Reliable Online Prediction with Refuse Option

Ph.D. Dissertation Defense

Mustafa Anil Kocak

NYU Tandon School of Engineering

Introduction

- Machine learning and prediction algorithms are the building blocks of automation and forecasting.

- Reliability is crucial in risk-critical applications.
    - Analytics, risk assessment, credit decisions.
    - Health care, medical diagnosis.
    - Judicial decision making.

- Basic idea: Create a meta-algorithm that takes predictions from underlying machine learning algorithms and decides whether to pass them on to higher level applications.
- Goal: Achieve *robust correctness guarantees* for the predictions emitted by the meta-algorithm.

What Does It Mean to Refuse?

- The implications of refusing to make a prediction may vary according to the application of interest.
  - do more tests / collect more data
  - request user feedback or ask for a human expert to make the decision.
- Want to refuse seldom while still achieving the error bound.

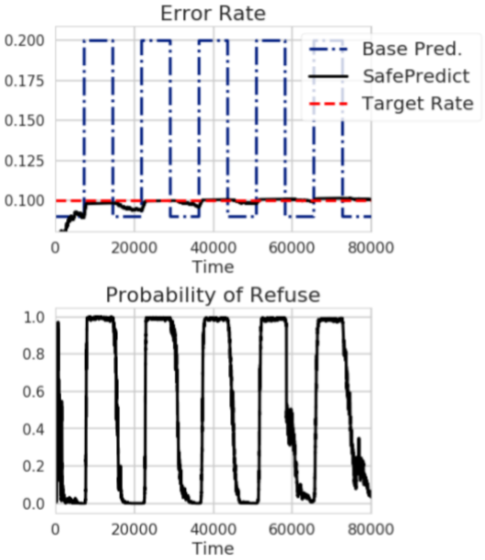
Thesis Outline

- Part I (Covered in the Area Exam) : *Conjugate Prediction*
  - Assumes exchangeable data stream.
  - Kocak, Mustafa Anil, Dennis E. Shasha, and Elza Erkip. "Conjugate Conformal Prediction for Online Binary Classification." *UAI*. 2016.
- Part II (Today) : *SafePredict*
  - Adversarial data, i.e. no data assumptions.
  - Kocak, Mustafa A., David Ramirez, Elza Erkip, and Dennis E. Shasha. "SafePredict: A Meta-Algorithm for Machine Learning That Uses Refusals to Guarantee Correctness." *Submitted to IEEE TPAMI*. August 2017.

SafePredict: Novelty and Teaser

- *SafePredict* achieves a desired error bound without any assumption on the data or the base predictor.

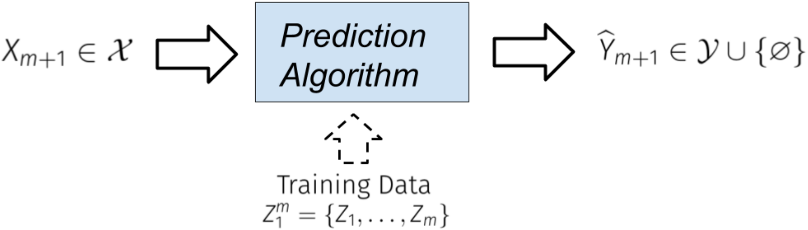
- Tracks the changes in the error rate of the base predictor to avoid refusing too much.



Literature Review

Batch Setup

Data: $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y} \sim$ i.i.d. $\mathcal{D}$ for all $i = 1, \ldots, m+1$.



Probability of Error (P_e)

$$P\left(\widehat{Y}_{m+1} \notin \{Y_{m+1}, \varnothing\} \mid Z_1^m\right)$$

Probability of Refusal (P_r)

$$P\left(\widehat{Y}_{m+1} = \varnothing \mid Z_1^m\right)$$

Batch Setup Goal: Minimize $P_e + \kappa\, P_r$, where κ is the cost of a refusal relative to an error.

Related Work (Batch Setup)

- *Chow, 1970:* Assuming $\mathcal{D}$ is known, the optimal refusal mechanism is:

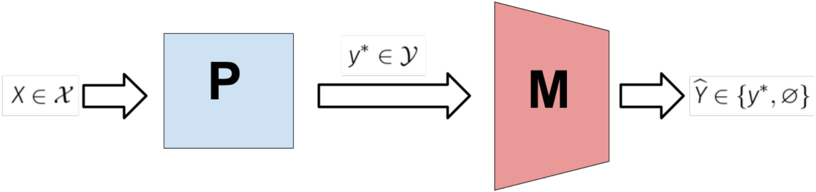
$$\widehat{Y}(X) = \begin{cases} y^* & \text{if } P(Y = y^* | X = x) \geq 1 - \kappa \\ \varnothing & \text{otherwise} \end{cases},$$

  where $y^* = \arg\max_y P(Y = y^* | X)$ is the MAP predictor.

- For unknown $\mathcal{D}$, instead minimize $\widehat{P}_e + \kappa \, \widehat{P}_r$.
  - *Wegkamp et al., 2006-2008:* Rejection with hinge loss and lasso.
  - *Wiener and El Yaniv, 2010-2012:* Relationship with active learning and selective SVM.
  - *Cortes et al., 2016-2017:* Kernel based methods and boosting.

Refuse Option via Meta-Algorithms

In practice, a meta-algorithm approach is much more common.



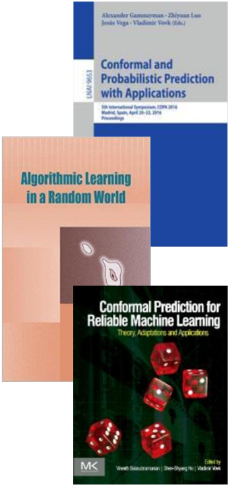
Base Predictor P is characterized by a scoring function S:

- $S(X, Y)$: How typical/probable/likely is (X, Y)?
- $y^* = \arg\max_{y \in \mathcal{Y}} S(X, y)$

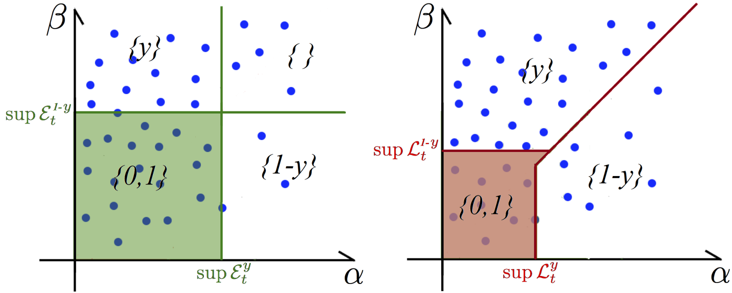
Meta-algorithm M characterized by τ: $\widehat{Y}(X) = \begin{cases} y^* & \text{if } S(X, y^*) \geq \tau \\ \varnothing & \text{otherwise} \end{cases}$.

Conformal Prediction (Vovk et al., 2005):

- Conformity score, $S(x, y)$, measures how well (x, y) conforms with the training data.
  - e.g. distance to the decision boundary, out-of-bag scores, other probability estimates.
- Strong guarantees in terms of coverage, i.e. $P_e \leq \epsilon + o(1)$.
  - Errors in first N points $\sim Binom(N, \epsilon)$.
- Probability of refusal is asymptotically minimized if S is consistent estimate of $\mathcal{D}$.



Conjugate Prediction: Reminder



Conformal Predictor ***Conjugate Conformal Predictor***

Conjugate Prediction (Kocak et al., 2016):

- Multi-dimensional scoring functions.
- Fewer refusals for stable S.

Online/Adversarial Setup

- *Online:* First observe $x_1, \ldots, x_t$ and $y_1, \ldots, y_{t-1}$, then predict $\hat{y}_t$.
    - *For each $t = 1, \ldots, T$:*
        - *i.* *Observe x_t.*
        - *ii.* *Predict $\hat{y}_t$.*
        - *iii.* *Observe y_t and suffer $l_t \in [0, 1]$.*
- *Adversarial:* Assume nothing about the data.
    - Instead assume access to a set of predictors : $P_1, P_2, \ldots, P_N$.

Related Work (Online/Adversarial Setup)

i. Realizable Setup: Assume there exists a perfect predictor in the ensemble.

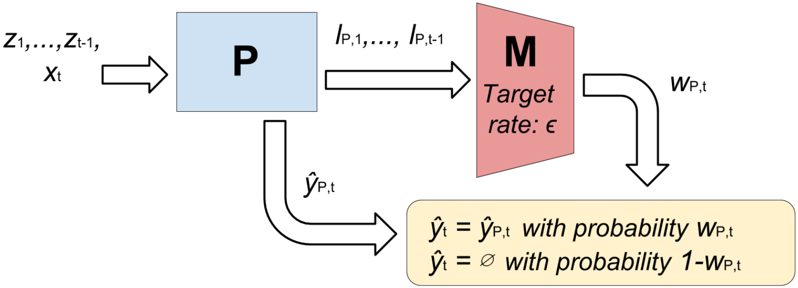
- *"Knows What it Knows"* (Li et al., 2008): Minimize the number of refusals without allowing any errors.
- *"Trading off Mistakes and Don't-Know Predictions,"* (Sayedi et al, 2010): Allow up to k errors and minimize the refusals.

ii. l-bias Assumption: One of the predictors makes at most l mistakes.

- *"Extended Littlestone Dimension"* (Zhang et al., 2016): Minimize the refusals while keeping the number of errors below k.

SafePredict

Meta-Algorithms in Online Prediction Setup



- Base-algorithm P makes prediction $\hat{y}_{P,t}$ and suffer $l_{P,t} \in [0,1]$.
- Meta-algorithm M makes a (randomized) decision to refuse ($\varnothing$) or predict $\hat{y}_t$, to guarantee a target error rate ϵ.
  - M predicts at time t with probability $w_{P,t}$.

Validity and Efficiency

- We use the following * notation to denote the averages over the randomization of M, i.e.

  T^*: Expected number of (non-refused) predictions, $\sum_{t=1}^{T} w_{P,t}$.

  L_T^*: Expected cumulative loss of M, $\sum_{t=1}^{T} l_{P,t} w_{P,t}$.

Validity

M is valid if $\lim \sup_{T^* \to \infty} \frac{L_T^*}{T^*} \leq \epsilon$.

Efficiency

M is efficient if $\lim \inf_{T^* \to \infty} \frac{T^*}{T} = 1$.

SafePredict Goal: *M should be valid for **any** P and be efficient when P performs well.*

Background: Expert Advice and EWAF

- How to combine expert opinions $P_1, \dots, P_N$ to perform almost as well as the best expert?

Exponentially weighted average forecasting (EWAF)

(Littlestone et al., 1989) (Vovk, 1990)

Intuition: Weight experts according to their past performances.

0. Initialize $(w_{P_1,1}, \dots, w_{P_N,1})$ and choose a learning rate $\eta > 0$.
1. For each $t = 1, \dots, T$
   1.1. Follow P_i with probability $w_{P_i,t}$.
   1.2. Update the probability $w_{P_i,t+1} \propto w_{P_i,t} e^{-\eta l_{P_i,t}}$.

- Regret Bound: $L_T - \min_i L_{P_i,T} \leq \sqrt{T \log(N)/2}$ where L_T and $L_{P_i,T}$ are the cumulative losses of EWAF and P_i.

Dummy and SafePredict

- We compare P with a dummy predictor (D) that refuses all the time.

$$l_{D,t} = \epsilon, \quad y_{D,t} = \varnothing.$$

- SafePredict is simply running EWAF over the ensemble $\{D, P\}$.

- EWAF regret bound implies $L_T^*/T^* - \epsilon = O\left(\sqrt{T}/T^*\right)$.

Therefore, for validity, we need a better bound and a more careful choice of η.

Theoretical Guarantees (Validity)

Theorem (Validity) [1]

Denoting the variance for the number of predictions with V^* and choosing $\eta = \Theta\left(1/\sqrt{V^*}\right)$, SafePredict is guaranteed to be valid for any P. Particularly,

$$\frac{L_T^*}{T^*} - \epsilon = O\left(\frac{\sqrt{V^*}}{T^*}\right) = O\left(\frac{1}{\sqrt{T^*}}\right),$$

where $V^* = \sum_{t=1}^{T} w_{P,t} w_{D,t}$.

[1] *In practice, V^* can be estimated via so called "doubling trick".*

Theoretical Guarantees (Validity)

Outline of the proof:

First define $l_t = w_{P,t} l_{P,t} + w_{D,t}\epsilon$ and $e^{-\eta m_t} = w_{P,t} e^{-\eta l_{P,t}} + w_{D,t} e^{-\eta\epsilon}$.

Key Fact 1: For $M_T = \sum_t m_t$ and $L_{P,T} = \sum_t l_{P,t}$, we have

$$e^{-\eta M_T} = w_{P,1} e^{-\eta L_{P,T}} + w_{D,1} e^{-\eta \epsilon T}$$

$$e^{-\eta M_T} \geq w_{D,1} e^{-\eta \epsilon T} \Longrightarrow M_T \leq \epsilon T - \frac{\log w_{D,1}}{\eta}.$$

Key Fact 2: $\sum_t l_t - m_t \leq \sum_t \eta w_{P,t} w_{D,t} (1-\epsilon)^2 = \eta (1-\epsilon)^2 V^*$.

Finally, combine these two facts and choose $\eta = \Theta\left(1/\sqrt{V^*}\right)$

$$\sum_t l_t \leq \epsilon T - \frac{\log w_{D,1}}{\eta} + \eta (1-\epsilon)^2 V^* \Longrightarrow L_T^* \leq \epsilon T^* + O\left(\sqrt{V^*}\right).$$

Theoretical Guarantees (Efficiency)

SafePredict is efficient as long as P has an error rate less than ϵ and η vanishes slower than $1/T$. Formally,

Theorem (Efficiency)

If $\limsup_{t\to\infty} L_{P,t}/t < \epsilon$ and $\eta T \to \infty$, then SafePredict is efficient. Furthermore, the number of refusals are finite almost surely.

Theoretical Guarantees (Efficiency)

Key Lemma: $w_{D,t+1} \leq \frac{w_{D,1}}{w_{P,1}} e^{\eta(L_{P,t} - \epsilon t)}.$

Outline of the proof:

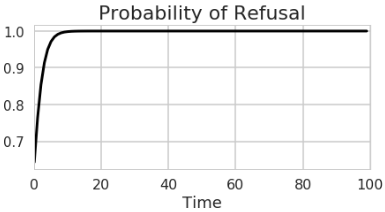
$$\lim_{T\to\infty} \frac{T - T^*}{T} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} w_{D,t}$$
$$\leq \lim_{T\to\infty} \frac{1}{T} \left(w_{D,1} + \frac{w_{D,1}}{w_{P,1}} \sum_{t}^{T-1} e^{\eta(L_{P,t} - \epsilon t)} \right)$$
$$< \frac{\text{constant}}{\infty} = 0. \quad \Longrightarrow \quad \lim_{T\to\infty} \frac{T^*}{T} = 1.$$

The last part directly follows from $\sum_t w_{D,t} < \infty$ and the Borel-Cantelli Lemma.

Making SafePredict Adaptive

- Probability of making a prediction decreases exponentially fast if the base predictor has a higher error rate than ϵ. Therefore, it is hard to recover from long sequences of mistakes.
  - Probability of refusal only depends on the cumulative loss of P.
  - e.g. cold starts, concept changes.

- *Toy example:*

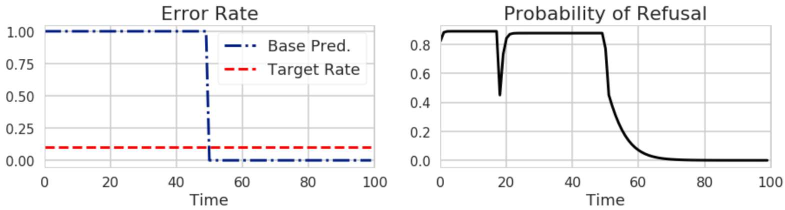


Making SafePredict Adaptive: Weight Shifting

Weight-shifting: At each step, shift α portion of the D's weight towards P, i.e.

$$w_{P,t} \leftarrow \; w_{P,t} + \alpha w_{D,t} = \alpha + (1-\alpha) w_{P,t}.$$

- Guarantees that $w_{P,t}$ is always greater than α.
- Inspired by Fixed Share algorithm (Herbster et al., 1998).
- *Toy example:*



Weight Shifting Guarantees

Validity Corollary:

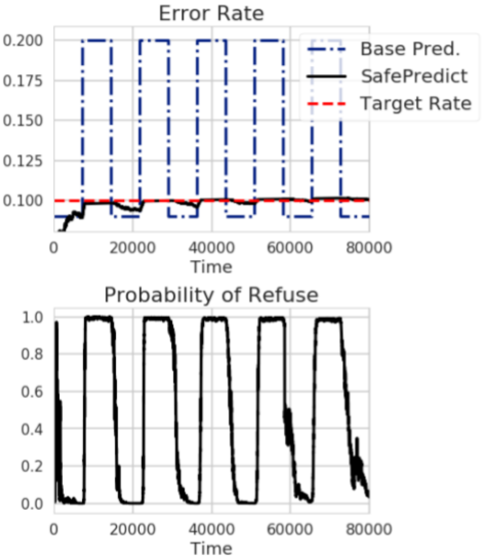
The validity guarantee is preserved for $\alpha = O(1/T)$,

$$\frac{L_T^*}{T^*} - \epsilon = O\left(\frac{\sqrt{V^*}}{T^*}\right).$$

Efficiency Lemma:

Refuse probability decreases exponentially, if P performs better than D after t_0,

$$w_{D,t} \leq e^{\eta\left(\sum_{\tau=t_0}^{t-1} l_{P,\tau} - \epsilon(t-t_0)\right)}/\alpha.$$



SafePredict: Implementation

Weight-Shifting SafePredict with Doub. Trick
Base predictor: P; *Initial weight*: $w_{P,1} \in (0,1)$
Target error rate: $\epsilon \in (0,1)$; *Adaptivitiy Parameter*: $\alpha \in [0,1)$

Initialize $t = 1$
for each $k = 1, 2, \ldots$ **do**
    Reset $w_{P,t} = w_{P,1}$, $V_{sum} = 0$, and

$$\eta = \sqrt{-\log\left(w_{D,1}\,(1-\alpha)^{T-1}\right)/(1-\epsilon)^2/2^k}$$

    while $V_{sum} \leq 2^k$ **do**
        Predict with probability $w_{P,t}$, refuse otherwise,

$$\hat{y}_t = \begin{cases} \hat{y}_{P,t} & \text{with prob. } w_{P,t} \\ \varnothing & \text{otherwise} \end{cases}$$

        Update the prediction probability:

$$w_{P,t+1} = \alpha + (1-\alpha)\frac{w_{P,t}e^{-\eta l_{P,t}}}{w_{P,t}e^{-\eta l_{P,t}} + w_{D,t}e^{-\eta\epsilon}}$$

        Compute $V_{sum} \leftarrow V_{sum} + w_{P,t+1}w_{D,t+1}$
        Increment t by 1, i.e. $t \leftarrow t + 1$

Choose the learning rate to minimize the excess error bound $\eta = K/\sqrt{V^*}$.

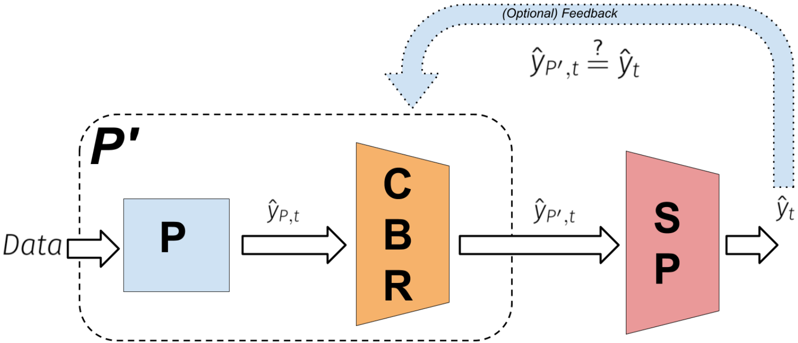
Use the *“Doubling trick”* to estimate V^*.

The validity guarantee is loosened by only a constant multiplicative factor of $\sqrt{2}/(\sqrt{2}-1)$.

Hybrid Approach and Amnesic Adaptivity

- SafePredict uses only the loss values while deciding to refuse or predict. Therefore, it only infers when it is safe to predict.
  - Robust validity under any conditions.
- Conformity based refusal mechanisms (CBR) use the data itself and pick out the easy predictions ***assuming*** all the data points are coming from (roughly) the same distribution.
  - Higher efficiency when the data is i.i.d.
- **Hybrid Approach:** Employ SafePredict on top of other refusal mechanisms for the best of the both worlds.

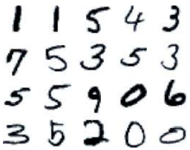
Hybrid Approach and Amnesic Adaptivity



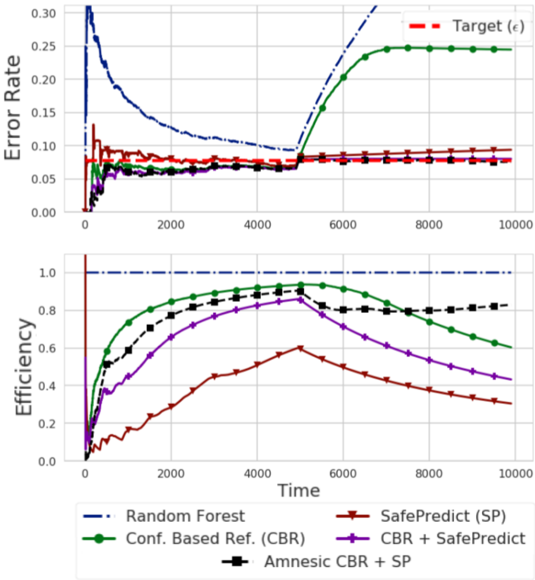
- If Confidence Based Refusal (CBR) mechanism predicts but SafePredict refuses, interpret this as violation of *i.i.d.* assumption.
  - **Amnesic adaptation:** if 50% of the last 100 predicted data points are refused by SafePredict, forget the history and reset the P'.

Numerical Experiments

Numerical Experiment (MNIST)

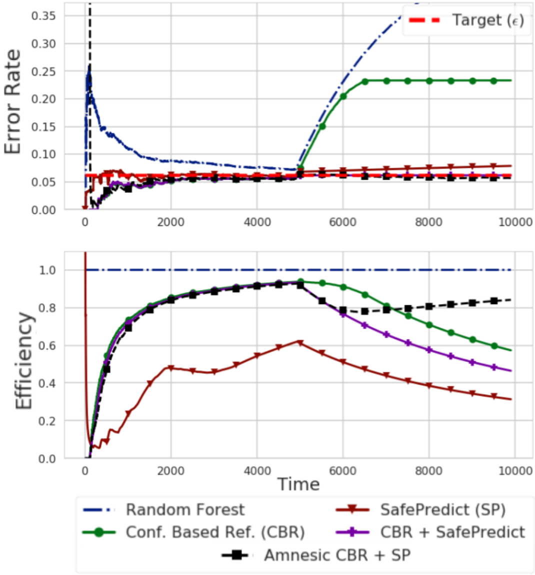


- $T = 10,000$.
- $\alpha = 10/T = 0.01$.
- P: Random forest retrained at every 100 data points.
- Change Point at $t = 5000$ (random permutation of labels).



Numerical Experiment (COD-RNA)

- Detection of non-coding RNAs *(Uzilov, 2006)*
- $T = 10,000$.
- $\alpha = 10/T = 0.01$.
- P: Random forest retrained at every 100 data points.
- Change Point at $t = 5000$ (random permutation of labels).



Conclusion

- We recast the exponentially weighted average forecasting algorithm to be used as a method to manage refusals.
- SafePredict works with any base prediction algorithm and asymptotically guarantees an upper bound on the error rate for non-refused predictions.
- The error guarantees do not depend on any assumption on the data or the base prediction algorithm.
- In changing environments, weight-shifting and amnesic adaptation heuristics boost efficiency while preserving the validity.
- Paper : https://arxiv.org/abs/1708.06425
  I-Python Notebooks : https://tinyurl.com/yagw3xzx

Questions?

Back-up Slides

Conformity Based Refusals

1. Split the training set as **core training**, Z_1^n, and **calibration**, Z_{n+1}^{n+l}, sets where $n + l = m$.
2. Train the base classifier P on the **core training set.**
3. Choose the smallest threshold that gives an empirical error probability on the **calibration set** less than ϵ, i.e.

$$\tau^* = \inf\left\{\tau : \sum_{i=m+1}^{m+l} 1_{\widehat{Y}_i \notin \{Y_i, \varnothing\}} / \sum_{i=m+1}^{m+l} 1_{\widehat{Y}_i \neq \varnothing} \leq \epsilon\right\}.$$

   - This operation takes $O(l)$ computational time.

Then we have the following guarantee:

Theorem

We have $P_e \leq \epsilon + \frac{1}{1 - P_r}\sqrt{\frac{\log(l/\delta)}{2l}}$ with probability at least $1 - \delta$.

CBR: Experiments