Online Conformal Prediction with Efficiency Guarantees

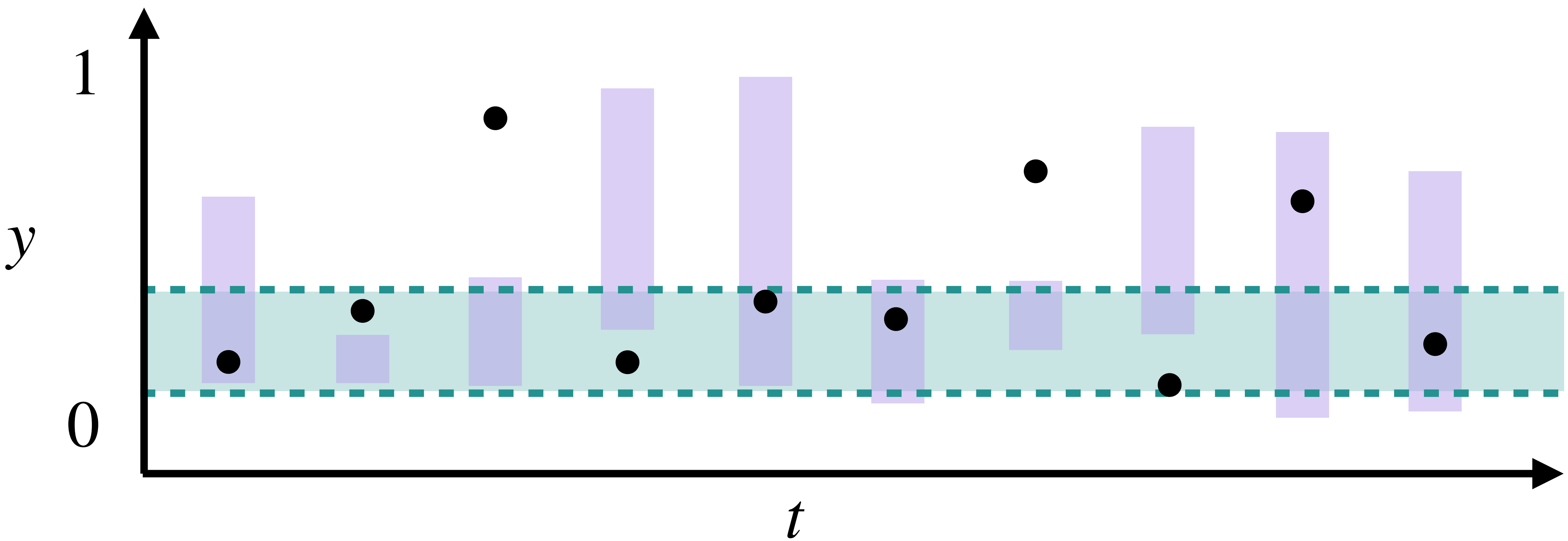
Vaidehi Srinivas

Chicago Junior Theorists Workshop, December 2025

Generating Prediction Sets Online

For each day t, generate **prediction set** $C_t \subseteq [0,1]$
to achieve **coverage** $1 - \alpha$ over $y_t \in [0,1]$

Illustration: $1 - \alpha = 0.7$



Goals:

(1) **Coverage:** capture $1 - \alpha$ fraction of the points

(2) **Efficiency:** play average volume close to the **best fixed interval in hindsight** that achieves coverage $1 - \alpha$

Motivation: Conformal Prediction

Strategy to ensure **reliability** of ML models in practice

Standard ML setup:

See $(x_1, y_1), (x_2, y_2) \ldots, (x_T, y_T)$. Now see x_{T+1}. What is y_{T+1}?

Learning theory answer:

Assume $(x_i, y_i) \sim \mathscr{D}$ i.i.d.. Then, for a reasonable hypothesis class $\mathscr{F}$, can learn function $f^\star \in \mathscr{F}$, $\hat{y} = f^\star(x)$, that has the lowest error on $\mathscr{D}$ **(regression)**

Issues in practice:

(1) (x_i, y_i) are not i.i.d.

(2) If the function $f^\star \in \mathscr{F}$ is bad (high error), want to know now!
Evaluating error at the end of the game is too late to do anything about it
Need **uncertainty quantification**

Usual Strategy

"Wrap" **regression model** with conformal wrapper

Strategy:

(1) Train a regression model $f : X \to Y$ to predict $\widehat{y_i} = f(x_i)$

(2) Measure the error of the prediction according to a hand-chosen **non-conformity score** $s(y, \hat{y}) \in \mathbb{R}$

(3) Estimate τ, the $(1 - \alpha)$th quantile of the non-conformity score online

(4) For new x_i, predict set of y that would make f low-error

$$C_i = \{y \ : \ s(f(x_i), y) \leq \tau\}$$

Pros:

- Ensure coverage with no guarantees required of f (could be a neural network)
- Simple recipe

Cons:

- Need to design non-conformity score by hand for every new setting
- Efficiency (set size) depends a lot on non-conformity score!

  $\rightarrow$ no efficiency guarantees

Bottleneck: Set Size

Requires lots of work to design the non-conformity scores for new settings

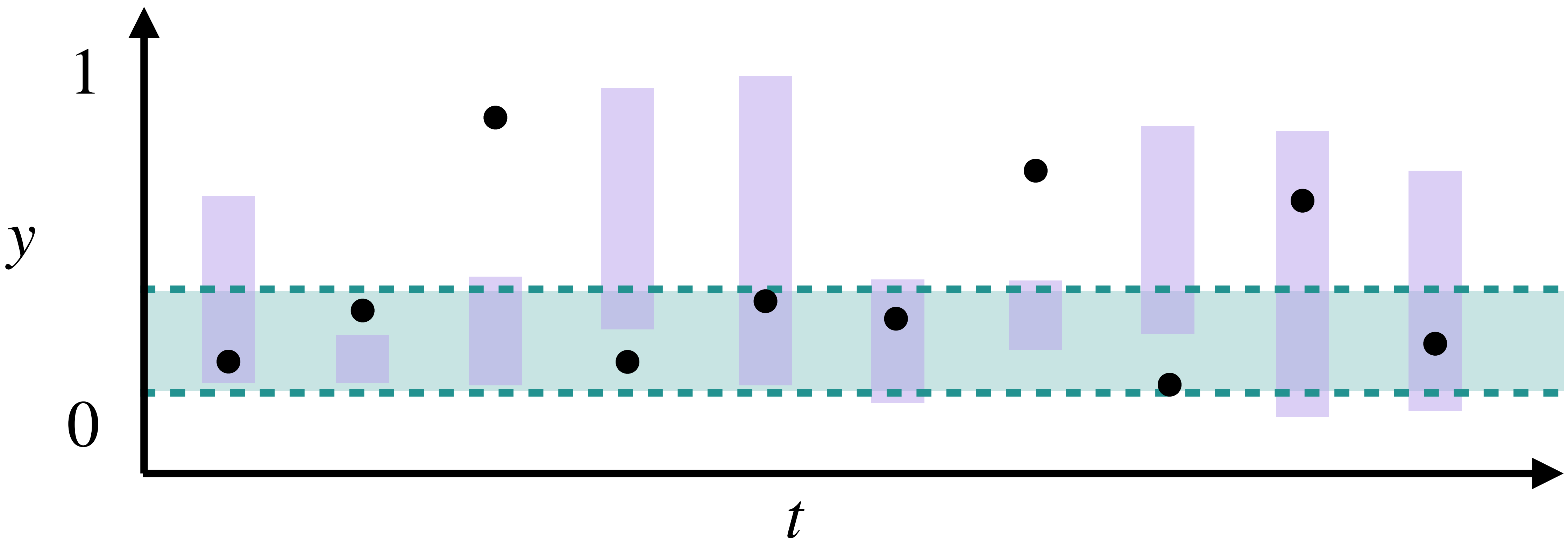
Questions:

- Can we automatically learn the smallest possible prediction sets?
- Is **regression** the right way to approach this problem?

This Work: Theoretical Study

Simplest setting:

Only have $y_i \in [0,1]$ (all features x_i are the same)

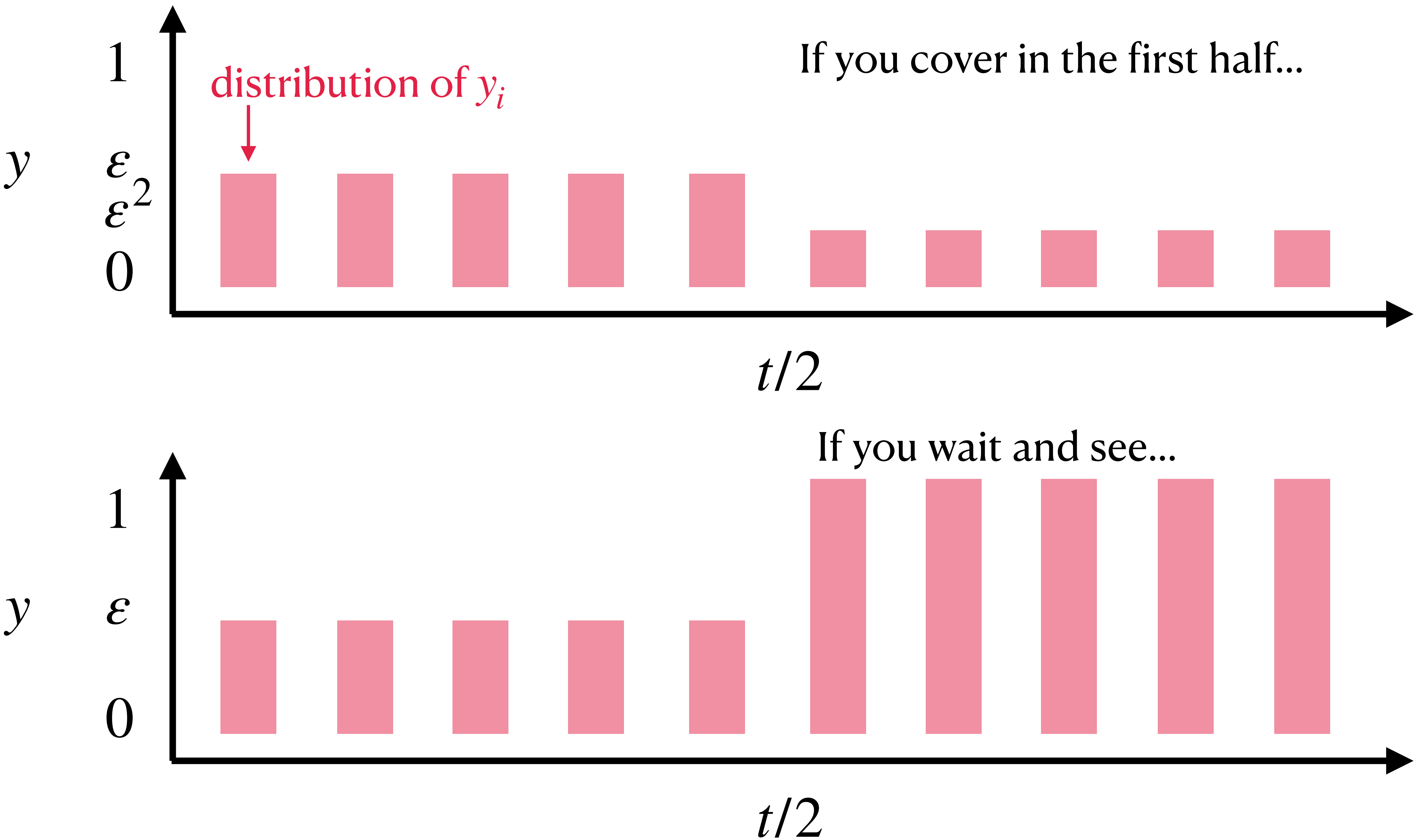


Sneak peek:

Prediction set problem very different from **regression/estimation** problem

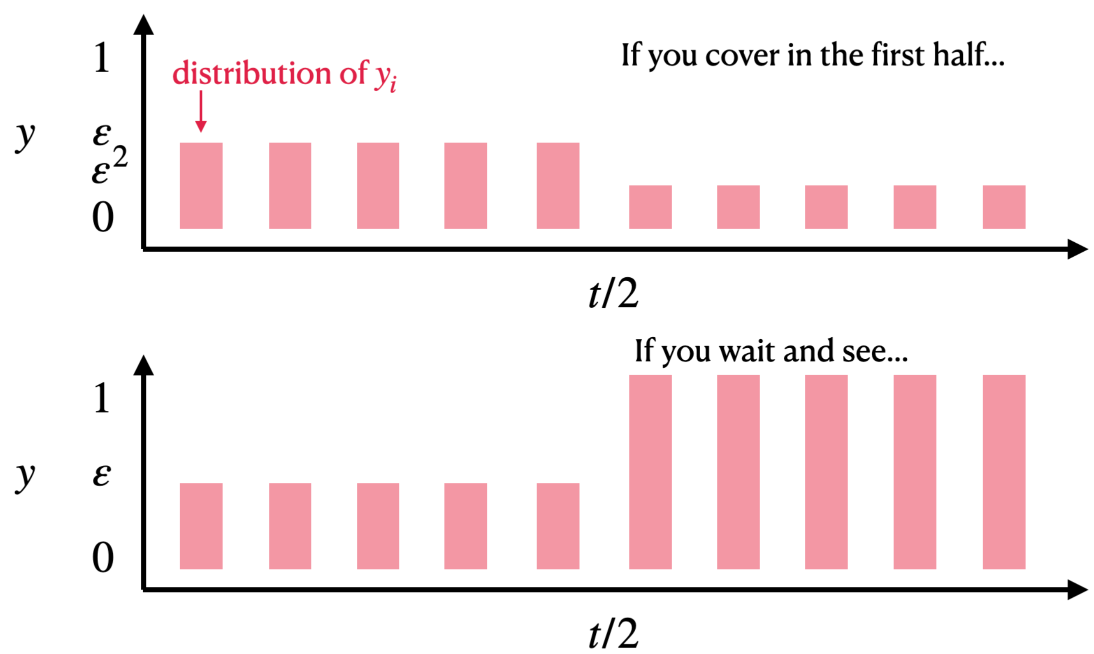
Hurdle

Goal: coverage $1 - \alpha = 0.5$



Have to lose in either **coverage** or **volume**

Are we toast?



Previous example:

For $1 - \alpha = 0.5$, can't simultaneously

- achieve non-trivial volume guarantee better than $1/\varepsilon$
- capture $1 - \alpha$ fraction of points

Relaxed objective:

Allow multiplicative approximation factors in

- **Volume:** compared to best interval that captures $(1 - \alpha)$-fraction of points
- **Miscoverage:** number of points not covered, compared to target αT
  - Interesting for $\alpha < 1/2$

Updated Question: What are the Pareto-optimal **bicriteria approximations**?

Result: Arbitrary-order Sequences

$\mathsf{Opt}_S(\alpha)$: volume of smallest interval in hindsight achieving coverage $1 - \alpha$ on S

Informal Theorem [S. '25]

For a given scale lower bound $\varepsilon > 0$, multiplicative volume approximation $\mu > 3$, target miscoverage rate $\alpha \geq 0$, and time horizon T,
we give a deterministic algorithm that on any sequence S of length T plays intervals of *maximum* volume

$$\leq \mu \max\{\mathsf{Opt}_S(\alpha),\ \varepsilon\},$$ **(efficiency)**

and makes number of mistakes bounded by

$$O\left(\frac{\log(1/\varepsilon)}{\log(\mu)}(\alpha T + 1)\right),$$ **(coverage)**

and this is near-optimal.

Stark tradeoff between coverage and efficiency, no **vanishing regret** possible!

Interpretation

Constrained Online Learning:

- Related to **binary classification** with hypothesis class of intervals
  - Learn best labeling of points as $+$ or $-$
- Think of every point y_i as labeled positive $\rightarrow (y_i, +)$
- Minimize volume of intervals played, subject to classification error $\leq \alpha$
- **Takeaway 1:** Unconstrained online learning admits **vanishing regret**, but constrained online learning looks very different!

Standard Recipe for Conformal Prediction:

- **Quantile regression** achieves coverage approaching $1 - \alpha$ as $T \rightarrow \infty$
- **Takeaway 2:** this strategy achieves unboundedly bad volume approximations in the worst case!
- Achieving small prediction set size requires a different approach

Algorithm: Volume

Goal: optimize **volume** with respect to a **coverage** constraint

Intuition: Optimizing **coverage** with a **volume** constraint would be easy!

Convert **feasibility** $\longleftrightarrow$ **optimization**

Informal Algorithm:

Input: $\alpha < 1/2$, vol. approx. factor μ

$I_{\text{current}} \leftarrow [0,0]$

For day t:

- If I_{current} missed more than αT points seen so far:
  - $I_t \leftarrow$ smallest interval that makes at most αT mistakes so far
  - $I_{\text{current}} \leftarrow \mu I_t$
- Predict I_{current}

Optimize coverage with volume constraint

Volume Approximation:

- Opt always feasible choice for I_t
- Never play intervals more than μ times bigger than Opt

Algorithm: Coverage

Informal Algorithm:

Input: $\alpha < 1/2$, vol. approx. factor μ

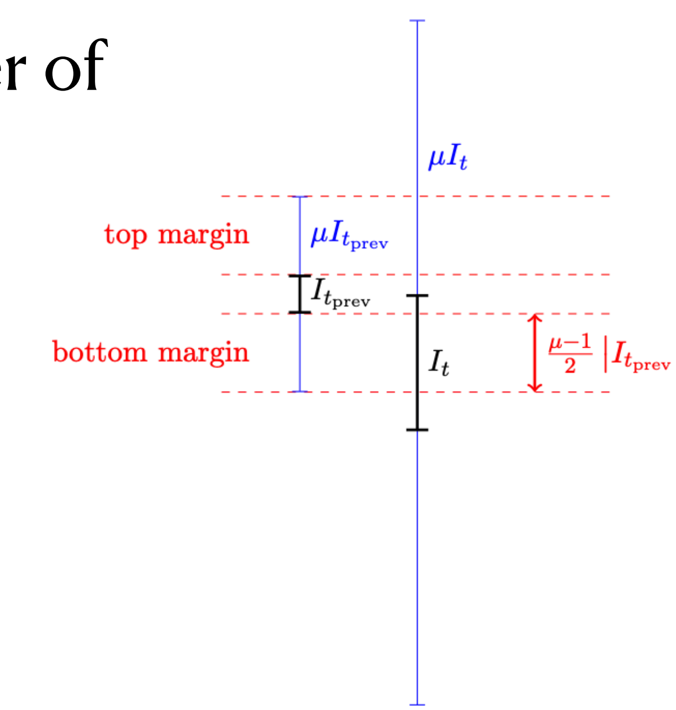
$I_{\text{current}} \leftarrow [0,0]$

For day t:

- If I_{current} missed more than αT points seen so far:
  - $I_t \leftarrow$ smallest interval that makes at most αT mistakes so far
  - $I_{\text{current}} \leftarrow \mu I_t$
- Predict I_{current}

Coverage Approximation:

- Each choice of I_{current} misses at most αT points
- Bound # of times we reset I_{current}
  - New I_t captures at least one point in the old I_t, and at least one point outside I_{current}
  - I_{current} grows by factor $\approx \mu$
- Bound number of iterations by $\dfrac{\log(1/\varepsilon)}{\log(\mu)}$



Algorithm: Coverage

Informal Algorithm:

Input: $\alpha < 1/2$, vol. approx. factor μ

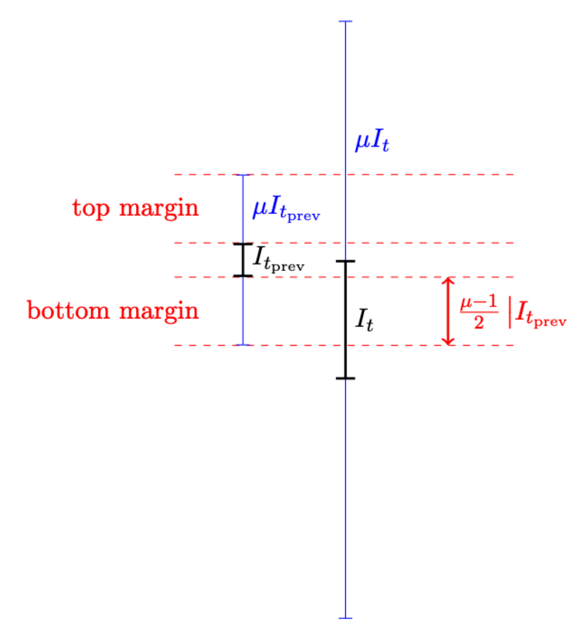
$I_{\text{current}} \leftarrow [0,0]$

For day t:

- If I_{current} missed more than αT points seen so far:
  - $I_t \leftarrow$ smallest interval that makes at most αT mistakes so far
  - $I_{\text{current}} \leftarrow \mu I_t$
- Predict I_{current}

Coverage Approximation:

- Each choice of I_{current} misses at most αT points
- Bound number of times we choose I_{current}
  - New I_t captures at least one point in the old I_t, and at least one point outside I_{current}
  - I_{current} grows by factor $\approx \mu$
- Bound number of iterations by $\dfrac{\log(1/\varepsilon)}{\log(\mu)}$



Result again: Arbitrary-order Sequences

Informal Theorem [S. '25]

We give a deterministic algorithm that on any sequence S of length T plays intervals of *maximum* volume

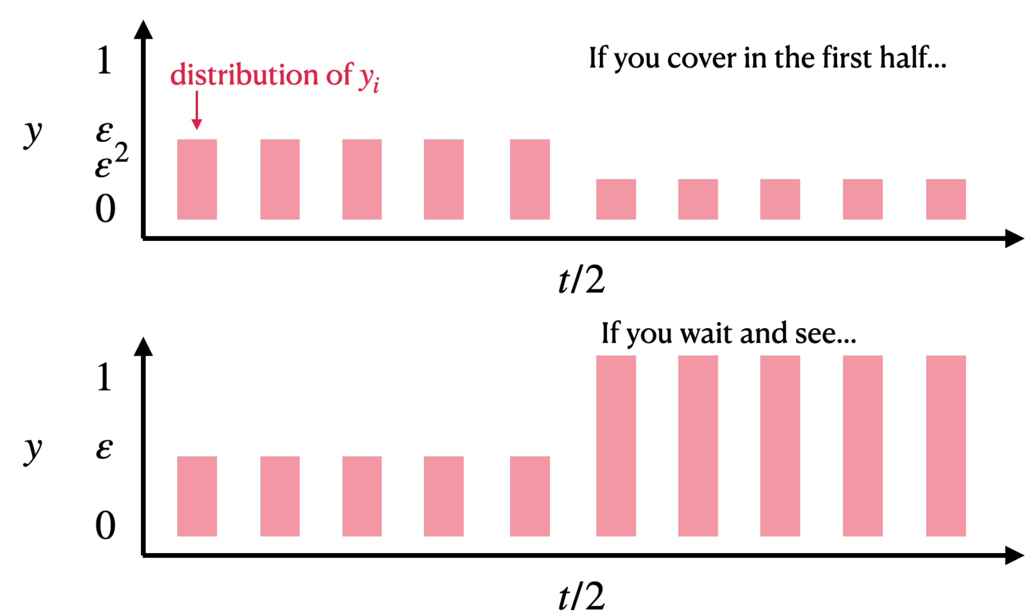
$$\leq \mu \max\{\mathsf{Opt}_S(\alpha),\ \varepsilon\}, \quad \textbf{(efficiency)}$$

and makes number of mistakes bounded by

$$O\left(\frac{\log(1/\varepsilon)}{\log(\mu)}(\alpha T + 1)\right), \textbf{(coverage)}$$

and this is near-optimal.

Lower bound is a generalization of earlier example, with more stair steps



Zoom Out: Results in this Work

Arbitrary order sequences:

- Must incur multiplicative factor approximations in **volume** and **miscoverage**

Random order sequences:

- Can approach optimal **volume** and **miscoverage** simultaneously!
- Close to **standard conformal prediction** (vs. online conformal prediction)
- Almost the same algorithm! (Reset I_{current} more aggressively, for lower error rates)

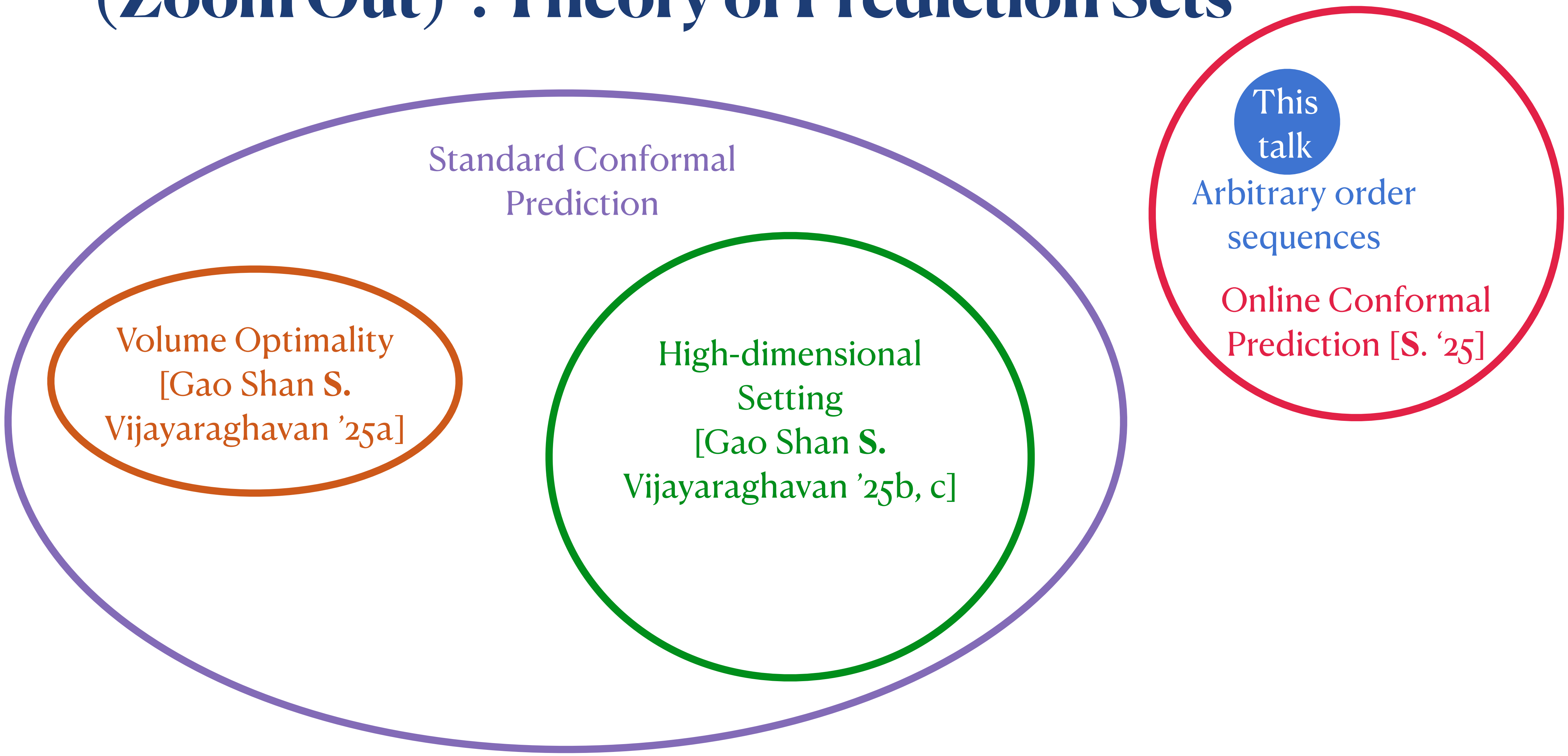
Why settle for **almost**?

No best-of-both worlds:

- No single algorithm can be optimal for both arbitrary and random-order sequences
- Can design algorithm to achieve the optimal trade-off

Informs what kinds of guarantees we can hope for

$(\text{Zoom Out})^2$: Theory of Prediction Sets



Learning **prediction sets** is fundamentally different than other learning tasks like estimation, regression, and classification, and requires **new theory**

Open problems: almost everything! Come join us :)

Thanks!