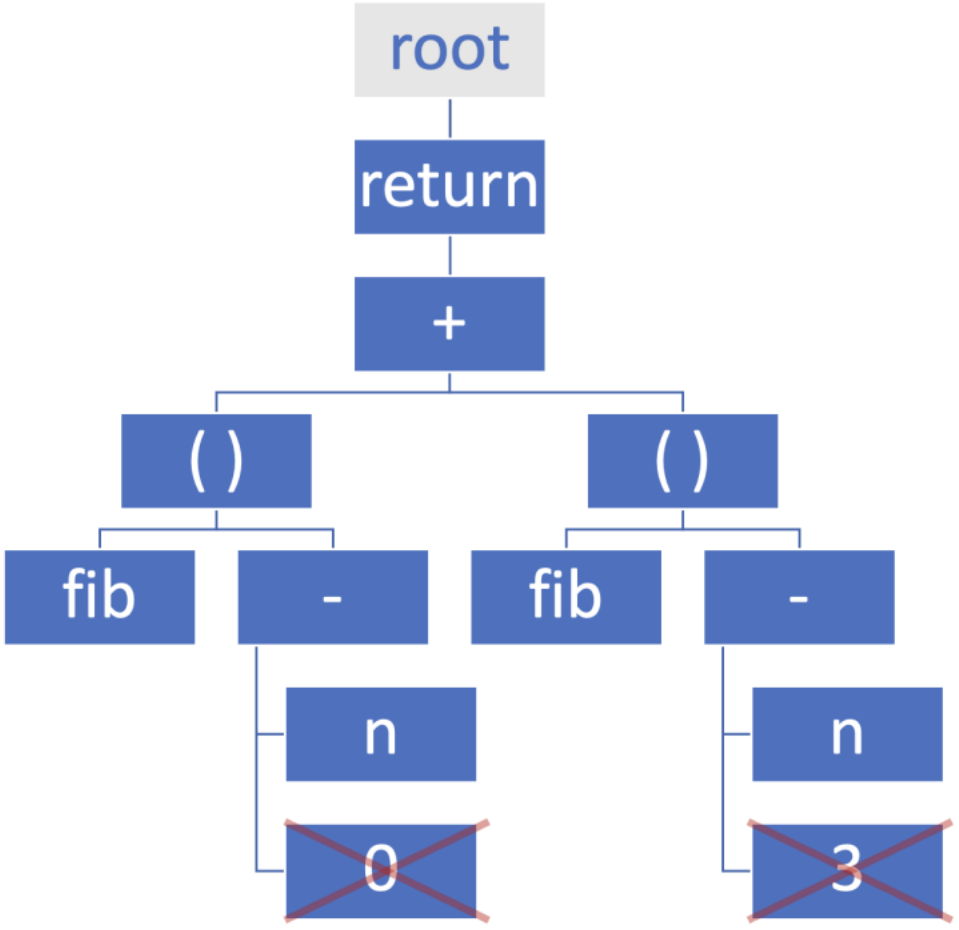
- **Generated program:**

```
return fib(n-??) + fib(n-??)
```

```
return fib(n-0) + fib(n-3)
```

Prediction Sets as Partial Programs

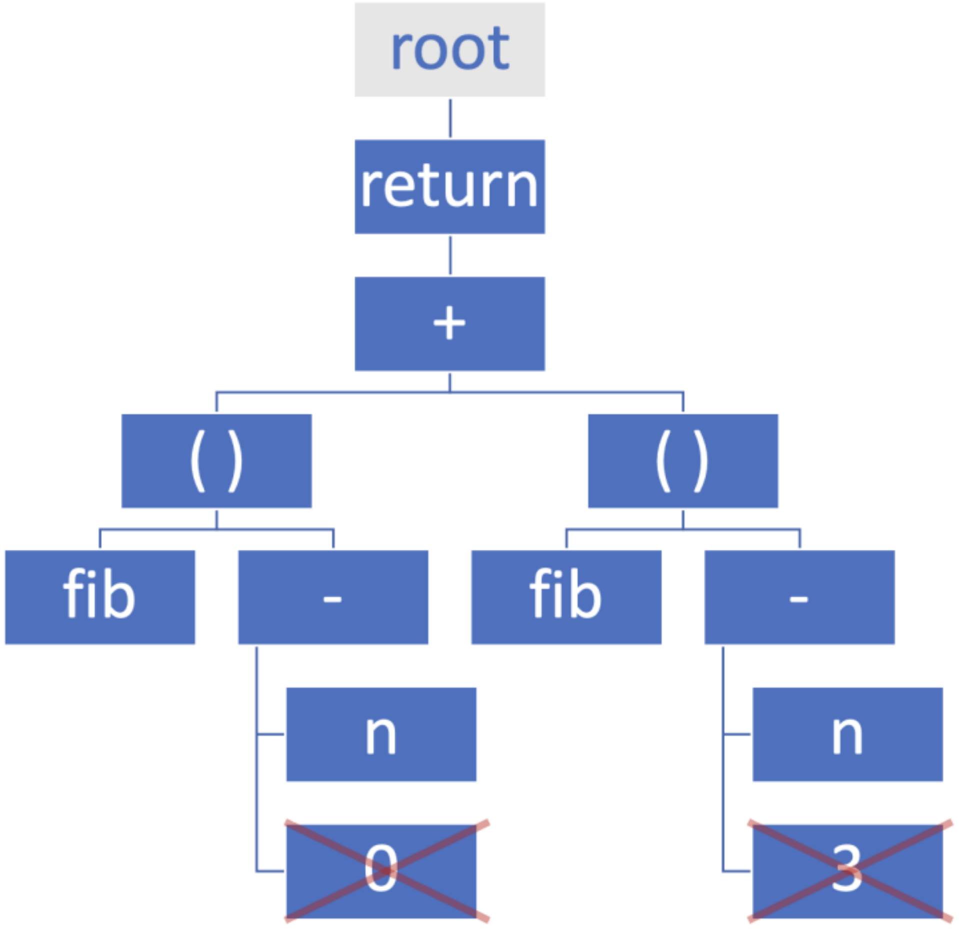
- **Strategy:** Remove AST nodes until probability mass removed exceeds τ



`return fib(n-0) + fib(n-3)`

Prediction Sets as Partial Programs

- **Strategy:** Remove AST nodes until probability mass removed exceeds τ




```
return fib(n-??) + fib(n-??)
```

Computing Prediction Sets

- Formulate as optimization problem:

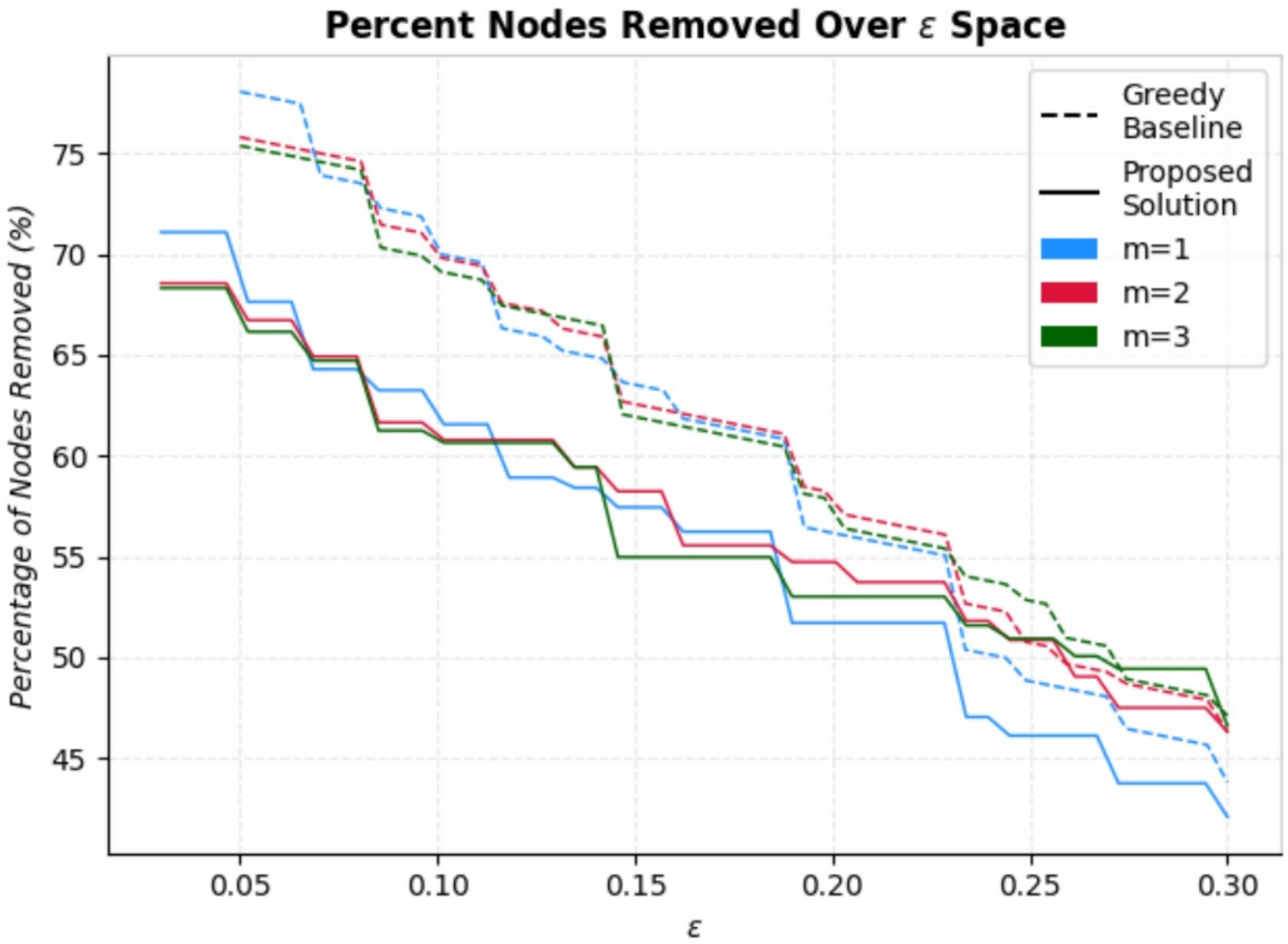
$$\sum_{v \in V} \alpha_{i,v} \leq m \qquad (\forall i \in [k])$$
$$\alpha_{i,v} \to \beta_{i,v} \qquad (\forall v \in V,\ i \in [k])$$
$$\beta_{i,v} \to \beta_{i,v'} \qquad (\forall (v, v') \in E)$$
$$\beta_{i,v} \to \alpha_{i,v} \vee \beta_{i,v'} \qquad (\text{where } (v', v) \in E)$$
$$\beta_{i,v} \to \beta_{i+1,v} \qquad (v \in V,\ \forall i \in \{2, ..., k\})$$
$$\sum_{v \in T} \ell_v \cdot (1 - \beta_{i,v}) \leq \tau_i \qquad (\forall i \in [k])$$

- We additionally impose constraint that number of holes $\leq m$

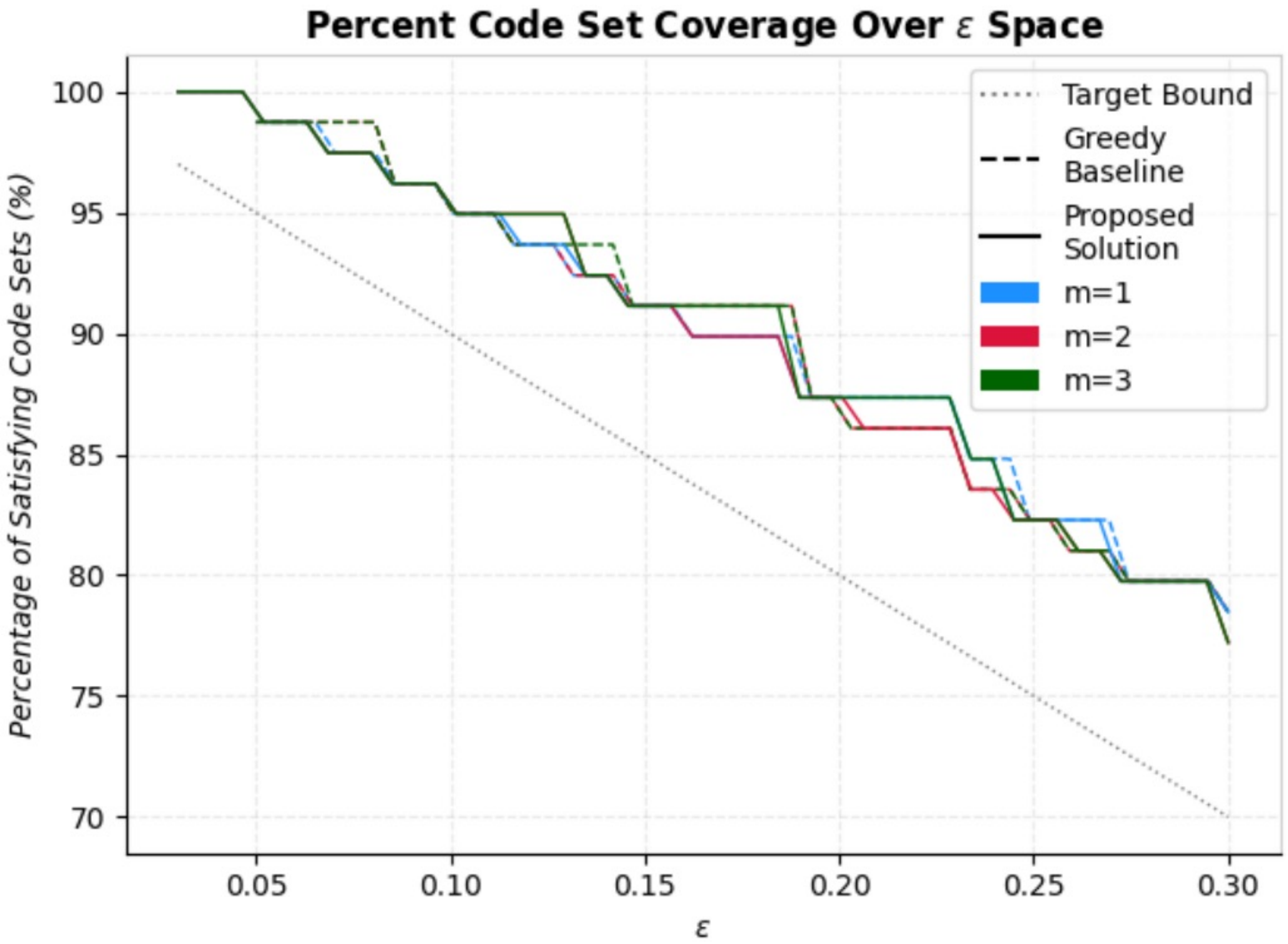
Evaluation

- **Dataset:** APPS program synthesis dataset
  - Similar results on text-to-SQL task
- **Model:** Codex
- **Baseline:** Greedy strategy for constructing prediction sets

Prediction Set Size



Coverage


Percent Code Set Coverage Over ε Space
Percentage of Satisfying Code Sets (%)
100
95
90
85
80
75
70
0.05
0.10
0.15
0.20
0.25
0.30
ε
Target Bound
Greedy Baseline
Proposed Solution
m=1
m=2
m=3

Example on SQL Query

```
SELECT COUNT(*) FROM countries AS t1
JOIN car_makers as t2 on t1.countryid = t2.country
WHERE t1.countryname = "usa";
```

Conformal Structured Prediction

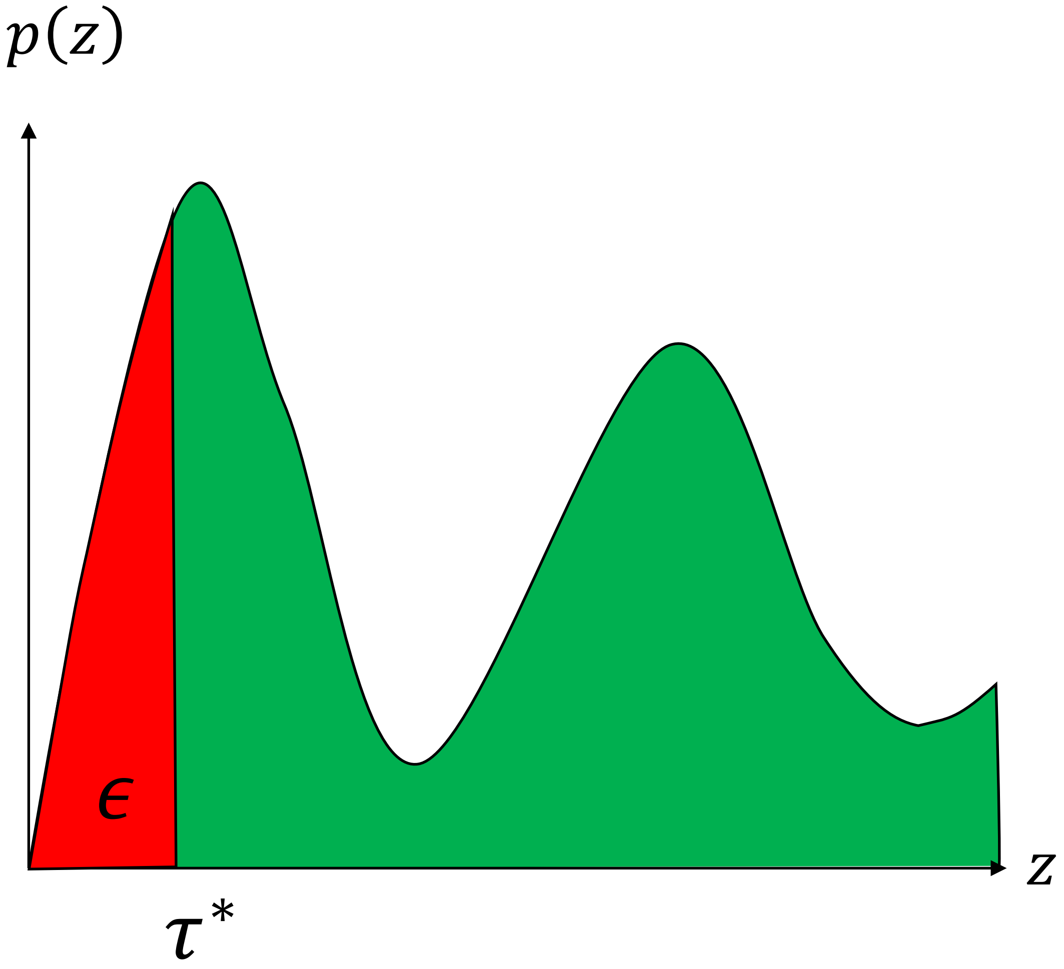
- Approach generalizes to any structured prediction problem

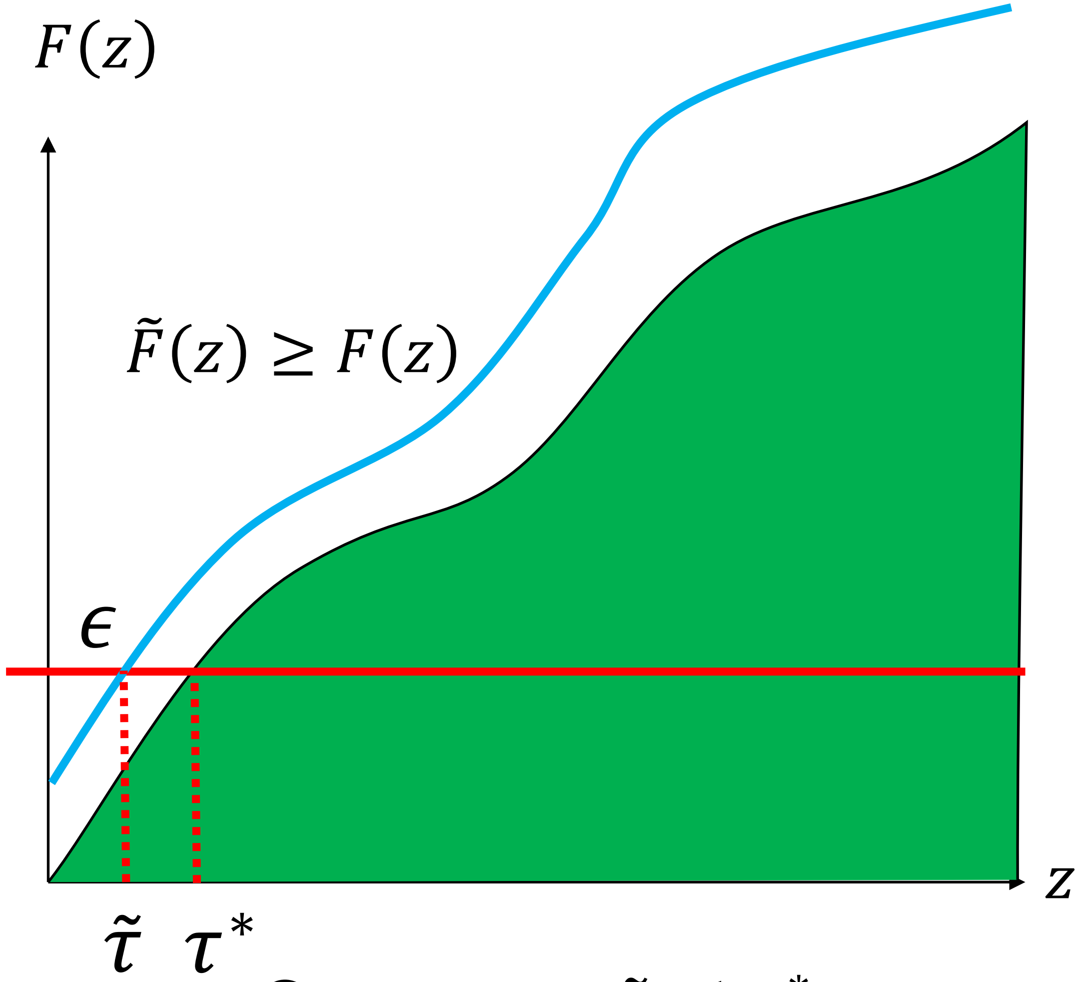
- **Examples**
  - Hierarchical classification
  - Open-ended question answering

Agenda

- Conformal prediction under distribution shift
- Composing conformal prediction sets
- Conformal structured prediction
- Uniform conformal prediction

Recall: Distribution of $z = f(y^* \mid x)$





Guarantee: $\tilde{\tau} \leq \tau^*$

Uniformly Valid Conformal Prediction

- **DKW Inequality (Massart 1990)**
  - Let P be a probability distribution and let $F(x)$ be its CDF
  - Given samples $z_1, \ldots, z_n \sim_{\text{i.i.d.}} P$, the **empirical CDF** is

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1(z_i \leq x)$$

  - **Theorem:** With probability $\geq 1 - \delta$, we have

$$\sup_{x\in\mathbb{R}}\left|\hat{F}(x) - F(x)\right| \leq \sqrt{\frac{\log(2/\delta)}{2n}}$$

Uniformly Valid Conformal Prediction

- **Input**
  - Calibration dataset $Z_{\text{val}} = \{(x_i, y_i^*)\}_{i=1}^n$
  - Error bound δ

- **Step 1:** Construct CDF upper bound

$$\tilde{F}(z) = \frac{1}{n}\sum_{i=1}^{n} 1(z_i \leq z) + \sqrt{\frac{\log(2/\delta)}{2n}}$$

- **Step 2:** Return $\tilde{\tau} = \tilde{F}^{-1}(\epsilon)$ (caveat: need to use pseudoinverse here)

Uniformly Valid Conformal Prediction

- **Original guarantee:** For all $\epsilon, \delta \in \mathbb{R}$, we have

$$\Pr_{Z_{\text{val}}}\left[\Pr_{(x,y^*)}\left[y^* \in \tilde{f}_{\tilde{\tau}(Z_{\text{val}})}(x)\right] \geq 1-\epsilon\right] \geq 1-\delta$$

- **New guarantee:** For all $\delta \in \mathbb{R}$, we have

$$\Pr_{Z_{\text{val}}}\left[\forall \epsilon \,.\, \Pr_{(x,y^*)}\left[y^* \in \tilde{f}_{\tilde{\tau}(Z_{\text{val}})}(x)\right] \geq 1-\epsilon\right] \geq 1-\delta$$

Uniformly Valid Conformal Prediction

Example x	$\epsilon = 0.01$	$\epsilon = 0.02$	$\epsilon = 0.03$	$\epsilon = 0.04$	$\epsilon = 0.05$
	{croquet ball, golf ball, nail, thimble, custard apple}	{golf ball}	{golf ball}	{golf ball}	{golf ball}
	{sulphur-crested cockatoo, Japanese spaniel, Sussex spaniel, bee, pot, vase, cardoon, daisy}	{daisy}	{daisy}	{daisy}	{daisy}
	{tabby, tiger cat, Egyptian cat, lynx}	{tabby, tiger cat, Egyptian cat}	{tabby, tiger cat, Egyptian cat}	{tabby, tiger cat, Egyptian cat}	{tiger cat, Egyptian cat}
	{tabby, tiger cat, Egyptian cat, lynx, leopard, snow leopard, jaguar, tiger, zebra}	{tabby, tiger cat, leopard, jaguar, tiger, zebra}	{tiger cat, jaguar, tiger}	{tiger cat, jaguar, tiger}	{tiger cat, tiger}

Agenda

- Conformal prediction under distribution shift
- Composing conformal prediction sets
- Conformal structured prediction
- Uniform conformal prediction